STA 291
Lecture 16

- **Normal distributions: (mean and SD) use table or web page.**
- **The sampling distribution of $\hat{p}$ and $\overline{X}$ are both (approximately) normal**

- **Sampling Distributions**
  - **Sampling Distribution of $\overline{X}$**
  - **Sampling Distribution of $\hat{p}$**
- **Central limit theorem: no matter what the population look like, as long as we use SRS, and when sample size *n* is large, the above two sampling distribution are (very close to) normal.**

- $\hat{p}$ is approximately normal with
- mean = p, SD = $\frac{\sqrt{p(1-p)}}{\sqrt{n}}$
- $\overline{X}$ is approximately normal with
- mean = μ, SD = $\frac{s}{\sqrt{n}}$

Central Limit Theorem

- For random sampling, as the sample size *n* grows, the sampling distribution of the sample mean $\overline{X}$ approaches a normal distribution. So does the sample proportion $\hat{p}$
- Amazing: This is the case even if the population distribution is discrete or highly skewed
- The Central Limit Theorem can be proved mathematically
- We will verify it experimentally in the lab sessions

Central Limit Theorem

- Online applet 1

http://www.stat.sc.edu/~west/javahtml/CLT.html

- Online applet 2

http://bcs.whfreeman.com/scc/content/cat_040/spt/CLT-SampleMean.html

Population distribution vs. sampling distribution

- Population distribution: = distribution of X_1, a sample of size one from the population.

- In a simple random sample of size 4:

$$X_1, X_2, X_3, X_4$$

each one has the distribution of the population. But the average of the 4 has a different distribution --– the sampling distribution of mean, when *n=4.*

$$\overline{x} = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

- Has a distribution different from the population distribution:
  (1) shape is more normal
  (2) mean remains the same
  (3) SD is smaller (only half of the population SD)

Population Distribution

- Distribution from which we select the sample
- Unknown, we want to make inference about its parameters
- Mean = ?
- Standard Deviation = ?

Sample Statistic

- From the sample $X_1, \ldots, X_n$ we compute descriptive statistics

- Sample Mean =
- Sample Standard Deviation =
- Sample Proportion =

They all can be computed given a sample.

Sampling Distribution of a sample statistic

- Probability distribution of a statistic (for example, the sample mean)
- Describes the pattern that would occur if we could repeatedly take random samples and calculate the statistic as often as we wanted
- Used to determine the probability that a statistic falls within a certain distance of the population parameter
- The mean of the sampling distribution of $\overline{X}$ is =
- The SD of $\overline{X}$ is also called Standard Error =

- The 3 features for the sampling distribution of *sample mean* also apply to *sample proportion. (1. approach normal, 2. centered at p; 3. less and less SD)*
- *This sampling distribution tells us how far or how close between "p" and* $\hat{p}$
- *One quantity we can compute, the other we want to know*

Central Limit Theorem

- For example:

  If the sample size is $n = 100$, then the sampling distribution of $\hat{p}$ has mean p and SD (or standard error) =

$$\frac{\sqrt{p(1-p)}}{\sqrt{100}} = \frac{\sqrt{p(1-p)}}{10}$$

Preview of estimation of p

- Estimation with error bound: Suppose we counted 57 "YES" in 100 interview. (SRS)
- Since we know just how far $\hat{p}$ is to p. (that is given by the sampling distribution)
- 95% of the time $\hat{p}$ is going to fall within two SD of p.
- SD = ?

- 57/100 = 0.57 = $\hat{p}$
- Sqrt of [0.57(1-0.57)] = 0.495
- $\frac{\sqrt{p(1-p)}}{10}$ = 0.495/10 = 0.0495

- Finally, with 95% probability, the difference between p and $\hat{p}$ is within 2 SD or
- 2(0.0495) = 0.099

Multiple choice question

The standard error of a statistic describes

1. The standard deviation of the sampling distribution of that statistic
2. The variability in the values of the statistic for repeated random samples of size *n.*

Both are true

Multiple Choice Question

The Central Limit Theorem implies that

1. All variables have approximately bell-shaped sample distributions if a random sample contains at least 30 observations
2. Population distributions are normal whenever the population size is large
3. For large random samples, the sampling distribution of the sample mean (X-bar) is approximately normal, regardless of the shape of the population distribution
4. The sampling distribution looks more like the population distribution as the sample size increases

- In previous page, 3 is correct.

Chapter 10

- Statistical Inference: Estimation of p
  - Inferential statistical methods provide predictions about characteristics of a population, based on information in a sample from that population
  - For quantitative variables, we usually estimate the population mean (for example, mean household income)
  - For qualitative variables, we usually estimate population proportions (for example, proportion of people voting for candidate A)

Two Types of Estimators

- **Point Estimate**
  - A single number that is the best guess for the parameter
  - For example, the sample mean is usually a good guess for the population mean
- **Interval Estimate (harder)**

  =point estimator with error bound
  - A range of numbers around the point estimate
  - To give an idea about the precision of the estimator
  - For example, "the proportion of people voting for A is between 67% and 73%"

Point Estimator

- A point estimator of a parameter is a sample statistic that predicts the value of that parameter
- A good estimator is
  - ***Unbiased***: Centered around the true parameter
  - ***Consistent***: Gets closer to the true parameter as the sample size gets larger
  - ***Efficient***: Has a standard error that is as small as possible (made use of all available information)

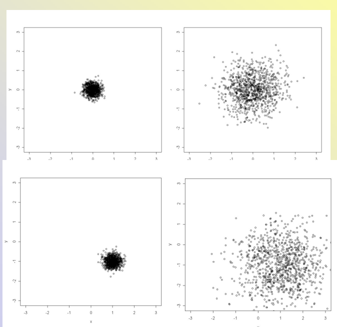
Efficiency

- An estimator is efficient if its standard error is small compared to other estimators
- Such an estimator has high precision
- A good estimator has ***small standard error and small bias*** (or no bias at all)

- The following pictures represent different estimators with different bias and efficiency
- Assume that the true population parameter is the point (0,0) in the middle of the picture

STA 291 - Lecture 16 22

Bias and Efficiency



Note that even an unbiased and efficient estimator does not always hit exactly the population parameter.

But in the long run, it is the best estimator.

23

- Sample proportion is an unbiased estimator of the population proportion.
- It is consistent and efficient.

STA 291 - Lecture 16 24

Example: Estimators

- Suppose we want to estimate the proportion of UK students voting for candidate A
- We take a random sample of size $n=400$
- The sample is denoted $X_1, X_2, \ldots, X_n$, where $X_i=1$ if the ith student in the sample votes for A, $X_i=0$ otherwise

Example: Estimators

- Estimator1 = the sample proportion
- Estimator2 = the answer from the first student in the sample (X_1)
- Estimator3 = 0.3
- Which estimator is unbiased?
- Which estimator is consistent?
- Which estimator has high precision (small standard error)?

Attendance Survey Question

- On a 4"x6" index card
  - Please write down your name and section number
  - Today's Question:
  - Table or web page for Normal distribution? Which one you like better?

Central Limit Theorem

- Usually, the sampling distribution of $\bar{X}$ is approximately normal for n = 30 or above
- In addition, we know that the parameters of the sampling distribution are "μ" and $s_{\bar{X}} = \frac{s}{\sqrt{n}}$
- For example:
  If the sample size is n=49, then the sampling distribution of $\bar{X}$ has mean μ and SD (or standard error) = $\frac{s}{\sqrt{49}} = \frac{s}{7}$

Cont.

Using the "empirical rule" with 95% probability $\bar{X}$ will fall within 2 SD of its center, mu.

(since the sampling distribution is approx. normal, so empirical rule apply. **In fact, 2 SD should be refined to 1.96 SD)**

- with 95% probability, the $\bar{X}$ falls between

$$m - 1.96\frac{s}{\sqrt{n}} = m - \frac{1.96}{7}s = m - 0.28s$$

$$\text{and } m + 1.96\frac{s}{\sqrt{n}} = m + \frac{1.96}{7}s = m + 0.28s$$

(m = population mean, s = population standard deviation)

Unbiased

- An estimator is unbiased if its sampling distribution is centered around the true parameter
- For example, we know that the mean of the sampling distribution of "X-bar" equals "mu", which is the true population mean
- So, "X-bar" is an unbiased estimator of "mu"

Unbiased

- However, for any particular sample, the sample mean "X-bar" may be smaller or greater than the population mean
- "Unbiased" means that there is no systematic under- or overestimation

Biased

- A biased estimator systematically under- or overestimates the population parameter
- In the definition of sample variance and sample standard deviation uses *n-1* instead of *n*, because this makes the estimator unbiased
- With *n* in the denominator, it would systematically underestimate the variance

Point Estimators of the Mean and Standard Deviation

- The sample mean is unbiased, consistent, and (often) relatively efficient for estimating "mu"
- The sample standard deviation is *almost* unbiased for estimating population SD (no easy unbiased estimator exist)
- Both are consistent