Probability &
Sampling Distributions

Administrative issues

- Reading chapter 5
  - Skip 5.4: choice of sample size
  - Skip 5.5: confidence intervals for median
  - (in general, skip all optional sections, marked *)
- Schedule pushed back one week
  - Updated syllabus on D2L

What we will do today

- Finish chapter 4
  - Z-scores and normal distribution
    - Finding probabilities associated w/ particular z-scores
  - Sampling distribution
    - A longer introduction than Agresti
- Discuss Teranishi (2004)

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
- 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, $\bar{x}$
- About 95% of obs fall within two std. dev of mean
- About 99% of obs fall within three std. dev of mean
- Show in OneNote

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, $\bar{y}$
  - $z_i = \frac{y_i - \bar{y}}{s}$
- Example: $\mathrm{y_i} = 30;\ \bar{y} = 18.2; s = 5.3$
  - $z_i = \frac{y_i - \bar{y}}{s} = \frac{30 - 18.2}{5.3} = 2.226$
  - The observation $\mathrm{y_i} = 30$ is 2.226 standard deviations from the mean

Standard Normal Distribution

- A *very* special distribution
  - A bell-shaped (i.e., normal) distribution that has mean=0 and standard deviation=1
- The value of each observation is already in terms of z-scores
  - Each observation shows how many standard deviations from the mean
- Question: if the variable has a standard normal distribution, would it be likely to see an observation with a value of 3?

Standard Normal Distribution

- Question:
  - If we have a variable with a roughly normal distribution (i.e., symmetrical), how could we transform it into a variable with a standard normal distribution (i.e., symmetrical with mean=0 and std deviation=1)?

Standard Normal Distribution

- Question: how transform a normal distribution into a standard normal distribution?
- Answer:
  - Find the z-score associated with each value
- Show in Stata
  - Show histogram of normal distribution variable
  - Show mean and std dev of normal distribution variable
  - Calculate z-score for normal distribution variable
    - Show list of observations
  - Show histogram of standard normal variable
  - Show mean and std deviation of standard normal variable

Z-scores and standard normal distribution

- Show standard normal distribution in OneNote
- We want to know the probability of picking an obs with a z-score value at least that high (e.g., 1.2)?
- Some basic rules:
  - 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 and std normal distribution

- Empirical rule and Z-scores
  - 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
- Show in OneNote

Table of Z-scores (inside cover Agresti)

- The distribution on the top is a standard normal distribution
  - i.e., values are in terms of standard deviations from the mean. i.e., values are z-scores
- Table shows the probability a random observation has a z-score at least as large
- How to read table:
  - Rows: first decimal place of z-score
  - Columns: second decimal place of z-score
  - Individual cells: probability of observing a z-score at least that large

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!
  - Write out probability you are looking for in words; draw normal distribution; label z-score(s) you are looking for; shade the region you are looking for; look at z-score table
- Questions:
  - What is the probability an observation has a z-score at least as large as: 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
    - $Pr(z>-1.5)=$

Z-scores continued

- Z-score table shows probabilities for "right half" of distribution
  - i.e., for observations greater than the mean, i.e., $z>0$
  - Because a normal distribution is symmetric (show picture), the probabilities are the same for the bottom half of the distribution.
- Find these probabilities (Draw a picture!)
  - What is the probability an observation has a z-score less than -1 (i.e., $Pr(z<-1)$)
  - What is the probability an observation has a z-score less than -2 (i.e., $Pr(z<-2)$)

In class exercises

- (Try a few until you feel comfortable)
- Find each 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>.49)?
  - Pr(z<-.49)?

Z-score probabilities continued

- What is probability of having z-score less than 1.5?
  - Draw picture!
  - 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)
- 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 (helpful to draw two pictures)!
- What is probability of having z-score greater than -1.96?
  - Draw picture (helpful to draw two pictures)!

Probability observation is *within* some range of z-scores

- What is probability obs is within one standard deviation of the mean? i.e., Pr(-1<z<1)
  - Draw picture; use symmetry
- What is probability obs is within two standard deviations of the mean? i.e., Pr(-2<z<2)
  - Draw picture; use symmetry
- What is probability obs is within three standard deviation of the mean? i.e., Pr(-3<z<3)
  - Draw picture; use symmetry

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 $\bar{x} - s \; and \; \bar{x} + s$
- About 95% of obs fall within two std. dev of mean
  - i.e., between $\bar{x} - 2s \; and \; \bar{x} + 2s$
- About 99% of obs fall within three std. dev of mean
  - i.e., between $\bar{x} - 3s \; and \; \bar{x} + 3s$
- Show in OneNote

Probability observation is *outside* some range of z-scores

- What is probability an obs is more than one standard deviation away from the mean? i.e., $Pr(z<-1 \text{ or } z>1)$
  - Draw picture; use symmetry
- What is probability an obs is more than 1.64 stand deviations away from the mean?
  - Draw picture; use symmetry

In class exercises (just try a few)

- Find each probability; draw a picture for each:
  - Probability that z is less than 1.96 [i.e., Pr(z<1.96)]
  - Probability that z is greater than -1.96 [i.e., Pr(z>-1.96)]
  - Probability that z is less than 2.33 [i.e., Pr(z<2.33)]
  - Probability that z is greater than -2.33 [i.e., Pr(z>-2.33)]
  - Probability that z is *within* .78 standard deviations of the mean?
  - Probability that z is more than .78 standard deviations away from the mean?
  - Probability that z is within 1.82 standard deviations of the mean?
  - Probability that z is more than 1.82 standard deviations away from the mean?

Sampling Distribution

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	$\bar{x}$	μ or μ_{x}
Standard deviation	s or s_x	σ or σ_{x}
Sample size	n	N

Population vs. sample

- IPEDS data has the entire population of postsecondary institutions
- Consider the variable "avg. SAT score of enrolled freshmen"
  - Let's call this variable the "institutional SAT score"
  - Assuming that there are no missing values, this value is a population mean, μ, not a sample mean, $\bar{x}$.

Sampling Distributions: Intro

- In the case of IPEDS data, we know the population mean value of “institutional SAT score”
  - Show population mean value in Stata
- Usually, we don’t know the population. Instead, we have samples.
  - So we know sample mean instead of population mean
- Usually, we only have one sample. But the sample we have is one out of many possible
- We use samples to make predictions about populations, but our predictions will differ from one sample to another

Sample Mean Changes from Sample to Sample

- Example:
  - Start with a population (institutional SAT score)
- Take a random sample of the population
  - Random sample has a probability distribution and a sample mean (show in Stata)
- Each time we take a random sample, get a different sample mean
- Imagine that we take 1,000 random samples
  - We would have 1,000 sample means
  - Consider each sample mean to be an observation
  - We would have a variable called “sample mean” with 1000 obs
  - We could plot this variable and we would get a distribution of sample means

Sampling Distribution

- A sampling distribution (of sample means) is a relative frequency distribution where each observation is a sample mean)
  - Imagine we draw n (e.g., 1000) random samples
  - For each sample, we record the sample mean
  - We create a frequency distribution of sample means
    - X-axis=value; Y= number of times a particular sample mean (e.g., $\bar{y} = 1050$ is observed)
- A sampling distribution can be created for any sample statistic (e.g., mean, median, a regression coefficient)

Sampling distribution pictures

- Show Applet:
  - Very useful web application
    - http://onlinestatbook.com/stat_sim/sampling_dist/index.html
  - Show applet for normal population distribution
  - Show applet for skewed population distribution

Sampling Distribution

- The sampling distribution shows how the value of a sample statistic varies from sample to sample
  - For example, each Presidential Election Poll represents a single sample mean from a single random sample
  - If values from each individual poll are close to one another, then we have more faith that the sample mean from one poll is close to population mean.
  - If values from each poll are far apart, then we wouldn’t put too much faith in the sample mean from any single poll

Sampling Distribution

- Consider the sample mean, $\bar{y}$, to be a variable, because its value varies from sample to sample
- If we take random samples the value of the each sample mean, $\bar{y}$, fluctuates around the population mean, μ
- If we took a large number of samples (e.g., 1,000) then the mean of all sample means, $\bar{y}_{\bar{y}}$, would be equal to the population mean, μ
  - $\bar{y}_{\bar{y}} = \mu$
  - Draw a picture of population distribution (label mean, std dev) over sampling distribution

Standard Error

- Standard deviation
  - Population standard deviation, σ, of a variable, y, is the average distance of an observation from the population mean, μ
- Standard error
  - Average distance of a single sample mean, $\bar{y}$, from the mean of the sample means, $\bar{y}_{\bar{y}}$
  - Standard error, $\sigma_{\bar{y}}$, is the standard deviation of the sampling distribution.
- Draw pictures:
  - population distribution (show mean, std dev); sampling distribution (show mean, std err)

Standard Error

- Standard error, $\sigma_{\bar{y}}$
  - Average distance of a single sample mean, $\bar{y}$, from the mean of the sample means, $\bar{y}_{\bar{y}}$
  - Same as, average distance of a single sample, $\bar{y}$, mean from the population mean, μ
  - $\sigma_{\bar{y}} = \frac{std\ dev}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$
- Ex: $\mu = 100;\ \sigma = 23; n = 100;\ \sigma_{\bar{y}} = \frac{23}{\sqrt{100}} = 2.3$
  - On average, each sample mean is 2.3 away from the population mean
- Note that standard error, $\sigma_{\bar{y}}$, is a population parameter because it depends on population standard deviation, σ
  - i.e., we usually don't know it; we will learn a sample version

Standard Error and election polls

- Why is standard error important?
  - Standard error tells us how much statistics derived from a sample are likely to diverge from population parameters
- Sample mean, $\bar{y}$, is best estimate of the population mean, μ, pct of people who will vote for Obama (e.g., $\bar{y} = 51\%$)
- Standard error provides an indication of how far away each sample mean is likely to be from the population mean
  - Standard error=10%: On average, the sample mean from each poll is likely to be 10% away from population mean
  - Standard error=2%: On average, the sample mean from each poll is likely to be 2% away from population mean
- Do we want standard error to be large or small? Why?

Properties of Standard Error

- $\sigma_{\bar{y}} = \frac{std\ dev}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}$
- Standard error decreases as size of your sample increases
  - If you have a large sample size, the sample mean is likely to be close to population mean.
    - When sample means are close to close to population mean, then standard error is small
    - Example: what is mean income in U.S.:
  - e.g., standard error smaller in 30 samples of sample size 500 compared to 30 samples of sample size 100
  - Show in Applet: http://onlinestatbook.com/stat_sim/sampling_dist/index.html

- Standard error increases when standard deviation, σ, increases (i.e., $\uparrow variability\ \rightarrow \uparrow std. error$)

Central Limit Theorem

- Sampling distribution
  - A sampling distribution of a the sample mean, $\bar{y}$, is the probability distribution associated with specific values of the sample mean.
  - It has a mean of μ and a standard error of $\sigma_{\bar{y}}$
- **Central Limit Theorem**
  - For random sampling with a large sample size n, the sampling distribution of the sample mean , $\bar{y}$, is approximately normally distributed
  - What is a "large" sample size
    - Agresti: n>=30 (approximately)
  - Restated: no matter what the distribution of the variable, the sampling distribution will have a normal distribution
- Show in Applet
  - http://onlinestatbook.com/stat_sim/sampling_dist/index.html
  - Change sample sizes; use skewed distribution

Shape of distribution: 3 distributions

- Three distributions
  - Population distribution
  - Sample data distribution
  - Sampling distributions
- Draw pictures for three types of variables (assume sample size > 30)
  - Normal distribution
  - Skewed distribution
  - A "proportion" variable (i.e., a 0/1 variable such as vote for Obama or Romney)
- What is shape of sampling distribution

In class exercise

- Play with the "applet"
  - http://onlinestatbook.com/stat_sim/sampling_dist/index.html
  - Note: "distribution of means" is sampling distribution
- (1) Choose "normal distribution";
  - click on "animated" several times to get several random samples; click on "5" to get five random samples at once; click on "1,000" to get 1,000 random samples
    - Watch how the sampling distribution changes as you add more sample means
- (2) Choose "skewed" distribution instead of "normal"
  - Repeat above exercises in (1)
- (3) Choose "skewed" distribution, select "N=25" instead of "N=5"
  - Repeat above exercises in (1)
  - How does shape of sampling distribution differ from (2)? What does this have to do with the central limit theorem

Khan Academy on Sampling Distributions

- http://www.khanacademy.org/math/statistics/v/sampling-distribution-of-the-sample-mean

Chapter 5
Statistical Inference: Estimation

Point and Interval Estimation

- Parameter
  - A summary of the population; usually unknown
- Estimates (sometimes called statistics)
  - A summary of the sample; used to make predictions about the population
  - Point estimate
    - A single number that is the best guess for the parameter (e.g., Obama approval = 46%)
  - Interval estimate
    - An interval around the point estimate, within which the parameter value is believed to fall (e.g., we are 95% sure that Obama's approval rating is between 44% and 48%)

Properties of good estimators

- Unbiased
  - An estimator is unbiased if its sampling distribution centers around the parameter
    - parameter=population mean μ;
    - Estimator= sample mean, $\bar{y}$.
    - sampling distribution= distribution of sample means, $\bar{y}$
  - If an estimator is unbiased, the mean of the sampling distribution equals the parameter value
    - e.g., population mean, μ, is equal to the mean of the sampling distribution, $\bar{y}_{\bar{y}}$
  - Show in Applet: http://onlinestatbook.com/stat_sim/sampling_dist/index.html

Properties of good estimators

- Biased estimator (not good)
  - A biased estimator tends to underestimate or overestimate the value of a parameter
  - Bias often occurs because of non-random sampling or non-random missing variables
    - If missing obs are like the rest of the population, then no bias; if missing obs tend to be different from the population, then there is bias.
    - This is where a lot of funny business happens!
  - Show in Stata
    - Show population distribution
    - Show parameter estimates based on biased sample

Properties of good estimators

- Efficient estimator
  - An efficient estimator is an estimator with a low standard error
  - The more efficient your estimator (lower standard error) the closer your estimates (e.g., sample mean $\bar{y}$) are likely to be to the parameter value (e.g., population mean μ)
  - Estimates become more precise
  - Show example in Applet
    - Remember that standard error is the standard deviation of the sampling distribution
    - Show SE with different sample sizes

Means vs. Proportion (for this book)

- Mean
  - Refers to a quantitative variable
- Proportion
  - Refers to a categorical variable with two categories

Means vs. Proportion

- Proportion: refers to a categorical variable
  - e.g., proportion of people who are married; proportion of Americans with baccalaureate
  - This book usually uses 0/1 variables when referring to proportions
    - Note that you can create a 0/1 variable from a variable with more than two categories; e.g., create a 0/1 variable called “bachelor” from an input variable called “highedu”

Means vs. Proportion

- Why have we been making all these 0/1 variables?
- Special properties of “proportion” variables
  - The relative frequency of observations that equal 1 is the same as the mean
  - Show examples in Stata:
    - When variable coded as 0/1
    - When variable coded as 2/1

Interval estimate

- Point estimate: single number that is best guess of parameter (e.g., sample mean)
- Interval estimate: interval of numbers around point estimate, within which the parameter is believed to fall (e.g., Confidence interval)
  - E.g., from a sample of 15 students, we are 95% sure that the average number of hours spent on statistics homework is between 2.5 and 3.4

Confidence intervals

- Very important for this entire course
- Method of teaching
  - Define confidence interval
  - Explain conceptually with pictures (most important)
  - Show how to calculate using formulas

Interval Estimate: Confidence Interval

- Confidence interval:
  - A confidence interval for a parameter is an interval of numbers within which parameter is believed to fall (e.g., we are 95% sure that Obama’s approval rating is between 44% and 48%)
  - Has the form: point estimate $\pm$ margin of error
  - The “confidence level” is the probability that the confidence interval contains the parameter.
  - Higher the confidence level, wider the confidence interval

Confidence intervals

- Confidence interval for a proportion (explain first)
  - $\pi = population\ proportion$
  - $\hat{\pi} = sample\ proportion$
- Confidence interval for a mean (explain second)
  - $\mu = population\ mean$
  - $\bar{y} = sample\ mean$

Stuff to skip

Basic Probability Rules

- P(A) means probability that event A occurs
  - Example: Event A= coin is “tails”; P(A) = 0.5
  - Example: Event A= dice roll is “1”; P(A)= 1/6
  - Example: A = graduate from college; P(A) = ?
  - Example: A = adopt an MBA program; P(A) = ?