STA 291
Lecture 17

- **Chap. 10 Estimation**
  - **Estimating the Population Proportion p**

  - **We are not predicting the next outcome (which is random), but is estimating a fixed number --- the population parameter.**

Review: Population Distribution, and Sampling Distribution

- Population Distribution
  - Unknown, distribution from which we select the sample
  - Want to make inference about its parameter , like p
- Sampling Distribution
  - Probability distribution of a statistic (for example, sample mean/proportion)
  - Used to determine the probability that a statistic falls within a certain distance of the population parameter
  - For large *n*, the sampling distribution of the sample mean/proportion looks more and more like a normal distribution

Chapter 10

- Estimation: Confidence interval
  - Inferential statistical methods provide estimates about characteristics of a population, based on information in a sample drawn from that population
  - For quantitative variables, we usually estimate the population mean (for example, mean household income) + (SD)
  - 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 (unknown) parameter
  - For example, the sample proportion/mean is usually a good guess for the population proportion/mean
- Interval Estimate
  - 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 estimates 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 n gets larger
  - ***Efficient***: Has a standard error that is as small as possible

- Sample proportion, $\hat{p}$, is unbiased as an estimator of the population proportion p.

It is also consistent and efficient.

New: Confidence Interval

- An inferential statement about a parameter should always provide the accuracy of the estimate (error bound)
- How close is the estimate likely to fall to the true parameter value?
- Within 1 unit? 2 units? 10 units?
- This can be determined using the sampling distribution of the estimator/sample statistic
- In particular, we need the standard error to make a statement about accuracy of the estimator

- How close?
- How likely?

New: Confidence Interval

- **Example**: interview 1023 persons, selected by SRS from the entire USA population.

- Out of the 1023 only 153 say "YES" to the question "economic condition in US is getting better"

- Sample size n = 1023, $\hat{p}$ = 153/1023=0.15

- The sampling distribution of $\hat{p}$ is (very close to) normal, since we used SRS in selection of people to interview, and 1023 is large enough.

- The sampling distribution has mean = p , and
- SD = $\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} = \dfrac{\sqrt{0.15(1-0.15)}}{\sqrt{1023}} = 0.011$

- The 95% confidence interval for the unknown p is

- [0.15 – 2x0.011, 0.15 + 2x0.011]

- Or [0.128, 0.172]

- Or "15% with 95% margin of error 2.2%"

Confidence Interval

- A confidence interval for a parameter is a range of numbers that is likely to cover (or capture) the true parameter.
- The probability that the confidence interval captures the true parameter is called the confidence coefficient/confidence level.
- The confidence level is a chosen number close to 1, usually 95%, 90% or 99%

Confidence Interval

- To calculate the confidence interval, we used the Central Limit Theorem
- Therefore, we need sample sizes of at least moderately large, usually we require both *np > 10 and n(1-p) > 10*

- Also, we need a $z_{a/2}$ that is determined by the confidence level
- Let's choose confidence level 0.95, then $z_{a/2}$ *=1.96 (the refined version of "2")*

- confidence level 0.90, ←→ $z_{a/2}$ =1.645
- confidence level 0.95 ←→ $z_{a/2}$ =1.96
- confidence level 0.99 ←→ $z_{a/2}$ =2.575

Confidence Interval

- So, the ***random*** interval between

$$\hat{p} - 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \quad \text{and} \quad \hat{p} + 1.96\frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

Will capture the population proportion p with 95% probability

- This is a confidence statement, and the interval is called a **95% confidence interval**

Confidence Interval: Interpretation

- “Probability” means that “in the long run, 95% of these intervals would contain the parameter”
- If we repeatedly took random samples using the same method, then, in the long run, in 95% of the cases, the confidence interval will cover the true unknown parameter
- For one given sample, we do not know whether the confidence interval covers the true parameter or not.
- The ***95% probability*** only refers to the ***method*** that we use, but not to this individual sample

Confidence Interval: Interpretation

- To avoid the misleading word "probability", we say:

  "We are 95% confident that the true population p is within this interval"

- Wrong statements:
- 95% of the p's are going to be within 12.8% and 17.2%

Statements

- 15% of all US population thought “YES”.
- It is probably true that 15% of US population thought “YES”
- We do not know exactly, but we know it is between 12.8% and 17.2%

- We do not know, but it is probably within 12.8% and 17.2%

- We are 95% confident that the true proportion (of the US population thought “YES”) is between 12.8% and 17.2%

- You are never 100% sure, but 95% or 99% sureis quite close.

Confidence Interval

- If we change the confidence level from 0.95 to 0.99, the confidence interval changes
- Increasing the probability that the interval contains the true parameter requires increasing the length of the interval
- In order to achieve 100% probability to cover the true parameter, we would have to increase the length of the interval to infinite -- that would not be informative
- There is a tradeoff between length of confidence interval and coverage probability. Ideally, we want short length and high coverage probability (high confidence level).

Confidence Interval

- Confidence Interval Applet

- http://bcs.whfreeman.com/scc/content/cat_040/spt/confidence/confidenceinterval.html

Simpson's Paradox

- Suppose five men and five women apply to two different departments in a graduate school.

	Men	Women
Arts	3 out of 4 accepted (**75%**)	1 out of 1 accepted (**100%)**
Science	0 out of 1 accepted (**0%**)	1 out of 4 accepted (**25%**)
Totals	3 out of 5 accepted (**60%**)	2 out of 5 accepted (**40%**)

Although each department separately has a higher acceptance rate for women, the combined acceptance rate for men is much higher.

Success rate

- Another example: hospital A is more expensive, with the state of the art facility. Hospital B is cheaper.

	Hospital A	hospital B
Ease case	99 out of 100	490 out of 500
Trouble case	150 out of 200	30 out of 60

Hospital A is doing better in each category, but overall worse.

- The reason is that hospital A got the most trouble cases while hospital B got mostly the ease cases.

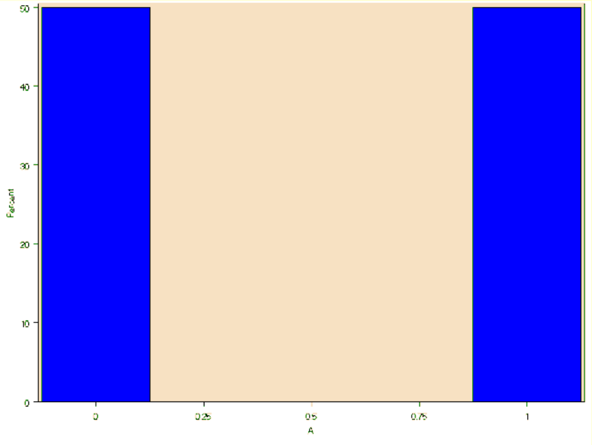
- Since hospital B is cheaper, when every indication of an ease case, people go to hospital B.

Attendance Survey Question 17

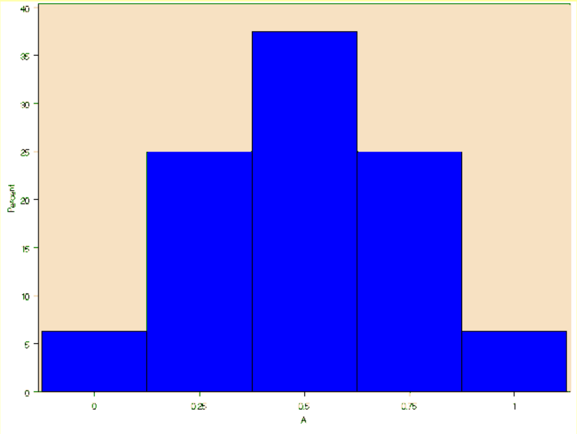
- On a 4”x6” index card
  - Please write down your name and section number
  - Today’s Question:

  - Can a hospital do better in both sub-categories, but overall do worse than a competitor?
  - Enjoy Spring Break

Review: Population, Sample, and Sampling Distribution



Population distribution



Sampling Distribution for the sample mean for *n*=4

•The population distribution is P(0)=0.5, P(1)=0.5.

•The sampling distribution for the sample mean in a sample of size *n*=4 takes the values 0, 0.25, 0.5, 0.75, 1 with different probabilities.

Example: Three Estimators

- Suppose we want to estimate the proportion of UK students voting for candidate A
- We take a random sample of size $n=100$
- The sample is denoted $X_1, X_2, \ldots, X_n$, where $X_i=1$ if the *i*th student in the sample votes for A, $X_i=0$ otherwise

Example: Three Estimators

- Estimator 1 = the sample mean (sample proportion)
- Estimator 2 = the answer from the first student in the sample (X_1)
- Estimator 3 = 0.3
- Which estimator is unbiased?
- Which estimator is consistent?
- Which estimator has high precision (small standard error)?

Point Estimators of the Mean and Standard Deviation

- The sample mean, X-bar, is unbiased, consistent, and often (relatively) efficient

- The sample standard deviation is slightly biased for estimating population SD
- It is also consistent (and often relatively efficient)

Estimation of SD

- Why not use an unbiased estimator of SD?
- We do not know how to find one……

- The sample variance (one divide with n-1) is unbiased for population Variance though……

- The square root destroy the unbias.
- The bias is usually small, and goes to zero (become unbiased) as sample size grows.

Confidence Interval

- Example from last lecture (application of central limit theorem):
- If the sample size is n=49, then with 95% probability, the sample mean falls between

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

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

$$(m = \text{population mean, } s = \text{population standard deviation})$$

Confidence Interval, again

- With 95% probability, the following interval will contain sample mean, X-bar

$$m - 1.96\frac{s}{\sqrt{n}} \text{ and } m + 1.96\frac{s}{\sqrt{n}}$$

$$(m = \text{population mean, } s = \text{population standard deviation})$$

- Whenever the sample mean falls within this interval, the distance between X-bar and mu is less than

$$1.96\frac{s}{\sqrt{n}}$$

When sigma is known

- Another way to say this: with 95% probability, the distance between mu and X-bar is less than 0.28sigma

- If we use X-bar to estimate mu, the error is going to be less than 0.28sigma with 95% probability.

Confidence Interval

- The sampling distribution of the sample mean $\bar{X}$ has mean m and standard error

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

- If *n* is large enough, then the sampling distribution of $\bar{X}$ is approximately normal/bell-shaped

(Central Limit Theorem)