Topic 6 - Confidence intervals based on a single sample

- Sampling distribution of the sample mean
- Confidence interval for a population mean
- Confidence interval on matched pairs (special case)
- Confidence interval for a population proportion

Confidence Intervals

- To use the CLT in our examples, we had to know the population mean, μ, and the population standard deviation, σ.

- It is okay to estimate the population parameters if we have a huge amount of sample data, using $\bar{x}$ and s respectively.

- In most cases, the primary goal of the analysis of our sample data is to estimate and to determine a range of likely values for these population values.

Confidence interval for μ with σ known

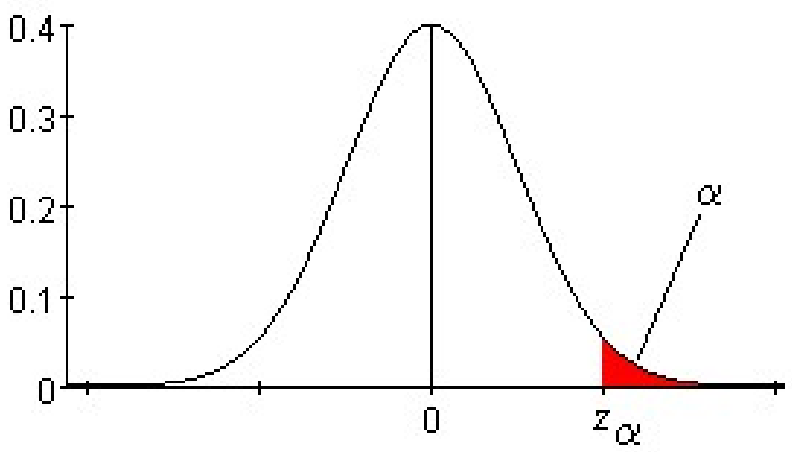
- Suppose for a moment we know σ but not u **(theoretical)**

- The CLT says that $\bar{X}$ is approximately normal with a mean of μ and a variance of σ^2/n.

- So, $Z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ will be standard normal.

- For the standard normal distribution, let z_α be such that

  $P(Z > z_\alpha) = \alpha$



Confidence interval for μ with σ known

For example, if you want a 95% confidence interval, the area between the extremes, limits or boundaries specified is 95% and the area outside, or alpha, is 5%.

Again, theoretical. If sigma is known, we know that the sampling distribution is Normal and can use the following formula;

$$P(\bar{x} - Z_{\alpha/2}\,\sigma/\sqrt{n} \leq \mu \leq \bar{x} + Z_{\alpha/2}\,\sigma/\sqrt{n}) \Rightarrow \bar{x} \pm Z_{\alpha/2}\,\sigma/\sqrt{n}$$

Confidence interval for μ with σ unknown

- For all sampling distribution of means where samples are taken **from a normal population** (or "assumed normal")

$$\bar{x} \pm t_{\alpha/2,n-1}\, s/\sqrt{n}$$

is a (1-α)100% confidence interval for μ.

- The value $t_{\alpha,n-1}$ is the appropriate quantile from a *t* distribution with *n*-1 degrees of freedom.

- Demonstration Z at .05 vs. varying levels of T in Rstudio, using the "qt" function.

Acid rain data CI example

- The EPA states that any area where the average pH of rain is less than 5.6 on average has an acid rain problem. The pH values collected at Shenandoah National Park are listed in acidrain.csv.
- Calculate 95% and 99% confidence intervals for the average pH of rain in the park.

```
> mean(pH_level)
[1] 4.827888889
> sd(pH_level)
[1] 0.289189081
> nrow(dta)
[1] 90
```

On the test, you may be required to calculate or identify these, or it's more likely that you'd be given the information in the problem.

$$\bar{x} \pm t_{\alpha/2,n-1}\left(s/\sqrt{n}\right)$$

4.8279 +/- (-1.987)(0.2891891/sqrt 90)

= (4.767 ; 4.888)

Light bulb data CI example

- The lifetimes in days of 10 light bulbs of a certain variety are given below. Give a 95% confidence interval for the expected lifetime of a light bulb of this type. Do you trust the interval?

Summary statistics:				
Column	**n**	**Mean**	**Variance**	**Std. Dev.**
lifetime	10	0.5580	0.9767	0.9883

Using T $\quad T_{.025,9} = \pm 2.2622$

$$0.5580 \pm 2.2622\left(\frac{0.9883}{\sqrt{10}}\right)$$

$$(-0.1499; 1.2650)$$

Interpreting confidence statements

For any % CI, a valid interpretation is,

"If we perform this study a large number of times, we would expect the estimated parameter to fall within our confidence interval that % of the time."

The "confidence" deals with the long-term view, not the percentage on any specific trial.

Also, note that the interpretation is on the process and is technically not a probability relating to any specific generated CI.

"about 95% will be correct and about 5% will be incorrect"

Demonstration in R shown through simulation on next slide

R code to simulate simulateous confidence intervals

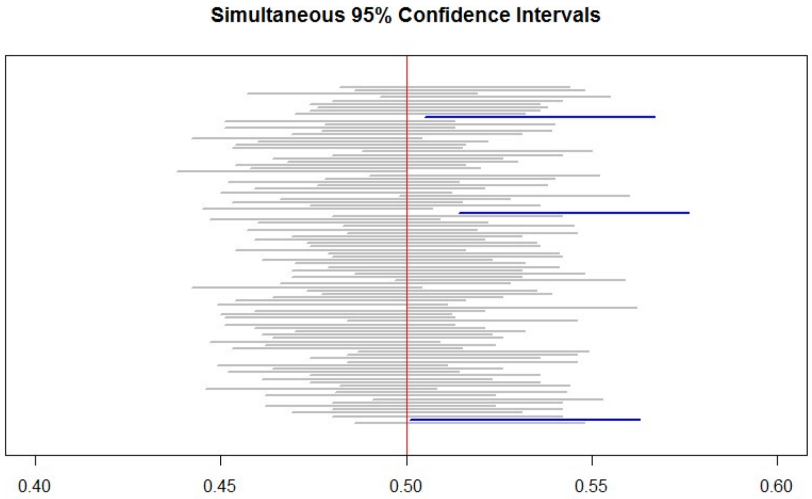
```
par(mfrow=c(1,1))
p=.5
out = c()
#Randomly generate 100 vectors of 0s and 1s of length 1015 (with
population proportion .5) and compute the sample proportion

for (i in 1:100) out[i] =
round(mean(sample(c(0,1),1015,prob=c(p,p),replace=T)),3)

#Compute the approximation to the margin of error.

ME = round(1/sqrt(1015),3)

#plot a blank graph window

plot(0,0,col='white',ylim=c(0,1.1),xlim=c(.4,.6),axes=F,ylab='100 Repeated
Samples',xlab='Sample Proportions',main='Simultaneous 95% Confidence
Intervals')
spots = seq(.05,1.05,length.out=100)
```

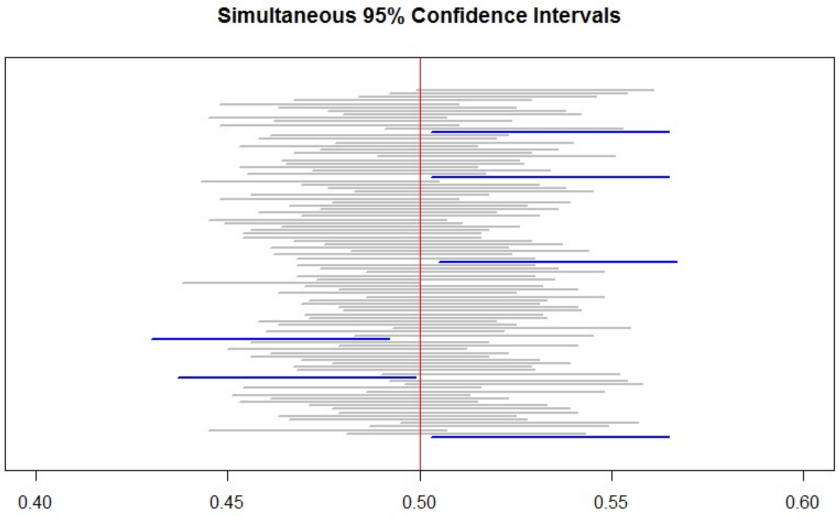
R code to simulate simulateous (cont)

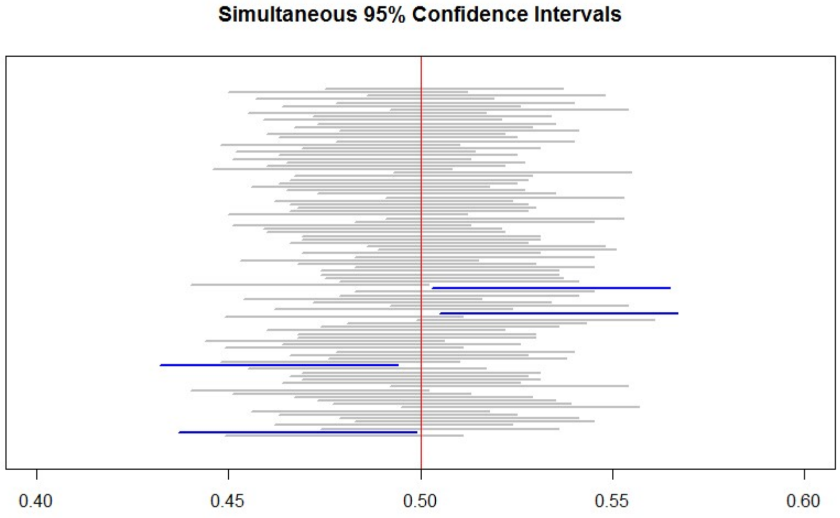
```
#Determine if the parameter is included in the interval and plot either a
blue (excluded) or grey (included) line

for (k in 1:100) {
  if (out[k]-ME > p | p >out[k]+ME) {
    segments(out[k]-ME,spots[k],out[k]+ME,spots[k],col='blue',lty=1,lwd=2)
  } else {
    segments(out[k]-ME,spots[k],out[k]+ME,spots[k],col='grey',lty=1,lwd=2)
  }
}
#Add a vertical red line at the parameter value

abline(v=p,col='red')

#Redraw the box and give approrpiate x-axis ticks.
box()
axis(1,at=seq(.4, .6,length.out=5))
```

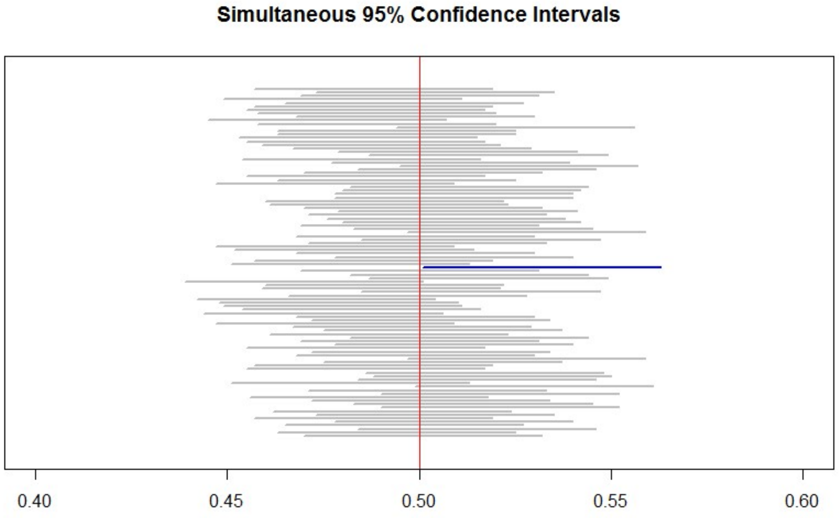
Sample R outputs of simulation runs

It will provide a graph of simultaneous confidence intervals. If the CI contains the parameter being estimated, the line is black. If not, it's blue.









Sample size effect

- As sample size increases, the width of our confidence interval clearly decreases (of course, as long as you're dealing with sample means).

- If σ is known, the width of our interval for μ is

$$L = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Solving for n,

$$n = (2z_{\alpha/2}\frac{\sigma}{L})^2$$

- If σ is known or we have a good estimate, we can use this formula to decide the sample size we need to obtain a certain interval length.

Paint example

- Suppose we know from past data that the standard deviation of the square footage covered by a one gallon can of paint is somewhere between 2 and 4 square feet.

- How many one gallon cans do we need to test so that the width of a 95% confidence interval for the mean square footage covered will be at most 1 square foot?

$$n = (2z_{\alpha/2}\frac{\sigma}{L})^2 \text{ or } (2\,[1.96]\,[\frac{4}{1}])^2 = 245.86 \Rightarrow 246 \text{ cans}$$

- What if we weren't conservative and used $\sigma = 2$?

$$n = (2z_{\alpha/2}\frac{\sigma}{L})^2 \text{ or } (2\,[1.96]\,[\frac{2}{1}])^2 = 61.47 \Rightarrow 62 \text{ cans}$$

Confidence interval on matched pairs (paired difference)

- This would be a special case where two observations are taken on **the same subject or object**, both with and without a designated change.

- Sounds complicated, but it's really what we've been doing for $\bar{x}$ all along, you just do the same calculations on the set of **differences, frequently stated as the variable "d".**

- In the subsequent example, the differences between two groups is what is being focused on. You'll see a lot of similarities in confidence intervals and hypothesis tests here, but we're effectively using the variable "d" for differences, instead of the "x" values shown previously.

Confidence interval on matched pairs (paired difference)

The data is stored in a form like this;

TwinA,TwinB,Type

5.9375,5.5625,BG

5.1875,5.6875,GB

6.4375,5.8125,BB.....

And the following code reads that data, separates it into a TwinA vs. TwinB observation for only “boy boy” twins”, then performs a t test on the “differences” between the two.

```
dta=read.csv("Twins_Data.csv")
attach(dta)
names(dta)
options(digits=10)
twin1.sub=subset(dta$TwinA,Type=="BB")
twin2.sub=subset(dta$TwinB,Type=="BB")
t.test(twin1.sub,twin2.sub,alternative="two.sided",paired=TRUE,conf.level=.9)
detach(dta)
```

Confidence interval on matched pairs (paired difference)

```
dta=read.csv("Twins_Data.csv")
attach(dta)
names(dta)
options(digits=10)
twin1.sub=subset(dta$TwinA,Type=="BB")
twin2.sub=subset(dta$TwinB,Type=="BB")
t.test(twin1.sub,twin2.sub,alternative="two.sided",paired=TRUE,conf.level=.9)
detach(dta)
```

```
        Paired t-test

data:  twin1.sub and twin2.sub
t = -2.2156846, df = 8, p-value = 0.05756832
alternative hypothesis: true mean difference is not equal to 0
90 percent confidence interval:
 -0.9324055103 -0.0814833786
sample estimates:
mean difference
 -0.5069444444
```

Confidence interval for a proportion

- A binomial random variable, X, counts the number of successes in n Bernoulli trials where the probability of success on each trial is p.

- In sampling studies, we are often times interested in the proportion of items in the population that have a certain characteristic.

- We think of each sample, X_i, from the population as a Bernoulli trial, the selected item either has the characteristic ($X_i = 1$) or it does not ($X_i = 0$).

- The total number with the characteristic in the sample is then

$$X = \sum_{i=1}^{n} X_i$$

Confidence interval for a proportion

- The proportion in the sample with the characteristic is then

$$\hat{p} = X / n = \sum_{i=1}^{n} X_i / n$$

- The "CLT" says that the distribution of this sample proportion is then normally distributed for large *n* with

$$E(X / n) = E(\bar{X}) = \mu = 0(1-p) + 1p = p$$

$$Var(X / n) = \frac{Var(X)}{n} = \frac{E(X^2) - E(X)^2}{n} = \frac{0^2(1-p) + 1^2 p - p^2}{n} = \frac{p(1-p)}{n}$$

- **For the CLT to work well here, we need**
  - ***X*** **≥ 10 and** ***n*****-*****X*** **≥ 10**
  - ***n*** **to be much smaller than the population size**

Confidence interval for a proportion

- Example: In a survey of 277 randomly selected shoppers, 69 stated that if the advertised item is unavailable, they request a rain check. Construct a 95% CI for p.

- What is the underlying distribution? What is $\hat{p}$? What is p ?

$$\hat{p} = \frac{x}{n} = \frac{69}{277} = 0.249$$

$$n\hat{p} = 277(.249) = 69 \geq 10$$
$$n\hat{q} = 277(.751) = 208 \geq 10$$

$$\hat{p} \pm Z_{\alpha/2}\sqrt{(\hat{p}\hat{q})/n}$$

$$0.249 \pm 1.96\sqrt{(0.249 * 0.751)/277}$$

$$0.249 \pm 0.051 \quad \text{or} \quad (0.198; 0.300)$$

Murder case example

- Find a 95% confidence interval for the proportion of African Americans in the jury pool. (22 out of 295 African American in sample)

$$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} = \frac{22}{295} \pm [1.96]\sqrt{\frac{(0.074576)(0.925424)}{295}}$$

$$= .074576 \pm 0.02998 = (0.0446; 0.10456)$$

Sample size considerations

- The width of the confidence interval is

$$L = 2Z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

- To get the sample size required for a specific length, we have

$$n = 4Z^2_{\alpha/2}\hat{p}\hat{q}/L^2$$

CI sample size with no clue for actual proportion

- Often in TV/Media polls, they talk about a *margin of error of +/- 3%.* This margin of error is actually a 95% confidence interval. How large of a sample do we need to have to result in a margin of error of 3%?

$$\pm 3\% \quad \rightarrow \quad L = .06$$

$$n = 4Z^2_{\alpha/2}\frac{\hat{p}\hat{q}}{L^2}$$

$$n = 4(1.96^2)\frac{\hat{p}\hat{q}}{.06^2}$$

- How do we continue? What is $\hat{p}$? We have a catch-22 situation, because to find the sample size we need an estimate of the population proportion and vice-versa.

CI sample size – "no clue for p" (cont)

Let's try some different values of $\hat{p}\hat{q}$ *:*

$\hat{p}$	$\hat{p}\hat{q}$	$\hat{p}$	$\hat{p}\hat{q}$	$\hat{p}$	$\hat{p}\hat{q}$
.1	.09	.4	.24	.7	.21
.2	.16	.5	.25	.8	.16
.3	.21	.6	.24	.9	.09

It turns out that the value of $\hat{p}\hat{q}$ *maximizes when* $\hat{p}$ =.50. Therefore, even if we don't have a clue for the real value of p, we can solve for n by setting $\hat{p}$ =.50 as a worst case scenario…..

$$n = 4Z^2_{\alpha/2}\hat{p}\hat{q}/L^2$$

$$n = 4(1.96^2)(.5)(.5)/.06^2$$

$$n = 1,067.11 \Rightarrow 1,068$$

So, if you have a good ***supportable*** *estimate for the population proportion, use it. If not, you 0.50.*

Nurse employment case

How many sample records should be considered if we want the 95% confidence interval for the proportion all her records handled in a timely fashion to be 0.02 wide?

$$n = 4Z^2_{\alpha/2}\frac{\hat{p}\hat{q}}{L^2} = 4(1.96^2)(\frac{.9(.1)}{.02^2}) \Rightarrow 3457.44 \Rightarrow 3458$$

Where did the .9 come from? That was the original percentage from the nurses contract in Topic 3. If that wasn't in place on average, there would be chaos in the hospital.

The additional file for Topic 6 contains examples of confidence intervals, sample size calculations and prediction intervals.

Prediction intervals

- Sometimes we are not interested in a confidence interval for a population parameter but rather we are interested in a prediction interval for a new observation.

- For our light bulb example, we might want a 95% prediction interval for the time a new light bulb will last.

- For our paint example, we might want a 95% prediction interval for the amount of square footage a new can of paint will cover.

Prediction intervals

- Given a random sample of size n, our best guess at a new observation, x_{n+1}, would be the sample mean, $\bar{X}$.
- Now consider the difference, $x_{n+1} - \bar{X}$.
- $E(x_{n+1} - \bar{X}) =$
- $Var(x_{n+1} - \bar{X}) =$

Prediction intervals

- If σ is known and our population is normal, then using a pivoting procedure gives

$$\bar{x} \pm t_{\alpha/2,n-1}(\sigma\sqrt{1+\frac{1}{n}})$$

- If σ is unknown (far more likely) and our population is normal, a (1-α)100% prediction interval for a new observation is

$$\bar{x} \pm t_{\alpha/2,n-1}(s\sqrt{1+\frac{1}{n}})$$

Acid rain example

- For the acid rain data, the sample mean pH was 4.577889 and the sample standard deviation was 0.2892. What is a 95% prediction interval for the pH of a new rainfall?

$$4.57789 \pm t_{.025,89}(.2892)\sqrt{\frac{91}{90}}$$

$$4.57789 \pm 1.986979(.290802)$$

$$(4.0001; 5.1557)$$

- Does this prediction interval apply for the light bulb data?

Additional Prediction Interval

- A 7th generation citrus farmer is worried about revenues from this year's crop, specifically because if he doesn't get at least an 82% yield, he won't be able to make the payments on the family farm and the bank will foreclose.

- It's late spring and the last cold snap is coming tonight; forecast low is 30-degrees. He pulls up the Department of Agriculture data for citrus crops for the area and finds 30 matches for impacts on crop yields due to a 30-degree night, with an average of 85% and a standard deviation of 3.5%.

- Right before bed, he mentions this all to his wife and the last thing his wife says is, "Well, I'm sure you've run some sort of prediction interval to tell whether we're in trouble or not…….Zzzzzzz"

Additional Prediction Interval (cont)

Will the farmer sleep well tonight (ie., is there a statistical chance that he's not going to make 82% or more on a crop yield?), using a 95% Prediction Interval?

$$\bar{x} \pm t_{\alpha/2,n-1} s\sqrt{1+\frac{1}{n}} \qquad .85 \pm (2.0452) * 035\sqrt{1+\frac{1}{30}}$$

$$.85 \pm .0728 = (0.7772; 0.9228)$$

The answer would be no. His disaster number falls within the realm of reasonable values, given the historical values……

He could expect to lose the farm