Chapter 4
Probability

What we will do today

- Introduction to Stata
- Agresti Chapter 4: probability distributions
- Want additional resources for class?
  - Optional book by Urdan
  - Khan Academy
    - http://www.khanacademy.org/

Introduction to Stata

Stata: What we will do today

- Download dataset
  - Why it didn't work before
- Open dataset in Stata
- Mainly work with three commands
  - Using the command "cheat sheet" I created
  - Command options
  - "value labels"
  - Using "if" logic to isolate certain cases (e.g., summarize income for people with college degree)

Download and open dataset

- Download dataset
  - Use link I emailed to you today:
    - Click “download” then “direct download”
- Save dataset to place you can find (e.g., desktop)
- Open Stata
  - Click “check next time” when it asks about updates
- Within stata
  - Using mouse: file >> open, then browse for dataset

Describe, view data

- type “describe” on command line (no quotations), then press “enter” on keyboard
  - Take some time to look through the different variables
- View actual raw data in Stata
  - At top of screen, click on icon that looks lake a table with a magnifying glass

Will work with 3 commands today

- Summarize
  - Show mean, standard deviation, median, etc.
  - Good for continuous variables
  - Syntax
    - summarize varname1 varname2…
- Tabulate (one way)
  - Show frequency distribution for one variable
  - Good for variables with only a few categories
  - syntax:
    - tabulate varname
- Tabulate (two way)
  - Show frequency distribution for two variables (e.g., show relationship between gender and political party)
  - Good for variables with only a few categories
  - Syntax
    - tabulate varname1 varname2

summarize

- Summarize
  - Show mean, standard deviation, median, etc.
  - Good for continuous variables
  - Syntax
    - summarize varname1 varname2…
- Type this:
  - describe hrs1 age childs
  - summarize hrs1
  - summarize age
  - summarize childs
  - summarize hrs1 age childs

tabulate (one way)

- Tabulate (one way)
  - Show frequency distribution for one variable
  - Good for variables with only a few categories
  - syntax:
    - tabulate varname
- Type this
  - tabulate sex
  - tabulate racecen1
  - tabulate race
  - tabulate demrepind
  - tabulate grass
  - tabulate homosex
  - tabulate marhomo

Tabulate (two way)

- Tabulate (two way)
  - Show frequency distribution for two variables (e.g., show relationship between gender and political party)
  - Syntax
    - tabulate varname1 varname2
    - Note: varname1 is rows; varname2 is columns
- Type this:
  - tabulate grass demrepind
  - tabulate homosex demrepind
  - tabulate marhomo demrepind

In-class exercises

- Describe the following variables (“describe” command)
  - hrs age demrepind owngun discaff taxshare abany abrape
- Find the mean value for the following variables (i.e., summarize command)
  - hrs1, age
- Show the frequency distribution of the following variables (i.e., tabulate command (one-way)
  - owngun, discaff, taxshare, abany, abrape
- Show two-say frequency distribution for the following variables (i.e., tabulate command (two-way)
  - owngun & demrepind; discaff & demrepind

Command options and command help

- For this class, two main ways for learning about commands
  - (1) the stata help function. Type “help” and then any command name. e.g.,:
    - help summarize
    - Problem with this: can be complicated
  - (2) I have created a “cheat sheet” for most commands we will use in this class. This has:
    - Command syntax
    - Important command “options”
    - examples

Learning “options” by using cheat sheet

- How to read syntax for (summarize) command:
  - summarize [varlist] [, options]
  - You are only required to type in the part of the command that is underlined. e.g.,:
    - su hrs1
    - sum hrs1
  - Anything in brackets [] is optional
    - e.g., if you don’t include variables in [varlist] Stata summarizes all variables. Try typing “sum”
- All Stata commands have options
  - Most useful option for summarize command: detail
  - sum hrs1, detail
  - sum hrs1, d

Tabulate command (one-way)

- tabulate varname [, options]
  - ta demrepind
  - tab demrepind
- Important options for tabulate command
  - missing
    - Show missing values
    - tab demrepind, missing
    - tab demrepind, m
  - nolabel
    - Displays “numeric values” rather than “value labels”
    - tab demrepind, nolabel
    - tab demrepind, m nolabel

Working with “Value Labels”

- Value labels
  - Instead of showing numeric code assigned to a category (e.g., 1), Stata shows a label that is easy to understand
  - Value labels are usually used for variables that take on a few different values
- Which variables have value labels?
  - “describe” the dataset
  - The column “value label” indicates whether value labels have been attached to each value
- Why would you want to exclude value labels?
  - Sometimes you want to isolate obs w/ certain values (e.g., all white people). [will show you how to do this later]
  - tab race
  - tab race, nolabel

Tabulate command (two-way)

- tabulate varname1 varname2 [, options]
  - Note: varname1 is rows; varname2 is columns
  - tab grass demrepind
  - tab demrepind grass
- Important options for tabulate (two-way)
  - missing
  - nolabel
  - column
    - Show column percentages
    - tab grass demrepind, col
  - row
    - Show row percentages
    - tab grass demrepind, row
  - nofreq
    - Exclude frequencies
    - tab grass demrepind, col nofreq

Tabulate command (two-way)

- Note on doing two-way tabulations
  - I “mess around” with commands until I get it right
  - I always forget this!:
    - varname1 is rows; varname2 is columns
- Type this:
  - tabulate demrepind marhomo
  - tabulate marhomo demrepind
  - tabulate marhomo demrepind, col
  - tabulate marhomo demrepind, col nofreq

In Class Exercises

- Practice using the “cheat sheet” to help you
- Summarize the variable age using the detail option
  - What is the standard deviation?
  - 50% of people are younger than what age?
  - 75% of people are younger than what age?
- Do a tabulation of variable abrape
- Do a tabulation of variable abrape with the missing option
  - If we include missing cases, what percentage of people say “yes” on abortion in cases of rape?
- Do two way tabulations of the following variables. For each tabulation, use col, row, and nofreq options to make the output look right to you.
  - sex and demrepind; sex and sexhar; choose others that are interesting to you

Isolating particular observations

- For (nearly) all Stata commands, can run command for a certain subset of observations (e.g., all men, all democrats, all people older than 60, etc.)
- uses “if” logic combined with “operators”
- Type this:
  - tab marhomo if demrepind==1
    - (Note the use of two equal signs)
  - tab marhomo if demrepind==3
  - tab marhomo if age>=60
  - tab marhomo if age<=30
  - tab marhomo if age<=30 & demrepind==3
- Operators (see “cheat sheet”)
  - == equal; > greater than; >= greater than or equal to; < less than; <= less than or equal to; != not equal

In Class Exercises

- Use Cheat sheet!
- What is mean hours worked (hrs1) for people 65 and older (age)?
- What percentage of democrats (demrepind) think taxes for rich people (taxrich) are “much too low”?
  - Note: first figure out the numeric value assigned to democrats:
    - tab demrepind, nolabel
- What percentage of republicans (demrepind) “strongly disagree” that government should reduce income differentials (goveqinc)

Agresti Chapter 4: Probability

Logistical issues

- Material gets more difficult starting today
  - Important to stay on top of the material
  - This class will get much easier after the midterm
- Make appointments with Scott and I if you are having difficulty
  - Please give us at least 48 hours notice

Continuation from last lecture

- Variables have frequency distributions
- Measures of the center of a frequency distribution (e.g., mean, median)
- Measures of variability of a frequency distribution (e.g., standard deviation)

Sample Standard Deviation

- Definition (in words)
  - A measure of avg distance between an obs and the mean
- Definition (algebraic)

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

- Intuition behind the formula
  - 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

- Definition (in words)
  - measure of the average distance of an obs from the mean
- Some intuition behind the concept
  - Gives you a sense of whether observations tend to be spread out far apart from one another or tend to be close to on another
  - Draw picture of two samples on one note
    - Which has greater standard deviation?
  - Draw picture of two frequency distributions
    - Which has greater standard deviation?

Properties of Sample Standard Deviation

- Sample Standard deviation

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

- Some Properties
  - The bigger the standard deviation, the further apart observations are from one another
  - Standard deviation is sensitive to presence of outliers
  - Standard deviation does not necessarily increase or decrease in size as sample size increases

Introduction to probability

Probability

- Probability is a big topic; we will do a very short, simple introduction
- We build on discussion of “relative frequency distribution”
  - Show relative frequency distribution of Barrons2004 in Stata
    - What is probability of having event A?

Probability

- Definition of probability
  - The probability an observation has a particular outcome is the proportion of times that outcome would occur in a long sequence of events
- Examples:
  - Flip a coin 1,000 times. Proportion of heads would be close to .5 (but maybe not if you flip coin 10 times)
  - Consider the variable Barrons2004. Imagine that we pick observations at random. What is the probability that an observation has the value “very competitive”?
- For this class, you can consider probability as measuring the same thing as “relative frequency”

Discrete vs. continuous variables

- Differentiate discrete vs. continuous variables
- Discrete
  - Formal: variable that can have a finite number of values
  - Informal: variable that can take on only a few values
    - Examples: Barrons2004, number of siblings
- Continuous
  - Formal: variable that can take on any real value
  - Informal: variable that can take on many values
    - Examples: SAT score, total enrollments

Discrete & Continuous Probability Distributions

- Probability distribution
  - Lists the possible outcomes and their probabilities of occurring
  - Same as relative frequency distribution
- Discrete
  - Show example
- Continuous
  - Show example

Properties of probability distribution

- (Will give more properties later)
- If we sum the probabilities of all possible outcomes, that sum will always equal 1
  - Show in Stata (Barrons2004)
- Probability distributions have measures of central tendency (e.g., mean, median) and measures of variability (e.g., standard deviation)
  - Show in Stata

Normal Distribution and Z-scores

Normal Distribution

- A special distribution; we will use the normal distribution a lot in this class
- Definition: the normal distribution is a symmetric, bell shaped distribution.
  - Not left-skewed or right-skewed
- Show what it looks like in OneNote
  - First show normal distribution, then add standard deviation
- Normal distribution has very useful properties
  - If we can reasonably assume a variable has a normal distribution, then we know a lot about that variable

Standard Deviation: Empirical Rule

- If variable has an approximately normal distribution (i.e., approximately "bell shaped") then:
- About 68% of obs fall within one std dev, s, of mean, $\bar{x}$
  - i.e., between $\bar{x} - s \; and \; \overline{x} + \mathrm{s}$
- About 95% of obs fall within two std. dev of mean
  - i.e., between $\bar{x} - 2s \; and \; \bar{x} + 2\mathrm{s}$
- About 99% of obs fall within three std. dev of mean
  - i.e., between $\bar{x} - 3s \; and \; \bar{x} + 3\mathrm{s}$
- Show in OneNote

Std. dev away from the mean

- We can think of the value of observation in terms of standard deviations away from the mean
- Example
  - Imagine mean SAT score is 1000 and standard deviation is 100
  - How many std dev from the mean is a score of 1100?
  - How many std dev from the mean is a score of 1200?
  - How many std dev from the mean is a score of 900?

Z-Score

- Z-score: how many standard deviations away from mean
- Definition
  - The z-score, z_i, of an observation, y_i, is the number of standard deviations, s, away from the mean, $\bar{y}$
  - $z_i = \frac{y_i - \bar{y}}{s}$
    - Numerator: deviations from the mean
    - Denominator: standard dev (i.e., scale in terms of std dev)
- Example: $y_i = 30;\ \bar{y} = 18.2; s = 5.3$
  - $z_i = \frac{y_i - \bar{y}}{s} = \frac{30 - 18.2}{5.3} = 2.226$

Z-Score

- Z-score: how many std dev away from mean
- Definition
  - $z_i = \frac{y_i - \bar{y}}{s}$
- Example (Stata): out of state tuition at 4-yr publics
  - How many std dev from the mean is tuition of:
  - 20,000?
  - 10,000?
  - 5,000?
  - 35,000?

Z-scores for normal distribution

- Show normal distribution in OneNote
- What is probability of picking an observation with a particular value (e.g., 11)?
  - Sum of all probabilities=1
  - Probability of picking an obs with a value greater than mean=.5; probability of picking an obs with value less than mean=.5
  - Symmetric: probability above mean is same as probability below mean

Z-scores for normal distribution

- Show normal distribution with z-scores
- Remember empirical rule
  - About 68% of obs fall within one std dev, s, of mean, $\bar{x}$
  - About 95% of obs fall within two std. dev of mean
  - About 99% of obs fall within three std. dev of mean

Table of Z-scores (inside cover Agresti)

- Shows the probability a random observation has a z-score at least as large
- Examples:
  - Z= 1.00: there is a 0.1587 probability a random observation has a z-score at least as large as 1.00
  - Z= 2.00: there is a 0.0228 probability a random observation has a z-score at least as large as 2.00
- Always Draw Pictures!
  - Draw normal distribution; label z-score(s) you are looking for; shade the region you are looking for
- Questions:
  - What is the probability an observation has a z-score greater than: 1.50? 1.96? 2.33?

Z-scores: a note on notation

- When working with z-scores, pretend that the variable z is continuous, taking on an infinite number of real values. Pretend that no observation can have a z-score of *exactly* 1.
  - So probability z>1 is the same as probability of z>=1
- Notation I will use:
  - Probability z is greater than 1
    - Pr(z>1)=
  - Probability z is less than 1
    - Pr(z<1)=
  - Probability z is greater than -1.5
    - Pr(z>-1.5)=

Agresti shows probabilities for Right Half of Normal Distribution

- Because a normal distribution is symmetric (show picture), the probabilities are the same for the bottom half of the distribution.
- What is the probability an observation has a z-score at least as small as -1 (i.e., $Pr(z<-1)$)
- What is the probability an observation has a z-score at least as small as -2 (i.e., $Pr(z<-2)$)
- Draw picture!

Z-score probabilities continued

- What is probability of having z-score less than 1.5?
  - Equals 1 minus probability of having z-score greater than 1.5
  - Probability rule: Pr(not A) = 1 – Pr(A)
  - Probability rule: $Pr(z<1.5)= 1 – Pr(z>1.5)$
  - Draw picture!
- What is probability of having z-score less than 1.96?
  - Draw picture!

Z-score probabilities continued

- What is probability of having z-score greater than -1.5?
  - Draw picture!
- What is probability of having z-score greater than -1.96?
  - Draw picture!

In class exercises

- Find the specific probability; draw a picture for each:
  - Probability that z is greater than 1.96 [i.e., Pr(z>1.96)]
  - Probability that z is less than 1.96 [i.e., Pr(z<1.96)]
  - Probability that z is greater than 2.33 [i.e., Pr(z>2.33)]
  - Probability that z is less than 2.33 [i.e., Pr(z>2.33)]
  - Pr(z>1.64)
  - Pr(z<1.64)
  - Pr(z>-1.64)
  - Pr(z<-1.64)

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