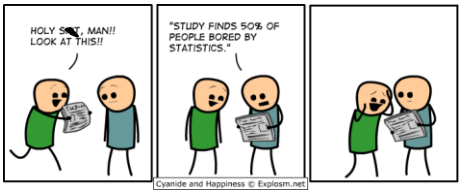
Chapter 6: Estimation and Confidence Intervals..

How to construct and interpret confidence intervals for:

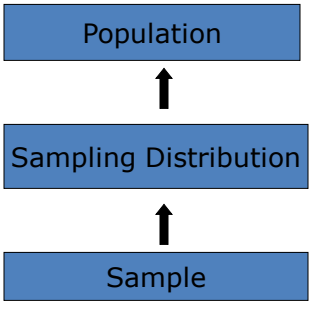
- Sample means
- Sample proportions



6-1

The 3 types of distributions in Inferential Statistics

- Every application of inferential statistics involves 3 different distributions.
  - **Population Distribution** empirical; typically unknown
  - **Sampling Distribution:** non-emprical; known via theory
  - **Sample Distribution:**
  - empirical; known through observation



Information from the sample is linked to the population via the sampling distribution.

6-2

Basic Logic of Estimation

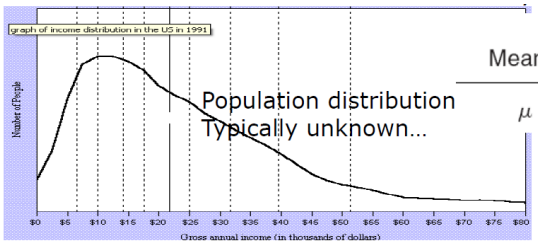
In estimation procedures, *statistics* calculated from random samples are used to estimate the value of population *parameters*, **with a varying level of success depending on:**

sample size and corresponding sampling error

Information on error is implied in "sampling distributions" with relatively large "standard errors" indicating lots of sampling error!!

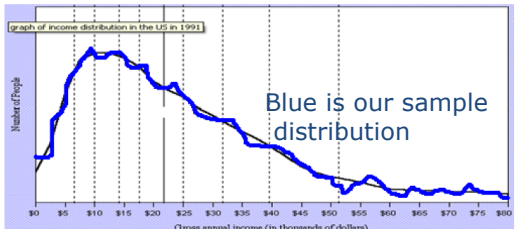
"This is interesting, 70% of the respondents to our survey said they don't respond to surveys."

- Reminder from last class, we have three basic types of distributions:



Mean	Standard Deviation	Proportion
μ	σ	P_u

Difference between the two typically involves what is referred to as "sampling error"

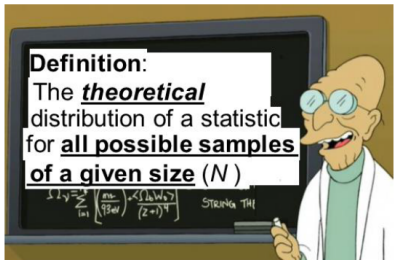


Mean	Standard Deviation	Proportion
$\overline{X}$	s	P_s

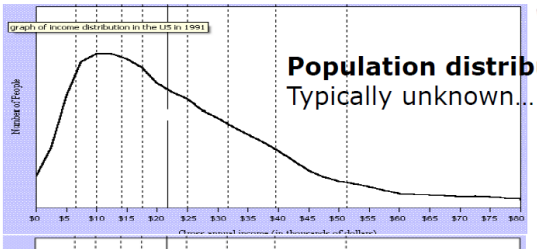
6-26

The 3rd type of distribution: sampling distribution

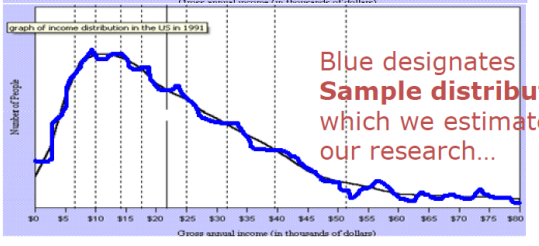
The single most important concept in inferential statistics (very different from the sample and population distribution)



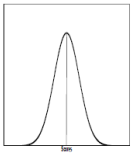
6-5



Population distribution
Typically unknown…



Blue designates
Sample distribution
which we estimate via
our research…



Sampling distribution

We can use this to estimate sampling error & "confidence intervals"!!! (this class)
Recall: its standard deviation is called the SE

6-6

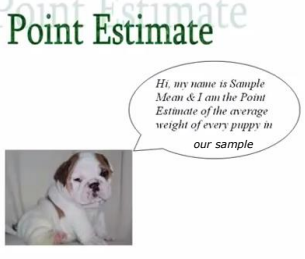
In statistics:

Two Estimation Procedures:
1. **Point estimates and**
2. **confidence intervals**

Point estimates

1. ***A point estimate*** is a sample statistic used to estimate a population value.
   Example: A random sample of puppies in Ontario documented that the average weight at 6 weeks is 2.5 pounds

Problem with point estimates:
In and of themselves, point estimates leave us with little information on the likely precision or efficiency of the estimate…



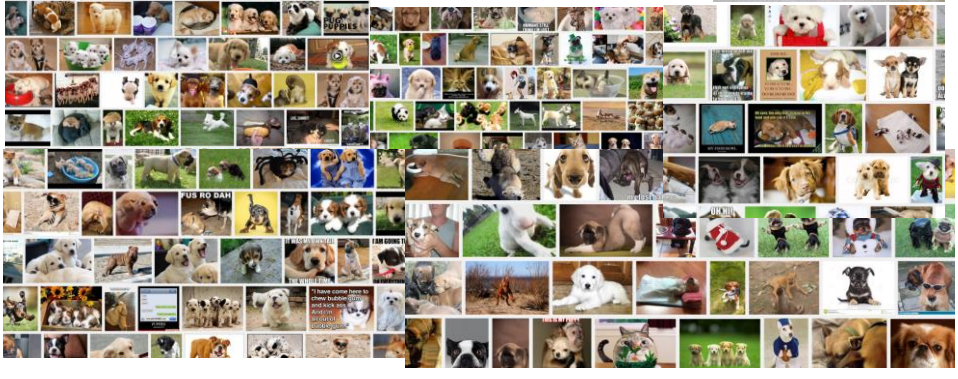
Is this likely to be very close to the population parameter?

Is this point estimate based on a very tiny sample?

hence; very inefficient.. Imprecise statistic???

Is this point estimate based on a larger sample?

Hence: higher quality estimate: highly efficient!!



2. ***Confidence intervals:***

They consist of a range of values.

Example: A random sample of puppies in Ontario documented that the average weight at 6 weeks as 2.5, or more specifically, 2.5 +/-0.3 pounds

(i.e. somewhere between 2.2 and 2.8 pounds)

… with a given level of "probability" e.g 95% of the time ***(19 times out of 20)!***

Provide us some sense as to the accuracy of statistics.

How wide is the range? Wide range, less accuracy!!

Mean is 2.5

One sample: CI is 2.2 pounds – 2.8 pounds, 95% of the time

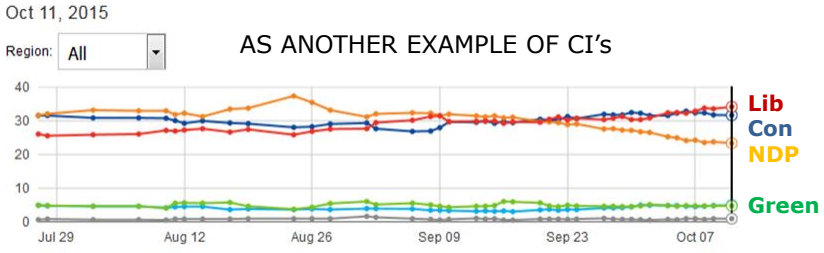
Next sample: CI is 2.4 pounds – 2.6 pounds, 95% of the time (more efficient)

NOTE:

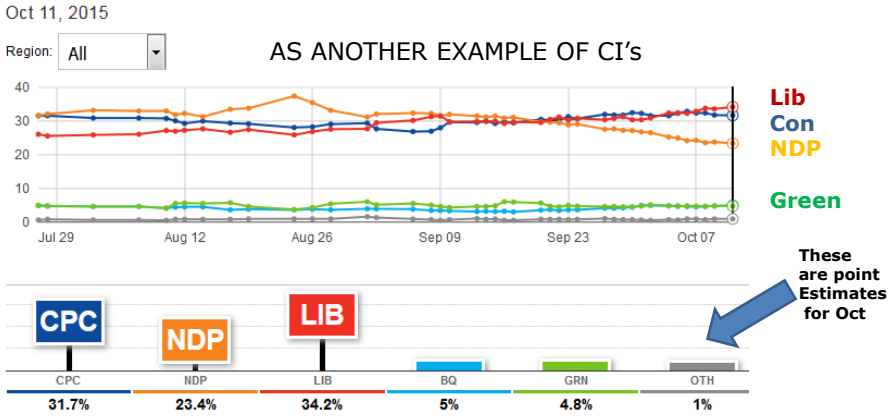
We take advantage of the "sampling distribution" in calculating these "intervals"..

(NORMAL DISTRIBUTION)

Last Federal election (polls leading up to a few weeks prior to the date)



Last Federal election (a few weeks prior to the date)?



MOE: +/- 2.5%, 19 times out of 20 **This tells us about potential confidence intervals**

Liberals are at: 34.2% +/- 2.5%, 19 times out of 20 (31.7 – 36.7%, 95% of the time)

Conservatives are at 31.7% +/- 2.5%, 19 times out of 20 (29.2 -34.2%, 95% of the time)

6-13

What does that mean? 19 times out of 20?

This refers to our "sampling distribution".. (our theoretical distribution) i.e. in 19 sample estimates out of 20!!!

If we were to repeatedly sample the Canadian population, again and again and again (in mid Oct, 2015) roughly 95% of the time our estimate of the percentage voting for the Liberals would fall between **31.7 – 36.7%**

5% of the time, we are wrong on this inference..

6-14

Constructing Confidence Intervals

We want to construct an interval working with a sample whereby the true population parameter likely lies…

Procedures:

1. Set what is called our "alpha level" (probability of being wrong: typically .05).
2. Find the associated *Z* score of the normal distribution that corresponds to this alpha (working with our sampling distribution).
3. Substitute values into the appropriate formula for constructing confidence intervals.. Several formulas are possible.. (details to come)

Relatively easy, but first I must first give you a bit more back ground..

Constructing Confidence Intervals

First step: Decide upon how much of a risk we are willing to take (of being wrong, with the true population parameter in reality being outside of our interval).

Called the "alpha level" (typically set at .05)

Also called:

the 95% confidence level interval,..

Correct 19 times out of 20

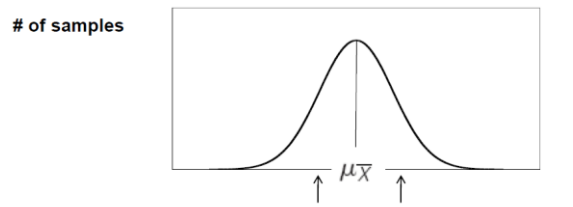
1 time out of 20, by chance, our interval doesn't contain the parameter (either below or above our interval)

Constructing Confidence Intervals

Secondly, we know stuff about our "sampling distribution" (review last week)

The Sampling Distribution

Normal in shape



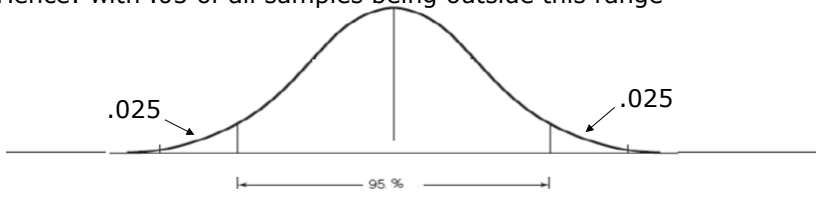
Innumerable different samples have many means

We are interested in knowing where 95% of our sample estimates would fall in our sampling distribution, by mere chance alone

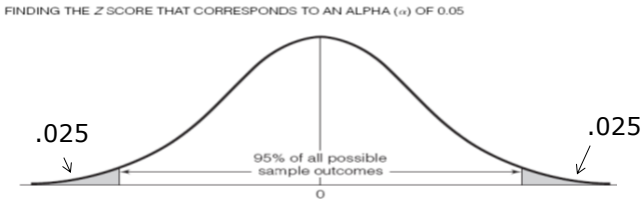
5% of all sample estimates would fall outside of this range, with an equal probability of being higher than the mean and an equal probability of being lower than the mean

AREAS UNDER THE THEORETICAL NORMAL CURVE

Hence: with .05 of all samples being outside this range



In a normal curve, what Z score would give us 95 percent of all sample outcomes?



What is the appropriate Z score?

Look to Appendix A, but start with
Column C rather than Column A (moving in the opposite direction), find .025 and identify the corresponding Z score…

AN ILLUSTRATION OF HOW TO FIND AREAS UNDER THE NORMAL CURVE USING APPENDIX A

(a) *Z*	(b) Area between Mean and *Z*	(c) Area beyond *Z*
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
⋮	⋮	⋮
1.00	0.3413	0.1587
1.01	0.3438	0.1562
1.02	0.3461	0.1539
1.03	0.3485	0.1515
⋮	⋮	⋮
1.50	0.4332	0.0668
1.51	0.4345	0.0655
1.52	0.4357	0.0643
1.53	0.4370	0.0630
⋮	⋮	

Column C tells us the area in the tail, right?

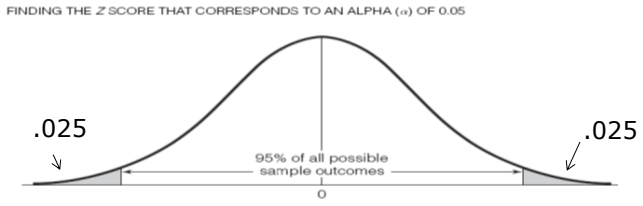
In this case, we are interested in .025..

1.96

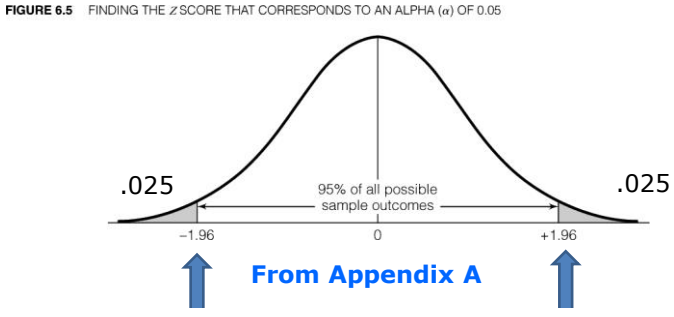
0.0250

Finding the Z score

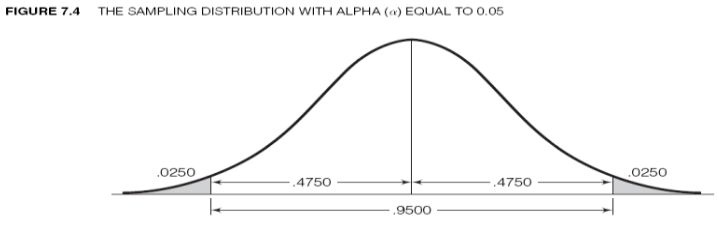
In a normal curve, what Z score would give us 95 percent of all sample outcomes? A situation whereby our sample outcome would Fall within a range, 19 times out of 20???

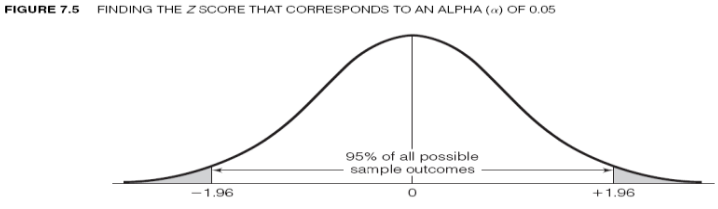


What is the appropriate Z score?



To be precise:
Theoretically, 95% of the sampling distribution falls with +/- 1.96 standard errors from the mean





Z-values for Various Alpha Levels

Confidence Level	α	α/2	Z-score
90%	.10	.0500	+/-1.65
95%	.05	.0250	+/-1.96
99%	.01	.0050	+/-2.58
99.9%	.001	.0005	+/-3.29

(**Note:** Z-scores are found in Appendix A using the area for α/2)

90% of the sampling distribution falls with +/- 1.65 standard errors from the mean
99% of the sampling distribution falls within +/- 2.58 standard errors from the mean, etc.

FOR INTRODUCTORY PURPOSES (I'll elaborate in the next slides):

3 different formulas will be used in this class to calculate confidence intervals:

1. Working with means: when we know our "population standard deviation"

FORMULA 6.1

$$\text{c.i.} = \overline{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

where c.i. = confidence interval
$\overline{X}$ = the sample mean
Z = the Z score as determined by the alpha level
$\frac{\sigma}{\sqrt{n}}$ = the standard deviation of the sampling distribution or the standard error of the mean

2. Working with means: when we do not know our "population standard deviation" but do know our sample standard deviation

FORMULA 6.2

$$\text{c.i.} = \overline{X} \pm Z\left(\frac{s}{\sqrt{n-1}}\right)$$

3. Working with proportions

FORMULA 6.3

$$\text{c.i.} = P_s \pm Z\sqrt{\frac{P_u(1-P_u)}{n}}$$

6-24

First set of calculation:

1. Constructing Confidence Intervals for Means (Population standard deviation known)

First, set the alpha, α (probability that the interval will be wrong).

Example: Setting alpha equal to 0.05, a 95% confidence level, means the researcher is willing to be wrong 5% of the time.

Second, find the *Z* score associated with alpha.

Example; If alpha is equal to 0.05, we would place half (0.025) of this probability in the lower tail and half (0.025) in the upper tail of the distribution. The Z score that corresponds to this will always be +/- 1.96!!!)

- **Third, substitute values into appropriate formulas for confidence intervals for sample means**

If σ known
Formula 6.1

FORMULA 6.1

$$\text{c.i.} = \overline{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

where c.i. = confidence interval
$\overline{X}$ = the sample mean
Z = the Z score as determined by the alpha level
$\frac{\sigma}{\sqrt{n}}$ = the standard deviation of the sampling distribution or the standard error of the mean

Let's get it real.. With a specific example:

A random sample of 178 households watch TV an average of 6 hours per day, with a population standard deviation of 3 (σ =3).

Let's create a 95% CI on this mean..

A random sample of 178 households (n=178) watch TV an average of 6 hours per day, with a population standard deviation of 3 (σ =3).

FORMULA 6.1

$$\text{c.i.} = \overline{X} \pm Z\left(\frac{\sigma}{\sqrt{n}}\right)$$

where c.i. = confidence interval
$\overline{X}$ = the sample mean
Z = the Z score as determined by the alpha level
$\frac{\sigma}{\sqrt{n}}$ = the standard deviation of the sampling distribution or the standard error of the mean

With alpha set to .05, the confidence interval is:

c.i. = 6.0 ±1.96(3/$\sqrt{178}$)

c.i. = 6.0 ±1.96(3/13.34)

c.i. = 6.0 ±1.96(.22)

c.i. = 6.0 ± .44

We can estimate that households in Canada average 6.0±.44 hours of TV watching each day.

Another way to state the interval:

$5.56 \leq \mu \leq 6.44$

We estimate that the population mean is greater than or equal to 5.56 and less than or equal to 6.44.

This interval has a .05 (5%) chance of being wrong.

Only rarely (5 times out of 100) will the interval *not* include μ.

2. Constructing Confidence Intervals for Means (Population standard deviation *unknown*)

First, set the alpha, α (probability that the interval will be wrong).

Example: Setting alpha equal to 0.05, a 95% confidence level, means the researcher is willing to be wrong 5% of the time.

Second, find the *Z* score associated with alpha.

Example; If alpha is equal to 0.05, we would place half (0.025) of this probability in the lower tail and half (0.025) in the upper tail of the distribution.

- Third, substitute values into appropriate formulas for confidence intervals for sample means

If σ unknown Formula 6.2

FORMULA 6.2

$$\text{c.i.} = \overline{X} \pm Z\left(\frac{s}{\sqrt{n-1}}\right)$$

where c.i. = confidence interval
$\overline{X}$ = the sample mean
Z = the Z score as determined by the alpha level

s = sample standard deviation
n= sample size

ON THE BASIS OF 1 SAMPLE ESTIMATE STANDARD ERROR!!!!!!!

IMPORTANT POINT!
Relative to Formula 6.1 σ is replaced by *s*. Further, *n* is replaced by *n−1* to correct for the fact that *s* is a biased estimator of σ.

Constructing Confidence Intervals for Means: An Example

A random sample of 500 puppies are found to weigh on average 2.5 pounds, with a sample standard deviation of 3

With alpha set to .05, the confidence interval is:

c.i. = 2.5 ±1.96(3/√500-1)
c.i. = 2.5 ±1.96(3/22.34)
c.i. = 2.5 ±1.96(.1343)
c.i. = 2.5 ± .26

$$\text{c.i.} = \overline{X} \pm Z\left(\frac{s}{\sqrt{n-1}}\right)$$

We can estimate that among Canadian puppies, their average weight is 2.5 pounds, plus or minus .26 pounds, 19 times out of 20.

Another way to state the interval:

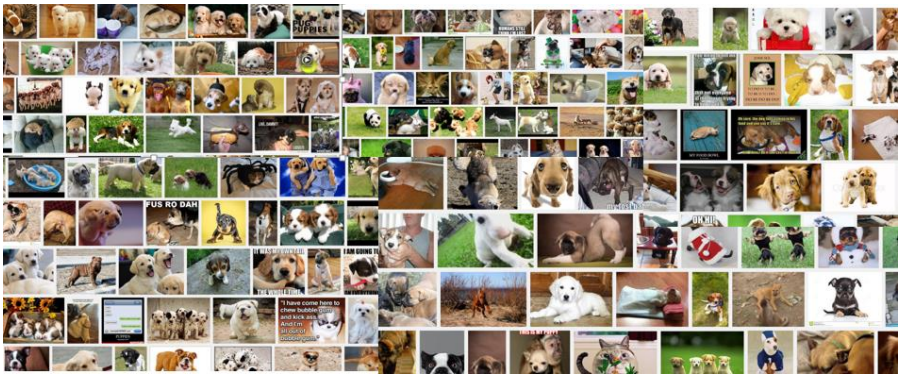
$2.24 \leq \mu \leq 2.76$

We estimate that the population mean is greater than or equal to 2.24 and less than or equal to 2.76.

This interval has a .05 (5%) chance of being wrong.

Only rarely (5 times out of 100) will the interval *not* include μ.

Same problem, but with 5000 puppies???



Constructing Confidence Intervals for Means: An Example

A random sample of 5000 puppies are found to weigh on average 2.5 pounds, with a sample standard deviation of 3

With alpha set to .05, the confidence interval is:

c.i. = 2.5 ±1.96(3/$\sqrt{5000}$-1)

c.i. = 2.5 ±1.96(3/70.71)

c.i. = 2.5 ±1.96(.042)

c.i. = 2.5 ± 0.08

$$\text{c.i.} = \overline{X} \pm Z\left(\frac{s}{\sqrt{n-1}}\right)$$

Note; with 500 it was:
c.i. = 2.5 ± .26

LARGER SAMPLE, NARROWER CI!!!

3. Constructing Confidence Intervals for Proportions (note also %'s)

Procedures:

1. Set alpha.
2. Find the associated *Z* score.
3. Substitute values into the formula for constructing confidence intervals for sample proportions:

*The procedures for constructing confidence intervals provided so far are only for samples of at least 100 persons

FORMULA 6.3

$$\text{c.i.} = P_s \pm Z\sqrt{\frac{P_u(1 - P_u)}{n}}$$

where P_s = sample proportion
Z = Z score as determined by the alpha level
P_u = population proportion (P_u is typically setting at .5)

$\sqrt{\frac{P_u(1 - P_u)}{n}}$ = standard deviation of the sampling distribution of sample proportions

Also called the "standard error" of the proportions, right?

Important point: If we don't know our "Population Proportion, which is typical, it is recommended that you substitute 0.5 for Pu

If 22% of a random sample of 764 adult Canadians smoke, provide a 95% confidence interval of what percentage of adult Canadians smoke?

If 22% of a random sample of 764 adult Canadians smoke, provide a 95% confidence interval of what percentage of adult Canadians smoke?

$$\text{c.i.} = P_s \pm Z\sqrt{\frac{P_u(1 - P_u)}{n}}$$

c.i. = .22 ±1.96 √.5(1-.5)/764
c.i. = .22 ±1.96 (√.25/764)
c.i. = .22 ±1.96 (√.00033)
c.i. = .22 ±1.96 (.018)
c.i. = .22 ±.04

Changing back to %'s, we can estimate that 22% ± 4% of Canadian adults smoke.

Another way to state the interval:

$18\% \leq P_u \leq 26\%$

We estimate the population value is greater than or equal to 18% and less than or equal to 26%.

This interval has a .05 chance of being wrong.

Z-values for Various Alpha Levels

Confidence Level	α	α/2	Z-score
90%	.10	.0500	+/-1.65
95%	.05	.0250	+/-1.96
99%	.01	.0050	+/-2.58
99.9%	.001	.0005	+/-3.29

(**Note:** Z-scores are found in Appendix A using the area for α/2)

Controlling the Width of Confidence Intervals

Confidence interval widens as **confidence level** increases:

TABLE 7.3 INTERVAL ESTIMATES FOR FOUR CONFIDENCE LEVELS ($\bar{X}$ = \$35,000, s = \$200, N = 500 throughout)

Alpha	Confidence Level	Interval	Interval Width
.10	90%		
.05	95%		
.01	99%		
.001	99.9%		

Controlling the Width of Confidence Intervals

Confidence interval widens as **confidence level** increases:

TABLE 7.3 INTERVAL ESTIMATES FOR FOUR CONFIDENCE LEVELS ($\bar{X}$ = \$35,000, s = \$200, N = 500 throughout)

Alpha	Confidence Level	Interval	Interval Width
.10	90%	\$35,000 ± 14.77	\$29.54
.05	95%	\$35,000 ± 17.55	\$35.10
.01	99%	\$35,000 ± 23.09	\$46.18
.001	99.9%	\$35,000 ± 29.45	\$58.90

Controlling the Width of Confidence Intervals

Confidence interval widens as **confidence level** increases:

TABLE 7.3 INTERVAL ESTIMATES FOR FOUR CONFIDENCE LEVELS ($\bar{X}$ = \$35,000, s = \$200, N = 500 throughout)

Alpha	Confidence Level	Interval	Interval Width
.10	90%	\$35,000 ± 14.77	\$29.54
.05	95%	\$35,000 ± 17.55	\$35.10
.01	99%	\$35,000 ± 23.09	\$46.18
.001	99.9%	\$35,000 ± 29.45	\$58.90

Confidence interval narrows as **sample size** increases:

TABLE 7.4 INTERVAL ESTIMATES FOR FOUR DIFFERENT SAMPLES ($\bar{X}$ = \$35,000, s = \$200, alpha = 0.05 throughout)

Sample	N
1	100
2	500
3	1,000
4	10,000

Controlling the Width of Confidence Intervals

Confidence interval widens as **confidence level** increases:

TABLE 7.3 INTERVAL ESTIMATES FOR FOUR CONFIDENCE LEVELS ($\bar{X}$ = \$35,000, s = \$200, N = 500 throughout)

Alpha	Confidence Level	Interval	Interval Width
.10	90%	\$35,000 ± 14.77	\$29.54
.05	95%	\$35,000 ± 17.55	\$35.10
.01	99%	\$35,000 ± 23.09	\$46.18
.001	99.9%	\$35,000 ± 29.45	\$58.90

Confidence interval narrows as **sample size** increases:

TABLE 7.4 INTERVAL ESTIMATES FOR FOUR DIFFERENT SAMPLES ($\bar{X}$ = \$35,000, s = \$200, alpha = 0.05 throughout)

Sample 1 (N = 100)	Sample 2 (N = 500)
c.i. = \$35,000 ± 1.96(200/$\sqrt{99}$) c.i. = \$35,000 ± 39.40	c.i. = \$35,000 ± 1.96(200/$\sqrt{499}$) c.i. = \$35,000 ± 17.55
Sample 3 (N = 1,000)	**Sample 4 (N = 10,000)**
c.i. = \$35,000 ± 1.96(200/$\sqrt{999}$) c.i. = \$35,000 ± 12.40	c.i. = \$35,000 ± 1.96(200/$\sqrt{9,999}$) c.i. = \$35,000 ± 3.92

Sample	N	Interval Width
1	100	\$78.80
2	500	\$35.10
3	1,000	\$24.80
4	10,000	\$ 7.84

In our leger poll

Population All adult Ontario voters

Sample 1000 persons selected

Statistic P_s = .32 (or 32%)

Parameter unknown.
The % of all adult Ontario residents who will vote for party X.

FORMULA 6.3

$$\text{c.i.} = P_s \pm Z\sqrt{\frac{P_u(1 - P_u)}{n}}$$

- If 32% of a random sample of 1000 Ontarians plan on voting for party X, provide a 95% confidence interval of what percentage will vote in this way.
  - c.i. = .32 ±1.96 ($\sqrt{.25/1000}$)
  - c.i. = .32 ±1.96 (.0158)
  - c.i. = .32 ±.03
  - 95% chance, between .29 and .35 or 29% and 35%

- What about southwestern Ontario? (N=250)
- If 32% of a random sample of 250 Ontarians from SW Ontario plan on voting for party X, provide a 95% confidence interval of what percentage of of residents will vote in this way?
  - c.i. = .32 ±1.96 ($\sqrt{.25/250}$)
  - c.i. = .32 ±.062
  - 95% chance, between .258 and .382 or 25.8% and 38.2%

  - What if a second party had 30%?
  - C.i. = .30+/- .062
  - 95% chance, between .238 and .362 or 23.8% and 36.2%

  - HEAVY OVERLAP OF CONFIDENCE INTERVALS ACROSS PARTIES +/- 6%!! (i.e. the differences are not significant!! Using scientific standards we can not say that support is different in the population (it may be, BUT we don't know since our sample is too small)

One more example

Working with a sample of 100,000 persons, we document that:

52% of persons aged 55-64 are not employed

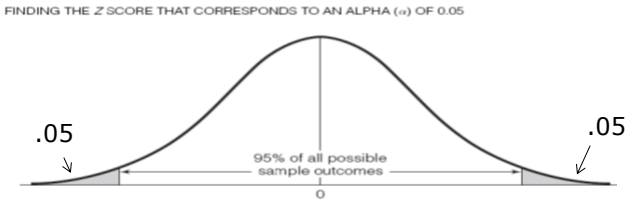
Provide me with a 90% CI on this estimate..

FORMULA 6.3

$$\text{c.i.} = P_s \pm Z\sqrt{\frac{P_u(1 - P_u)}{n}}$$

6-48

In a normal curve, what Z score would give us 90 percent of all sample outcomes?



What is the appropriate Z score?

Look to Appendix A, but start with
Column C rather than Column A (moving in the opposite direction), find .05 and identify the corresponding Z score…

6-49

AN ILLUSTRATION OF HOW TO FIND AREAS UNDER THE NORMAL CURVE USING APPENDIX A

(a) *Z*	(b) Area between Mean and *Z*	(c) Area beyond *Z*
0.00	0.0000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
0.03	0.0120	0.4880
⋮	⋮	⋮
1.00	0.3413	0.1587
1.01	0.3438	0.1562
1.02	0.3461	0.1539
1.03	0.3485	0.1515
⋮	⋮	⋮
1.50	0.4332	0.0668
1.51	0.4345	0.0655
1.52	0.4357	0.0643
1.53	0.4370	0.0630
⋮	⋮	

Column C tells us the area in the tail, right?

In this case, we are interested in .050..

1.65

0.050

Finding the Z score

6-50

One more example

Working with a sample of 100,000 persons, we document that:

52% of persons aged 55-64 are not employed

Provide me with a 90% CI on this estimate..

FORMULA 6.3

$$\text{c.i.} = P_s \pm Z\sqrt{\frac{P_u(1 - P_u)}{n}}$$

c.i. = .52 +/- 1.65 $\sqrt{.25/100{,}000}$
= .52 +/- 1.65 (.00158)
= .52 +/- .002609
= .5174 -> .5226

– 90% chance, between 51.74 and 52.26% (i.e. very precise, no? Highly efficient given sample size)

6-51