Chapter 7—Estimates and Sample Sizes

Example: Public opinion polls provide a common source of data that are analyzed as proportions.

CBS NEWS POLL
The New York Times

CBS NEWS/NEW YORK TIMES POLL
For release: Tuesday July 15, 2008
6:30 P.M. EDT

THE TOP ISSUES: IMPROVEMENT IN IRAQ, ECONOMIC CONCERNS
July 7 - 14, 2008

While Americans remain pessimistic about the direction of the country in general and the state of the economy in particular, they are increasingly positive about the way things are going in Iraq.

Only 7% of Americans would say the war in Iraq is going *very* well, but 45% say it is going at least *somewhat* well. This marks the most positive assessment of the war since January, 2006, and a 10 point upswing since just last month.

HOW IS THE WAR IN IRAQ GOING?

	Now	6/2008	6/2007	1/2006
Well	45%	35%	22%	49%
Badly	51	62	77	49

However, 51% say the war is going at least somewhat badly. And six in 10 Americans continue to believe the United States should never have gotten involved in Iraq in the first place.

DID THE U.S. DO THE RIGHT THING GOING TO WAR WITH IRAQ?

	Now	4/2008
Right thing	36%	37%
Should have stayed out	59	57

ENERGY COSTS, NUCLEAR POWER AND THE 2008 ELECTION

One solution advocated by John McCain for energy problems - building more nuclear power plants to generate electricity - now gets new consideration from many Americans. 57% think more nuclear power plants should be built - the highest number since 1977, before the nuclear accidents at the Three Mile Island power plant in Pennsylvania and the disaster at Chernobyl. Americans were divided just a year ago.

DO YOU APPROVE OF BUILDING MORE NUCLEAR POWER PLANTS?

	Now	4/2007	6/2001	5/1986	7/1977
Yes	57%	45%	51%	34%	69%
No	34	47	42	59	21

But most voters do not believe the policies of the next president - whether it is Barack Obama or John McCain - will help bring

down gas prices. 30% think Obama's energy policies will help bring down the price of gas, and 25% think John McCain's will.

WILL THE CANDIDATES' ENERGY POLICIES BRING DOWN GAS PRICES?
(Among registered voters)

	Obama	McCain
Yes	30%	25%
No	50	56
Don't Know	20	19

Americans continue to feel the pinch of gas prices that average more than four dollars a gallon across the country. 65% of Americans say that gas prices have caused financial hardship for them or their households and more than a third characterize this hardship as serious.

HAVE GAS PRICES CAUSED FINANCIAL HARDSHIP IN YOUR HOUSEHOLD?

	Now	6/2008
Yes, serious	34%	36%
Yes, not serious	31	29
No	35	34

THE ECONOMY AND THE BUSH ADMINISTRATION

Overall, the economy and jobs remain the most important problem facing the country today, with the more specific issue of the cost of gas tied with the war in Iraq in second place. 39% of Americans now count the economy and jobs together as the country's most important problem, up five points from last month.

MOST IMPORTANT PROBLEM

	Now	6/2008
Economy & Jobs	39%	34%
Gas & Oil Prices	14	16
War in Iraq	14	15

When rating the national economy, only 19% say the condition of the economy is even somewhat good, slightly above the all-time low of 16% reached in May. Eight in 10 Americans rate the condition of the economy as bad, including 35% who say it is very bad. A year ago a majority of Americans said the condition of the economy was at least somewhat good.

CONDITION OF THE ECONOMY

	Now	6/2008	5/2008	7/2007
Good	19%	20%	16%	55%
Bad	80	78	83	43

Looking ahead, more than two thirds of all Americans think the economy is getting worse, while only 3% say it is getting better. 29% say the economy is staying the same.

IS THE ECONOMY…?

	Now	6/2008
Getting better	3%	3%
Getting worse	67	69
Staying the same	29	27

However, Americans remain positive about their own financial situation. While only 13% say their household finances are very good, nearly three in four say it is at least somewhat good.

Bush continues to receive dismally low marks on his handling of the economy. Now only 20% approve of the job he is doing with this issue - a new low. His overall job approval rating is 28%, up slightly from an all-time low of 25% reached last month.

BUSH'S JOB APPROVAL

	Now	6/2008
Overall	28%	25%
Economy	20%	21%

When looking at the direction of the country as a whole, eight in 10 say the country is off on the wrong track. Only 14% of Americans think the country is headed in the right direction, unchanged from last month's all-time low.

DIRECTION OF THE COUNTRY

	Now	6/2008
Right direction	14%	14%
Wrong track	81	83

This poll was conducted among a random sample of 1796 adults nationwide interviewed by telephone July 7-14, 2008. Oversamples of African Americans and Hispanics were interviewed, for a total of 297 interviews with African Americans and 246 interviews with Hispanics.

Respondents could be interviewed in either English or Spanish, and the poll included a sample of cell phones. The results then weighted in proportion to the racial composition of the adult population in the U.S. Census.

The error due to sampling could be plus or minus three percentage points based on the entire sample. The sampling errors could be plus or minus six percentage points for the African American and Hispanic samples.

The true population proportion of adults nationwide who believe that Obama's energy policies will bring down gas prices, designated as p, is estimated by

$$\hat{p} = 0.30 \approx \frac{702}{2339}$$

How reliable is the estimator $\hat{p}$? Can we characterize the reliability (or uncertainty)?

The true population proportion of adults nationwide who feel that gas prices have caused a financial hardship in their household in July 08 versus June 08, designated as p_{Jul} and p_{Jun}, respectively, are estimated by

$$\hat{p}_{\text{Jul}} = 0.34 \text{ and } \hat{p}_{\text{Jun}} = 0.36$$

Are the true values of p_{Jul} and p_{Jun} the same or are they different?

We need to know the sampling distribution of $\hat{p}$---if we were to draw samples of 2,339 people over and over, and each time calculate a new value of $\hat{p}$, what would the distribution look like?

Solution: view $\hat{p}$ as the mean number of successes per trial over the n trials:

1. Assign a success a value of 1 and a failure a value of 0
2. Define X to be the sum of all n sample observations
3. $\hat{p} = \frac{X}{n}$ is the mean number of successes in n trials

From section 6-6, if X is a binomial RV, as $n \to \infty$ X will be approximately normally distributed with $\mu = np$ and $\sigma = \sqrt{npq}$

Rule of thumb for assuming approximate normality is: $np \geq 5$ and $nq \geq 5$; or $n\hat{p} \geq 5$ and $n\hat{q} \geq 5$

What can we say about $\hat{p} = \frac{X}{n}$?

$\hat{p} = \frac{X}{n}$ will also be approximately normally distributed for large n since dividing by n does not change the shape of the distribution.

$$\hat{p} = \frac{X}{n} = \frac{1}{n}X$$

$$E(\hat{p}) = \frac{1}{n}E(X) = \frac{1}{n}np = p$$

$$\sigma^2(\hat{p}) = Var\left(\frac{1}{n}X\right) = \left(\frac{1}{n}\right)^2 Var(X)$$

$$= \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

$$\sigma(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

Properties of the Sampling Distribution of $\hat{p}$

1. Mean of sampling distribution equals mean of sampled population

$$\mu_{\hat{p}} = E(\hat{p}) = p$$

2. Standard deviation of sampling distribution equals

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

3. ***Standard error*** of sampling distribution equals

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

1. Estimating a population proportion

Requires

- SRS—reality will more likely be that we have some sort of random sample and most importantly the sample is representative of the population of interest
- Conditions for a binomial distribution are satisfied
- Conditions for normal approximation are satisfied

Define

$p =$ population proportion

$\hat{p} = \frac{x}{n} =$ sample proportion of x successes in a sample of size n

$\hat{q} = 1 - \hat{p}$

2. Point Estimate

A single value (statistic) used to estimate a population parameter.

3. Interval Estimate (confidence interval)

- A range of values used to estimate the true value of a population parameter.
- Abbreviated as CI
- Associated with a confidence level
- Confidence level defined as $1-\alpha$, where $1-\alpha$ is the proportion of times the CI does contain the population parameter—assuming that the estimation process is repeated a large number of times
- Also called degree of confidence or confidence coefficient
- What is alpha?—complement of the CL

Common choices:

- 0.90 with $\alpha = 0.10$
- 0.95 with $\alpha = 0.05$
- 0.99 with $\alpha = 0.01$

Often expressed as a percentage

4. Critical values

The number on the borderline separating sample statistics that are likely to occur from those that are unlikely to occur—a cutoff value

The number $z_{\alpha/2}$--positive z value that is the vertical boundary separating an area $\alpha/2$ In the right tail of the standard normal distribution. Also have $-z_{\alpha/2}$.

Example standard normal curve

5. Margin of Error

Denoted E—with probability $1-\alpha$, the maximum likely difference between the observed sample proportion $\hat{p}$ and the true value of the population proportion p

$$E = z_{\alpha/2}\left(\text{SE}_{\hat{p}}\right) = z_{\alpha/2}\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

Use E to construct a CI:

$$\hat{p} \pm E$$
$$\hat{p} - E < p < \hat{p} + E$$
$$\left(\hat{p} - E, \hat{p} + E\right)$$

6. General interval estimate

A general format for all confidence intervals is:

point estimator ± E ➔
point estimate ± (critical value × SE(point estimate))

Example 1

7. Interpreting a CI

Must be careful. We are 95% confident that the true population proportion falls in …

Not: there is a 95% probability that the true population proportion falls between…

Not: there is a 95% chance that the true population proportion falls between…

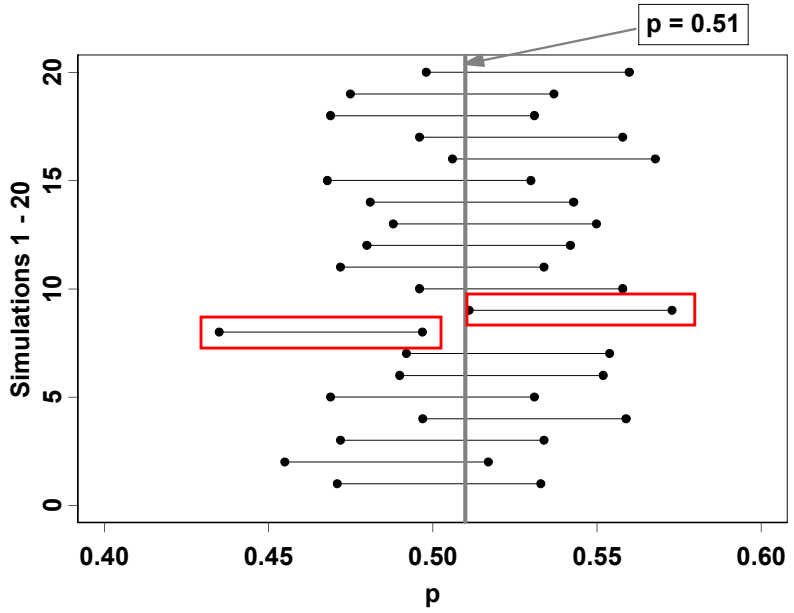
After you have used sample data to construct the interval, the interval either contains the truth or it does not. Confidence comes from the method/procedure.

Let's look at a simulation

1. Selected 100,000 independent trials from a binomial distribution with $n = 1$ and $p = 0.51$— this became the population
2. Randomly sampled $n = 1000$ observations from the "population"
3. Calculated $\hat{p}$ and SE($\hat{p}$)
4. Created 95% CI for p
5. Repeated steps 1 – 4 19 more times and plotted results

Drew n = 1000 from a population with p = 0.51

	lower.CL	phat	upper.CL
[1,]	0.4710099	0.502	0.5329901
[2,]	0.4550218	0.486	0.5169782
[3,]	0.4720102	0.503	0.5339898
[4,]	0.4970583	0.528	0.5589417
[5,]	0.4690097	0.500	0.5309903
[6,]	0.4900370	0.521	0.5519630
[7,]	0.4920425	0.523	0.5539575
[8,]	0.4350814	0.466	0.4969186
[9,]	0.5111192	0.542	0.5728808
[10,]	0.4960549	0.527	0.5579451
[11,]	0.4720102	0.503	0.5339898
[12,]	0.4800172	0.511	0.5419828
[13,]	0.4880321	0.519	0.5499679
[14,]	0.4810186	0.512	0.5429814
[15,]	0.4680097	0.499	0.5299903
[16,]	0.5060946	0.537	0.5679054
[17,]	0.4960549	0.527	0.5579451
[18,]	0.4690097	0.500	0.5309903
[19,]	0.4750119	0.506	0.5369881
[20,]	0.4980618	0.529	0.5599382



8. Determining sample size

We can use the expression for margin of error to help determine how large of a sample (minimum sample size) we need to obtain an estimate of p with a particular level of confidence.

$$E = z_{\alpha/2}\left(\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) \Rightarrow E^2 = \left(z_{\alpha/2}\right)^2\left(\frac{\hat{p}\hat{q}}{n}\right) \Rightarrow n = \frac{z_{\alpha/2}{}^2\left(\hat{p}\hat{q}\right)}{E^2}$$

If we are in the process of designing a study, where do we get an estimate of $\hat{p}$

- From a pilot study
- From the literature
- From an expert
- When in doubt, use $\hat{p} = 0.5$--the variance for a binomial random variable is maximum when $p = 0.5$

Example 2

Practice 1

The genetics and IVF Institute conducted a clinical trial of the YSORT method designed to increase the probability of conceiving a boy. During the study, 51 babies were born to parents using the YSORT method, and 39 of them were boys. Use the sample data to construct a 99% CI estimate of the percentage of boys born to parents using the YSORT method. ***Interpret the interval.***

Based on the result, does the YSORT method appear to be effective? Why or why not?

Practice 2

The music industry must adjust to the growing practice of consumers downloading songs instead of buying CDs. It therefore becomes important to estimate the proportion of songs that are currently downloaded. How many randomly selected song purchases must be surveyed to determine the percentage that was obtained by downloading? Assume that we want to be 95% confident that the sample percentage is within one percentage point of the true population percentage of songs that are downloaded.