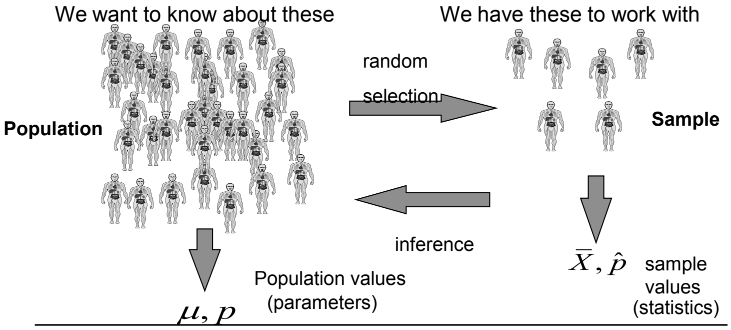
Stat 104: Quantitative Methods

Class 22: Confidence Intervals for Proportions

Review:



We want to know about these

We have these to work with

random selection

Population

Sample

inference

Population values (parameters)

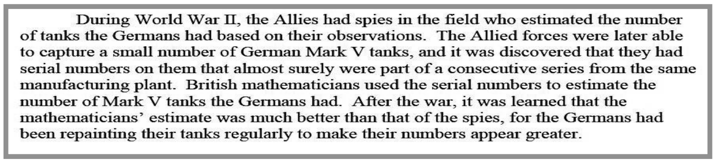
μ, p

$\bar{X}, \hat{p}$ sample values (statistics)

Unusual Estimators

- We are used to the sample mean and standard deviation as estimators.
- Another example of a an unusual estimator comes from World War II.
- Its called the German Tank Problem.

The German Tank Problem



During World War II, the Allies had spies in the field who estimated the number of tanks the Germans had based on their observations. The Allied forces were later able to capture a small number of German Mark V tanks, and it was discovered that they had serial numbers on them that almost surely were part of a consecutive series from the same manufacturing plant. British mathematicians used the serial numbers to estimate the number of Mark V tanks the Germans had. After the war, it was learned that the mathematicians' estimate was much better than that of the spies, for the Germans had been repainting their tanks regularly to make their numbers appear greater.

The problem set up

- Suppose there is a population of integers 1,2,3,….N
- You want to determine N
- You are given sample of *n* values from this population.
- How would you estimate N?
- Clearly your estimate should give you a number greater or equal to the max tank number in the sample.

Some R Information

- R can be extended via packages and there are two useful for computing confidence intervals.
- install.packages(“BSDA")
- install.packages(“binom")

IF you have data-the t.test command

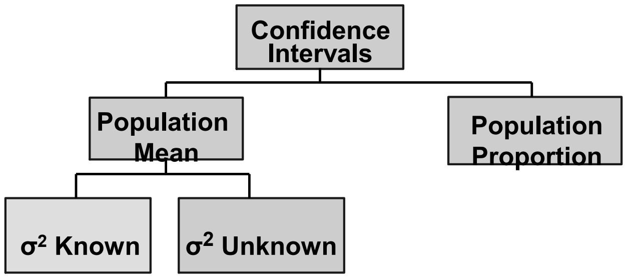
```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/stat111_survey.csv")
➢ sleep=mydata$sleep

> t.test(sleep)

        One Sample t-test

data:  sleep
t = 12.248, df = 89, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 6.669596 9.252626
sample estimates:
mean of x
 7.961111
```

IF you have summary statistics-tsum.test command

Suppose in a survey of 237 American first-graders it was found they spent on average 14 minutes per day on homework with a standard deviation of 4 minutes.. Construct a 95% confidence interval

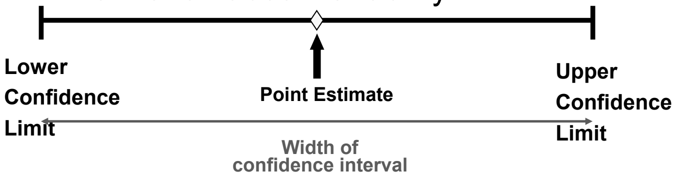
```
> library(BSDA)

> tsum.test(n.x=237,mean.x=14,s.x=4)

        One-sample t-Test

data:  Summarized x
t = 53.882, df = 236, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 13.48812 14.51188
```

Confidence Intervals



Recall Point and Interval Estimates

- A point estimate is a single number,
- a confidence interval provides additional information about variability



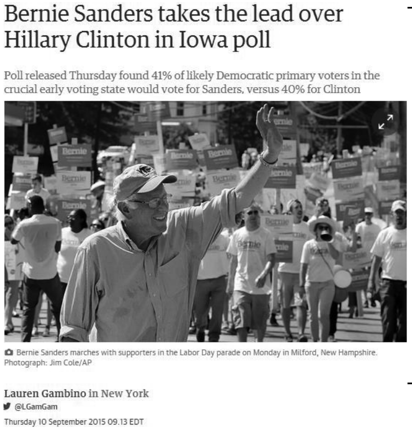
Confidence Interval for a Proportion

- The Gallup Poll periodically asks a random sample of U.S. adults whether they think economic conditions are getting better, getting worse, or staying about the same.
- When they polled 2976 respondents in March 2010, only 1012 thought economic conditions in the United States were getting better—a sample proportion of 1012/2976=34%. How wrong is this guess?

Can We Trust Opinion Polls? The Central Limit Theorem, Binomial Proportion Confidence Intervals, and Likely Voters

Ever since the 1824 straw poll of the U.S. presidential election showing Andrew Jackson leading over John Quincy Adams, opinion polls has become more and more popular. In 1936 George Gallup introduced the concept of *representative sampling*, which means that the people asked should be a mirror image of the population under scrutiny.

However, even if we draw a random sample of the population we cannot be certain that it is a true mirror image. The numbers we get from opinion polls are associated with statistical uncertainty. To understand the logic of this uncertainty we need to get acquainted with what is known as the *central limit theorem*. Let us say that we conduct a survey



Bernie Sanders takes the lead over Hillary Clinton in Iowa poll

Poll released Thursday found 41% of likely Democratic primary voters in the crucial early voting state would vote for Sanders, versus 40% for Clinton

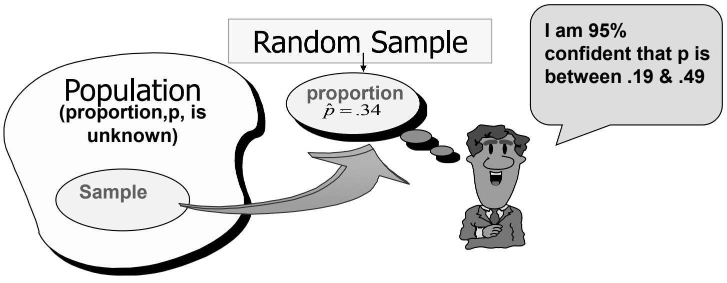
Bernie Sanders marches with supporters in the Labor Day parade on Monday in Milford, New Hampshire. Photograph: Jim Cole/AP

Lauren Gambino in New York
@LGamGam
Thursday 10 September 2015 09.13 EDT

In actuality, Bernie didn't exactly take the lead over Hillary Clinton. Instead, a Quinnipiac poll showed that 41% of likely Democratic primary voters in Iowa indicated that they would vote for Sanders, while 40% reported that they would vote for Clinton.

If you go to the original Quinnipiac poll, you can read that the actual data has a margin of error of +/- 3.4%, which means that the candidates are running neck and neck. Which, I think, would have still been a compelling headline.

The General Estimation Process



Random Sample

Population (proportion,p, is unknown)

Sample

proportion $\hat{p} = .34$

I am 95% confident that p is between .19 & .49

Point Estimate for Population p

Estimate Population Parameter…	with Sample Statistic
Proportion: p	$\hat{p}$

Point Estimate for p

The proportion of successes in a sample.

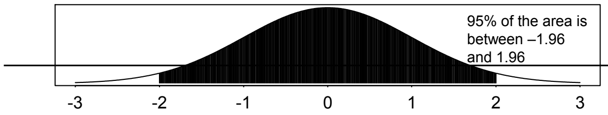
- Denoted by
  - ❑ $\hat{p} = \frac{x}{n} = \frac{\text{number of successes in sample}}{\text{sample size}}$
  - ❑ read as "*p* hat"

Review – standardization rule

$$If \quad X \sim N(Mean, Variance) \quad then$$

$$Z = \frac{X - Mean}{\sqrt{Variance}} \sim N(0,1)$$

$$For\ Z \sim N(0,1),$$
$$P(-1 \le Z \le 1) = 0.68$$
$$P(-1.96 \le Z \le 1.96) = 0.95$$
$$P(-3 \le Z \le 3) = 0.997$$



We construct the interval estimate using the CLT.

Recall that the CLT says that (for large *n*)

$$\boxed{\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)}$$

Then by the **Standardization Rule**:

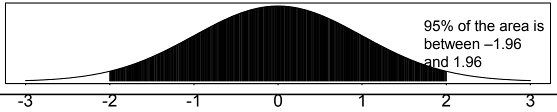
$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

We have

$$\frac{\hat{P} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

Then from what we know about the standard normal distribution:

$$P(-1.96 < \frac{\hat{P} - p}{\sqrt{p(1-p)/n}} < 1.96) = 95\%$$



We have

$$P(-1.96 < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < 1.96) = 95\%$$

By doing some **algebra**, we may rearrange stuff so that

$$P(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}) = 95\%$$

lower bound truth upper bound

The interval is random, p is not random.

Conclusion

- We are thus 95% confident that the true population proportion is in the interval

$$\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}, \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right)$$

- We are **assuming** that n is large, $n\hat{p} > 15$ $n\hat{q} > 15$ (at least 15 successes and 15 failures) and our sample size is less than 10% of the population size.

Wait!

- We can't use this formula as is…why??

$$\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}}, \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right)$$

- So the confidence interval formula is

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

Back to Gallup Example

- We have *n* =2976 and $\hat{p}$=34%
- The confidence interval for the true proportion of people who think the economy is improving is

$$0.34 \pm 1.96\sqrt{\frac{0.34(1-0.34)}{2976}} = (0.322, 0.357)$$

Example: A marketing research firm contacts a random sample of 100 men in Chicago and finds that 40% of them prefer the Gillette Sensor razor to all other brands. The 95% C.I. for the proportion of all men in Chicago who prefer the Gillette Sensor is determined as follows:

$$0.40 \pm 1.96\sqrt{\frac{0.40(1-0.40)}{100}} = 0.40 \pm 1.96(0.05) = (0.30398, 0.49602)$$

So with 95% confidence, we estimate the proportion of all men in Chicago who prefer the Gillette Sensor to be somewhere between 30 and 50 percent (pretty good market share).

Calculating the CI in R

- There are actually several ways to calculate a confidence interval for a proportion (more details in a few slides).
- This interval is called the Wald or asymptotic interval.
- This interval is taught in every intro stat class and more or less sucks.

Many Options for Proportions

```
> library(binom)
> binom.confint(40,100)
          method  x   n      mean     lower     upper
1  agresti-coull 40 100 0.4000000 0.3093314 0.4980673
2     asymptotic 40 100 0.4000000 0.3039818 0.4960182
3          bayes 40 100 0.4009901 0.3066746 0.4963954
4        cloglog 40 100 0.4000000 0.3040039 0.4940526
5          exact 40 100 0.4000000 0.3032948 0.5027908
6          logit 40 100 0.4000000 0.3088415 0.4986527
7         probit 40 100 0.4000000 0.3078765 0.4980788
8        profile 40 100 0.4000000 0.3074139 0.4976518
9            lrt 40 100 0.4000000 0.3074044 0.4976691
10     prop.test 40 100 0.4000000 0.3047801 0.5029964
11        wilson 40 100 0.4000000 0.3094013 0.4979974
```

How to Interpret

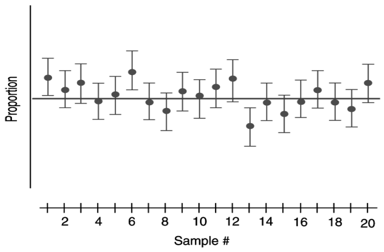
- What do we mean when we say we have 95% confidence that our interval contains the true proportion?
- Formally, what we mean is that "95% of samples of this size will produce confidence intervals that capture the true proportion."
- This is correct but a little long-winded, so we sometimes say "we are 95% confident that the true proportion lies in our interval."

What Does "95% Confidence" Really Mean?

- Each confidence interval uses a sample statistic to estimate a population parameter.
- But, since samples vary, the statistics we use, and thus the confidence intervals we construct, vary as well.

What Does "95% Confidence" Really Mean?

- The figure to the right shows that some of our confidence intervals capture the true proportion (the green horizontal line), while others do not:



What Does "95% Confidence" Really Mean?

- Our confidence **is in the *process*** of constructing the interval, not in any one interval itself.
- Thus, we expect 95% of all 95% confidence intervals to contain the true parameter that they are estimating.

Standard Error and Margin of Error

- The confidence interval is given by

$$\hat{p} \pm \boxed{1.96\sqrt{\frac{\hat{p}\hat{q}}{n}}}$$

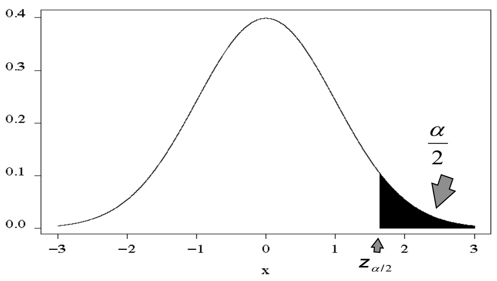
The Standard Error

The Margin of Error

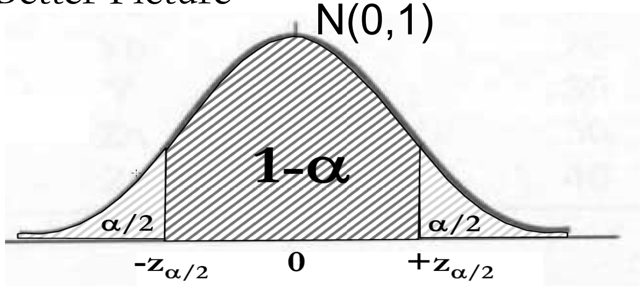
- The standard form of any confidence interval is estimate±(margin of error).

Some strange notation

- We define $z_{\alpha/2}$ to be the point on the normal curve as follows

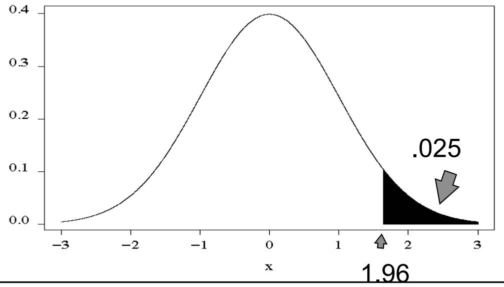


A Better Picture



Example of Strange Notation

For example, when α=5%, $z_{\alpha/2}$=1.96 and we have



Confidence Levels

- Any size confidence interval is then given by

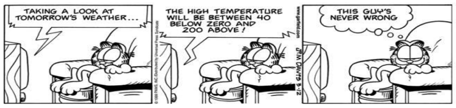
$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- Here are the most common values

Confidence Level	*Confidence Coefficient,* $1-\alpha$	*z value,* $z_{\alpha/2}$
80%	**.80**	**1.28**
90%	.90	1.645
95%	.95	1.96
98%	**.98**	**2.33**
99%	.99	2.58
99.8%	**.998**	**3.08**
99.9%	**.999**	**3.27**

Certainty versus Precision

- The more confident we want to be, the larger the margin of error must be.
- We can be 100% confident that any proportion is between 0% and 100%, but that's not very useful.
- Or we could give a narrow confidence interval, say, from 33.98% to 34.02%. But we couldn't be very confident about a statement this precise.
- Every confidence interval is a balance between certainty and precision.


TAKING A LOOK AT TOMORROW'S WEATHER...
THE HIGH TEMPERATURE WILL BE BETWEEN 40 BELOW ZERO AND 200 ABOVE!
THIS GUY'S NEVER WRONG
JPM DAVIS 3-12
www.garfield.com

Example : Survey Data

A CNN/USA Today/Gallup Poll asked 299 parents of K-12 children the following question (during March 2009):

Thinking about your oldest child, when he or she is at school, do you fear for his or her physical safety?

Of the parents surveyed, 136 (45.5%) answered "Yes" and 163 (54.5%) answered "No." The pollsters reported a margin of error of +/– 6 percent.

Where does this 6% come from?

Examine the Formula

- Using our binomial confidence interval formula, the confidence interval for the proportion of “yes” responses is

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.455 \pm 1.96\sqrt{\frac{.455(.545)}{299}} = 0.455 \pm 0.058$$

The 6% mentioned on the last slide. This is what pollsters call the **“margin of error”**

Determining Sample Size

- Say we want to perform a survey. How many people do we need to poll to be, oh, within 3% of the true value ?

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Longrightarrow 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = .03$$

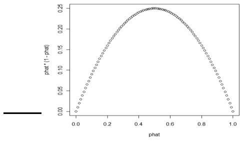
$$or \quad n = (1.96)^2\,\hat{p}(1-\hat{p})\,/\,(.03)^2$$

Example: Calculating Sample Size

- How do we find n? We need to know a value for $\hat{p}$:

$$n = (1.96)^2\,\hat{p}(1-\hat{p})\,/\,(.03)^2$$

- What’s the worst case scenario for phat?



The value is maximized when $\hat{p}$=0.5

Example: Calculating Sample Size

- So use $\hat{p}$=0.5 as the “worst case scenario” when performing sample size.
- So the desired sample size is

$$\begin{aligned} n &= (1.96)^2\,\hat{p}(1-\hat{p})\,/\,(.03)^2 \\ &= (3.84)(0.5)(0.5)\,/\,(.03)^2 \\ &= (0.96)\,/\,(.03)^2 \\ &= 1066.66 \end{aligned}$$

So we should sample 1067 people-we always round up.

Small samples and/or no successes?

- I once taught a short course to twenty students at General Electric
- As part of a class survey, I asked how many were vegetarians.
- None of them were.
- What happens to the confidence interval in this case?

Confidence Interval

- What does the asymptotic interval give if we observe no successes?

```
➢ binom.confint(0,20)
          method x  n       mean         lower      upper
1  agresti-coull 0 20 0.00000000 -2.868440e-02 0.18980956
2     asymptotic 0 20 0.00000000  0.000000e+00 0.00000000
3          bayes 0 20 0.02380952  0.000000e+00 0.09047643
4        cloglog 0 20 0.00000000  0.000000e+00 0.16843347
5          exact 0 20 0.00000000  0.000000e+00 0.16843347
6          logit 0 20 0.00000000  0.000000e+00 0.16843347
7         probit 0 20 0.00000000  0.000000e+00 0.16843347
8        profile 0 20 0.00000000  0.000000e+00 0.15022677
9            lrt 0 20 0.00000000  0.000000e+00 0.09156913
10     prop.test 0 20 0.00000000  0.000000e+00 0.20045335
11        wilson 0 20 0.00000000  1.164173e-17 0.16112516
```

- The interval is (0,0) ☹ not that useful

Other CI's for the Proportion

- R has several different types of CI's for the proportion.
- The one we just discussed and will use on hw's and exams is called the Asymptotic or Wald Interval.
- The Wald interval is ok (actually it sucks), but especially not great if either n is small and/or p is near 0 or 1.

The Agresti Interval

- Alan Agresti-sometimes teaches Stat 101 in the fall

- He wrote an extensive paper a few years ago comparing and discussing all the different confidence intervals for the proportion.

The Agresti Interval (cont)

- We usually define the sample proportion as:

$$\hat{p} = \frac{\text{x}}{\text{n}} = \frac{\text{number of successes in the sample}}{\text{sample size}}$$

- Under the Agresti approach, we define it as

$$\hat{p} = \frac{\text{x+2}}{\text{n+4}}$$

- The do the CI formula we have been using

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

R Example

- Consider the Gillete Example from earlier, with 40 out of 100 men preferring the Sensor razor.

```
> binom.confint(40,100)
         method  x   n      mean     lower     upper
1  agresti-coull 40 100 0.4000000 0.3093314 0.4980673
2     asymptotic 40 100 0.4000000 0.3039818 0.4960182
3          bayes 40 100 0.4009901 0.3066746 0.4963954
4        cloglog 40 100 0.4000000 0.3040039 0.4940526
5          exact 40 100 0.4000000 0.3032948 0.5027908
6          logit 40 100 0.4000000 0.3088415 0.4986527
7         probit 40 100 0.4000000 0.3078765 0.4980788
8        profile 40 100 0.4000000 0.3074139 0.4976518
9            lrt 40 100 0.4000000 0.3074044 0.4976691
10     prop.test 40 100 0.4000000 0.3047801 0.5029964
11        wilson 40 100 0.4000000 0.3094013 0.4979974
```

R Example

- Now consider 0 out of 20 people are vegetarians:

```
 > binom.confint(0,20)
         method x  n       mean         lower      upper
1  agresti-coull 0 20 0.00000000 -2.868440e-02 0.18980956
2     asymptotic 0 20 0.00000000  0.000000e+00 0.00000000
3          bayes 0 20 0.02380952  0.000000e+00 0.09047643
4        cloglog 0 20 0.00000000  0.000000e+00 0.16843347
5          exact 0 20 0.00000000  0.000000e+00 0.16843347
6          logit 0 20 0.00000000  0.000000e+00 0.16843347
7         probit 0 20 0.00000000  0.000000e+00 0.16843347
8        profile 0 20 0.00000000  0.000000e+00 0.15022677
9            lrt 0 20 0.00000000  0.000000e+00 0.09156913
10     prop.test 0 20 0.00000000  0.000000e+00 0.20045335
11        wilson 0 20 0.00000000  1.164173e-17 0.16112516
```

On homeworks and exams we will use the original confidence interval (its easier to do by hand).
In the real world use the Agresti interval.

Things you should know

❑What is a confidence interval

❑How to calculate CI's for proportions

❑How to determine required sample size

Truth	Guess	95% Confidence Interval
p	$\hat{p}$	$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$