Lecture 3
Descriptive Statistics

What We Will Do Today

- Descriptive statistics (Agresti Chapter 3)
  - An abridged version; some stuff I don't get to from chapter 3, will go back to later
- Introduction to Stata
- Discuss Jez (2012)

Administrative issue

- Homework due next Wednesday
  - Will put homework (and formatting requirements) on D2L tomorrow morning; what I assign depends on how far we get today
- We will move discussion of Teranishi (2004) from 9/12 class to 9/19 class
  - Have updated syllabus on D2L to reflect this

I love descriptive statistics!

- Advanced modeling techniques are difficult to understand
  - Results are worthless if implemented incorrectly
- Descriptive statistics are simple to understand
  - Results are usually not worthless
- There are many jobs in K-12 and higher education for people who can create simple tables and graphs of descriptive statistics
  - Example: Director institutional research (typically pays in excess of $100,000)

Descriptive Statistics

- What we will cover
  - Frequencies
  - Describing center of data (mean, median)
  - Shape of frequency distributions & skewness
  - Describing variability (standard deviation)

Frequencies

- Frequency:
  - Number of observations that have a particular value for a variable (e.g., number of students w/ SAT score= 1100)
- Frequency distribution:
  - Listing of possible values for a variable, with the number of observations having that value
- Relative frequency:
  - proportion or % of observations having a particular value
- Relative frequency distribution:
  - List of possible values for a variable, with proportion of observations at each value
- Show example in Stata w/ offerlev2 variable

Frequency & Relative Frequency

- Use IPEDS data
- Fore each variable,
  - What kind of variable is it (nominal, ordinal, continuous)?
  - What is the frequency (for a particular category)?;
  - What is the relative frequency (for a particular category)?
  - What is the frequency distribution?
  - What is the relative frequency distribution
- Show graphically?

Describing Center of Data: Mean and Median

First, introduce summation term $\sum_{i=1}^{i=n} y_i$

- Variable y_i has n observations:
  - i refers to each observation; $y_{i=3}$ refers to the 3rd observation; n refers to the total number of observations
- List of all observations:
  - $y_1, y_2, y_3, y_4, \ldots\ldots y_n$
  - Example: dataset with n=5 observations: $y_1 = 5, y_2 = 3, y_3 = 12, y_4 = 43, y_5 = 39$
- $\sum_{i=1}^{i=n} y_i = \mathrm{y}_1 + \mathrm{y}_2 + \mathrm{y}_3 + \mathrm{y}_4 + \cdots \mathrm{y}_\mathrm{n}$
- $\sum_{i=1}^{i=n} y_i = 5 + 3 + 12 + 43 + 39$
- Show on board: i denotes observations

Summation term $\sum_{i=1}^{n} y_i$

- Show summation term in MS Excel

Describing Center of the Data: Mean

- Sample mean of a variable x, denoted $\bar{x}$
  - A measure of the average value of x
  - $\bar{x} = \frac{sum\ of\ all\ obs}{(\#\ of\ obs)} = \frac{\sum_i^n x_i}{n}$
- Ex: variable, x, with 6 observations $x_1 \ldots x_6$
  - Obs: $x_1 = 6, x_2 = 12, x_3 = 19, x_4 = 17, x_5 = 4, x_6 = 10$
  - $\bar{x} = \frac{\sum_i^n x_i}{n} = \frac{(6+12+19+17+4+10)}{6} = \frac{68}{6} = 11.3333$
- Show in Microsoft Excel
- Population mean $= \mu$ ("mu"); sample mean= $\bar{y}$

Describing Canter of Data: Median

- Median
  - Order numbers from lowest to highest.
  - Median is the value of the middle observation(s)
  - Half the obs are below the median, half are above
- Example: variable with 5 observations
  - Obs= 6, 12, 19, 17, 4
  - Order from lowest to highest: 4, 6, 12, 17, 19
- Example: variable with 6 observations
  - Obs= 6, 12, 19, 17, 4, 10
  - Order from lowest to highest: 4, 6, 10, 12, 17, 19
  - Median is midpoint between two middle values=11

The Median and Percentiles

- Median is also referred to as 50th percentile
  - 50% of obs fall below this value
- Percentiles (first, order obs from low to high)
  - **The X percentile is the value at which X% of observations fall below this value**
  - 25th percentile: 25% of obs fall below this value
  - 10th percentile: 10% of obs fall below this value
  - 75th percentile: 75% of obs fall below this value
- Show examples in Stata

Mean vs. Median

- What is the better measure of the “average value”?
- Want to create an estimate of average income in class by taking a sample of 7 observations
- Show an example in Excel

Outlier

- Outlier
  - an observation with an extremely large or extremely small value
- Examples:
  - Income: Bill Gates
  - Endowment size: Harvard
  - Book sales: The Hunger Games
  - Others?
- Means are sensitive to outliers; Medians are not

Shape of frequency distributions

- (for now pretend we are working with continuous variables, e.g., income)
- Frequency distribution
  - For each value, the number of observations that have that value
- This can be displayed in table or graphically
  - Show in Stata (in-state tuition for 4-yr publics)

Shape of frequency distributions

- Bell shaped, symmetric “tails”
  - Show example using made-up variable
- Skew (which tail is longer)
- Right skew (right tail is longer than left)
  - Most obs have smaller values than the mean; more positive outliers than you would expect in bell shaped variable
  - Example: income; enrollment size, country population
- Left skew (left tail is longer than right tail)
  - Most obs have larger values than the mean; more negative outliers than you would expect in bell shaped variable
- Show in OneNote

Skew and Mean vs. Median

- Mean
  - Sensitive to outliers (extreme observations)
- Median
  - Insensitive to outliers
- Skew and mean vs. median
  - Symmetric (tails are symmetric)
    - Mean = median
  - Right skew (right tail longer)
    - Mean is greater than median
  - Left skew (left tail longer)
    - Median is greater than mean

Shape of frequency distributions

- Show examples of each in Stata
  - X-axis=value of observations
  - Y-axis=number of observations
- Ask class:
  - Right-skewed, left-skewed, bell-shaped?
  - What is the value of the median relative to the mean?

Annual Production of Master's Degrees

	University Low Select	University Med Select	University High Select	Liberal Arts Low/Med Select	Liberal Arts High Select
MEAN					
1970	142	353	609	5	36
1980	226	466	728	14	22
1990	219	483	871	30	18
2000	303	631	1,035	77	19
2009	412	790	1,349	143	21
MEDIAN					
1970	48	148	343	0	3
1980	121	256	442	0	0
1990	117	254	551	0	0
2000	164	367	708	11	0
2009	234	473	881	42	0
N 1990	**304**	**380**	**39**	**484**	**33**

Variability

- We just did measures of centrality
  - Mean, median
- Now we move to measures of variability
  - Deviation
  - Standard deviation

Variability: Sample Standard Deviation

- What we want to find out:
  - What is the average distance between each observation and the mean?
  - Let's say avg. income (in means) in class is $30,000. Standard deviation tells us, if we select a random student, how far away their income is likely to be from $30,000.
    - Note: $24K and $36K would both be $6K away from mean
- Concepts we will use to calculate sample std dev
  - Mean
  - deviation

Variability of Data: Deviation

- Deviation:
  - Difference between an observation, y_i, and the sample mean, $\bar{y}$
  - Deviation of $y_i = (y_i - \bar{y})$
    - e.g., $y_i = 6;\ \bar{y} = 4;\ (y_i - \bar{y}) = 6 - 4 = 2$
  - Show in excel
    - Example: income
    - $\text{deviation}_\text{i} = (income_i - \overline{income})$

Could this be a measure of sample standard deviation?

- Potential measure
  - $Sample\ standard\ deviation = \frac{\sum_i^n (y_i - \bar{y})}{n-1}$
  - Answer: No!
  - Why not?
- Show in excel
- Why (n-1) instead of n? don't worry about it for now

Could this be a measure of sample standard deviation?

- Is this a good measure of standard deviation?
  - $- = \frac{\sum_i^n (y_i - \bar{y})^2}{n-1}$
  - Answer: No!
  - Why not?
    - This is average *squared* distance from the mean
    - Average squared distance from the mean is called the "variance"
- Show in excel
- Why (n-1) instead of n? don't worry about it for now

Sample Standard Deviation

- Definition (in words)
  - Avg absolute distance between an obs and the sample mean
- Definition (algebraic)

$$- s = \sqrt{\frac{\sum_i^n (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{sum\ of\ squared\ deviations}{sample\ size\ -1}}$$

- Do in excel
- Intuition
  - Why the squared term? Because the sum of all deviations=0
  - Why the square root? To compensate for the squared term
  - Why divide by sample size? To get the average (like calculating a mean)
    - Why n-1 instead of n? Don't worry about it.

Sample Standard Deviation

- Standard deviation

$$-\ s = \sqrt{\frac{\sum_i^n (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{sum\ of\ squared\ deviations}{sample\ size\ -1}}$$

- Example: variable with 5 observations
  - Obs= 6, 12, 19, 17, 4; mean=$\bar{y} = 11.6$
- $s = \sqrt{\frac{(6-11.6)^2+(12-11.6)^2+(19-11.6)^2+(17-11.6)^2+(4-11.6)^2}{5-1}}$
- $s = \sqrt{\frac{173.2}{4}} = 6.58$

Properties of Sample Standard Deviation

- Sample Standard deviation
  - Average absolute distance between an obs and the sample mean
- Some Properties
  - Standard deviation is sensitive to presence of outliers
  - Standard deviation does not necessarily increase or decrease in size as sample size increases

In Class Exercise

- Standard deviation

$$-s = \sqrt{\frac{\sum_i^n (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{sum\ of\ squared\ deviations}{sample\ size\ -1}}$$

- Observations:
  - 23, 35, 12, 8, 49
- Questions. What is the:
  - Mean? Median? Standard Deviation?

Notation: estimates vs. parameters

- Parameter
  - Based on the population
- Estimate (also called “statistic”)
  - Based on a sample
  - An estimate is our best guess of the parameter
- Estimates use regular alphabet, parameters use Greek alphabet

	Estimate (sample)	**Parameter (population)**
Mean	$\bar{x}$	$\mu \; or \mu_{\mathrm{x}}$
Standard deviation	$s \; or \; s_x$	$\sigma \; or \sigma_{\mathrm{x}}$
Sample size	n	N

Break

Introduction to Stata

- What we will do
  - Open a dataset in Stata
  - Familiarize w/ Stata windows
  - Download dataset from D2L and open in Stata
  - Run a few commands

Open Stata

- You can use lab PC or your laptop
  - If you use your laptop, make sure you are connected to wireless internet
- Open Stata
- Discuss the different “windows”
  - Big window= “results” window
  - Bottom of screen= “command” line
    - A command is when you tell Stata to do something
  - Bottom left=list of variables
  - Top left= list of commands you ran previously

Open a dataset in Stata

- Three ways to enter commands in Stata
  - (1) **command line**; (2) point-and-click; (3) “do” file
- Typing commands in command line
  - Do not type “bullets”; type commands in lowercase; type the command; press “return” to enter command
- Type the following commands (one at a time):
  - webuse census
  - Describe
  - tabulate region
- Tell students how to read description of dataset

Open a dataset in Stata

- Show students where to see actual dataset
- Show students the variables window
- Show students the “previous command window”
- How to enter previously run commands on command line
  - Click previous command on “previous command window”
  - Press “page up” button on keyboard
    - “page down” button shows subsequent command
- Note: you can copy commands onto MS Word file

Download Stata dataset from D2L

- Go to D2L website for this class
- Datasets >> datasets for “small” stata
  - Download ipeds_2010_small_stata (zipped)
- Save the zipped dataset to a folder (any folder) you can find again
- “unzip” (i.e., un-compress) the dataset
  - Show class how to do this on your PC, may be different on theirs
  - save it in same folder as the zipped dataset

Open the ipeds dataset in Stata

- Open Stata
- Using your mouse, *click* on:
  - File >> open >> (browse for “ipeds_2010_small_stata” and click “open”
- See the command you just ran in the “results window” and in “previous command window”
- Look at actual data in “Data Editor (Browse)”

Learn three Stata commands

- Three commands
  - *describe*
    - *Describes variables in the dataset*
  - *tabulate* variable_name
    - shows frequencies and relative frequencies
  - *summarize* variable_name
    - shows means, standard deviations, etc.
- Type these commands:
  - describe
  - tabulate sector
  - tabulate offerlev2
  - summarize tfugoutst
  - summarize totfte

In class exercise

- Open Stata
- Type “clear” in command line (without quotations) to clear any open dataset
- Open the dataset ipeds_2010_small_stata
- Describe the dataset
- For the variable totbachv2, what is the mean value? What is the standard deviation?
- For the variable sector, how many obs are “public, 4-year or above”? What percentage of obs are “public, 4-year or above”?
- What is the mean and standard deviation of the variable satp50