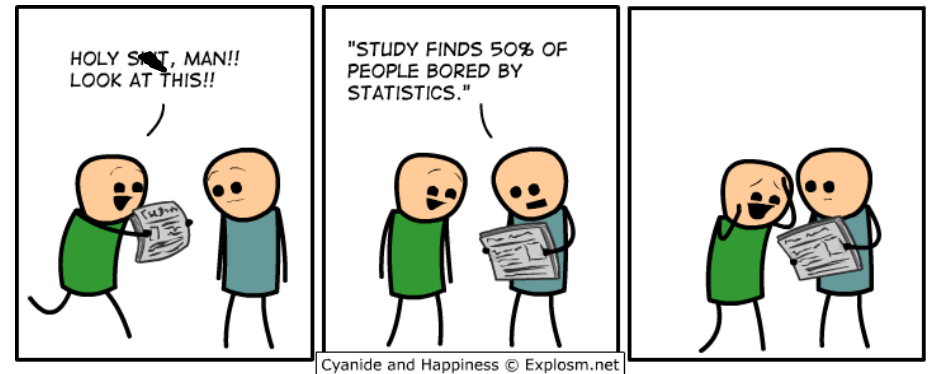


## Chapter 6: Estimation and Confidence Intervals..

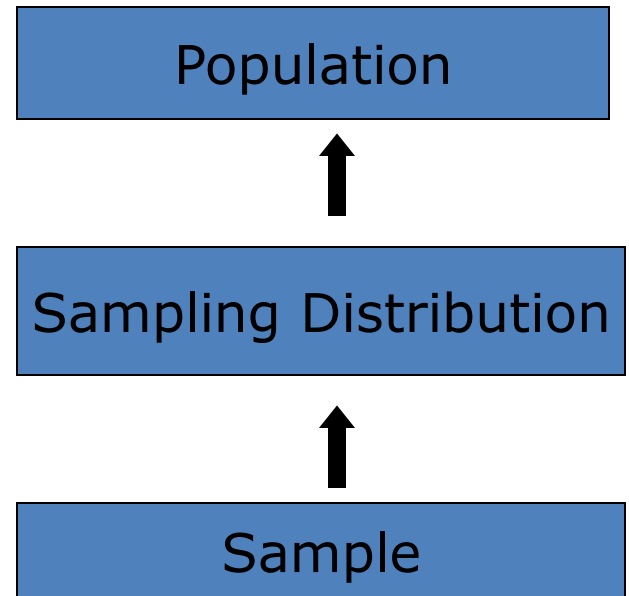
How to construct and interpret confidence intervals for:

- Sample means
- Sample proportions



# 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-empirical; known via theory
  - **Sample Distribution:**
    - empirical; known through observation



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

# 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:**

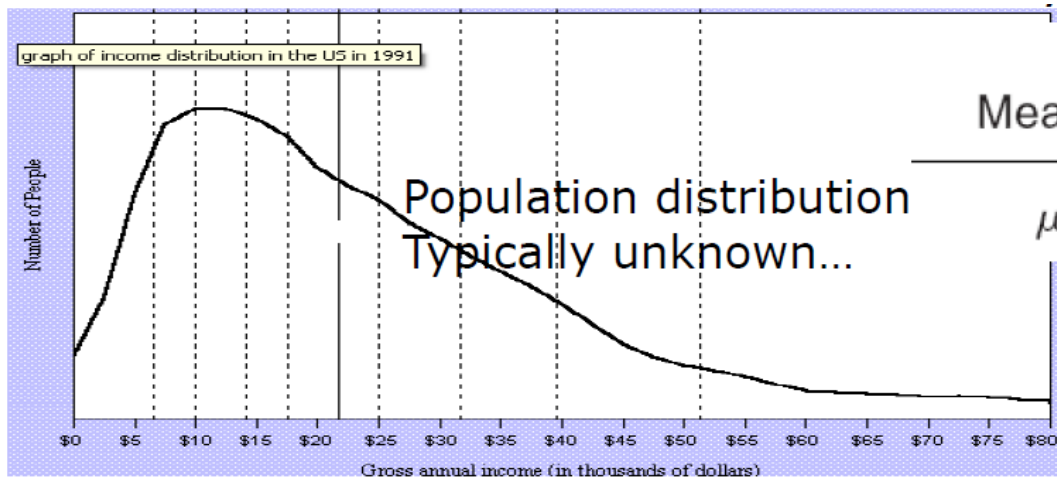
sample size and corresponding sampling error

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



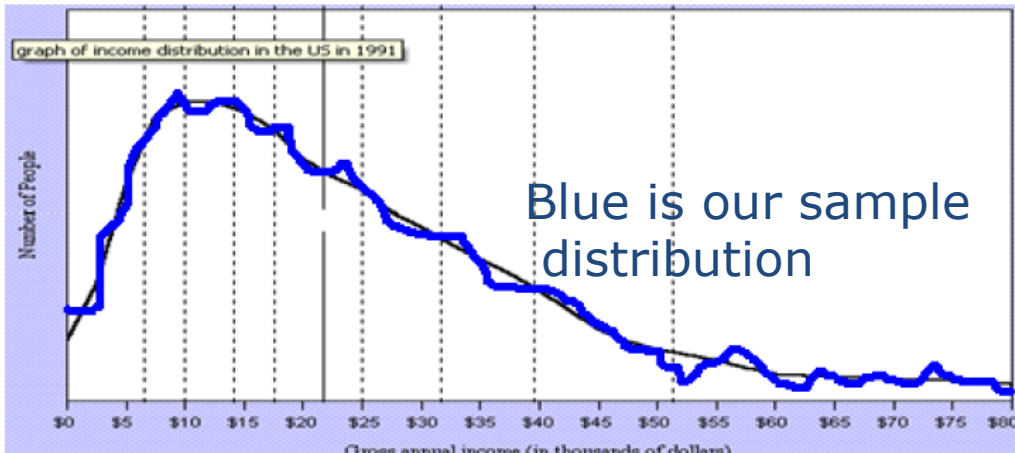
*“This is interesting, 70% of the respondents to our survey said they don't respond to surveys.”*

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



| Mean  | Standard Deviation | Proportion |
|-------|--------------------|------------|
| $\mu$ | $\sigma$           | $P_u$      |

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

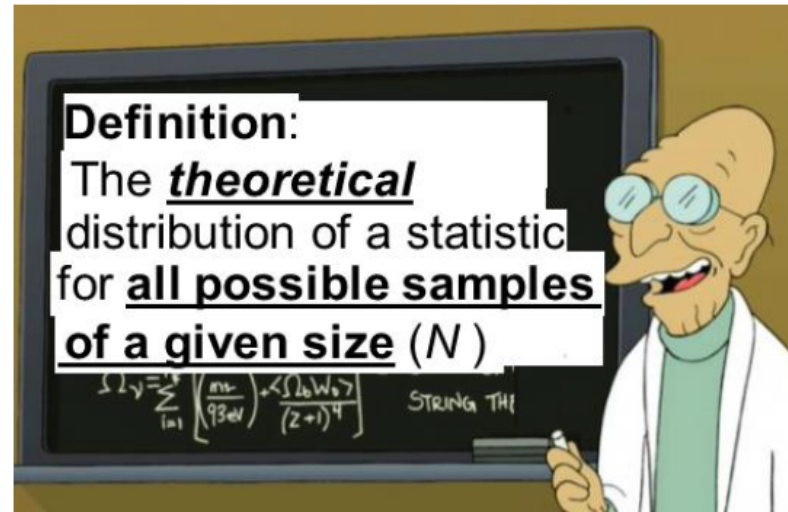


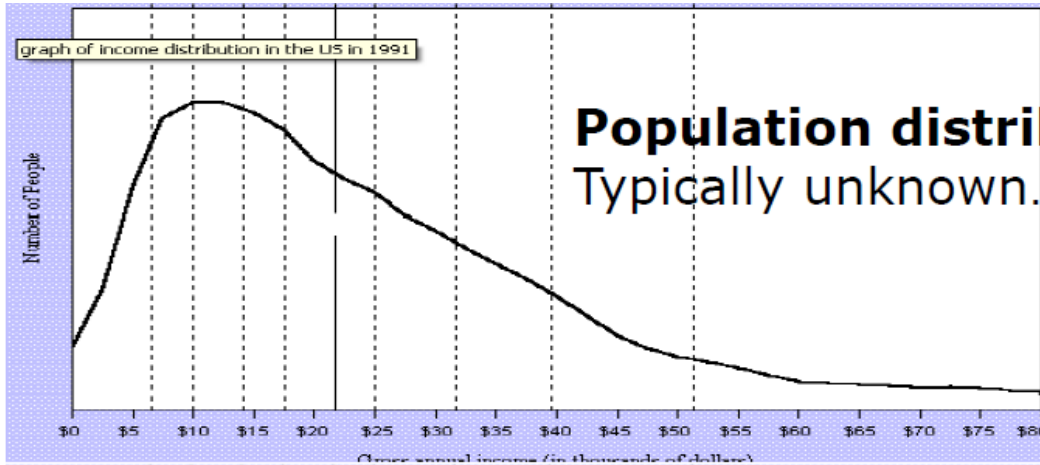
| Mean      | Standard Deviation | Proportion |
|-----------|--------------------|------------|
| $\bar{X}$ | $s$                | $P_s$      |

6-26

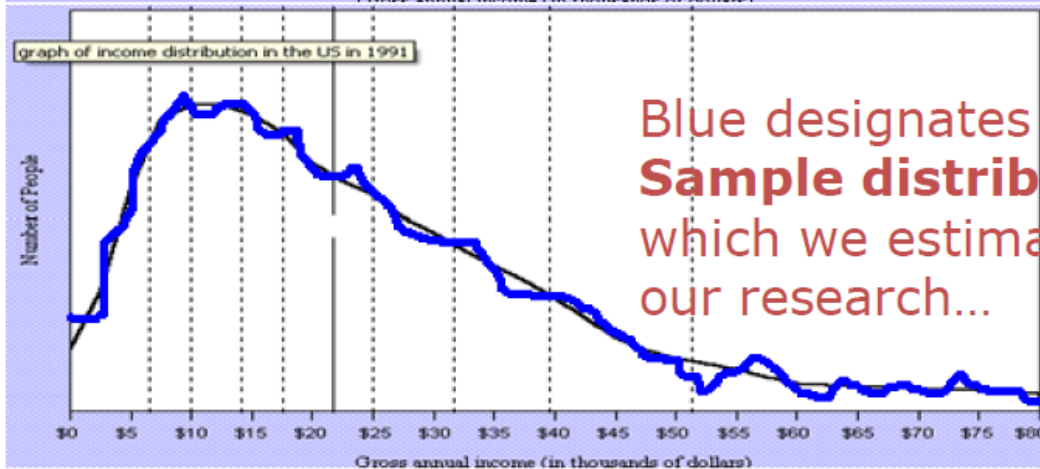
# The 3<sup>rd</sup> type of distribution: **sampling distribution**

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

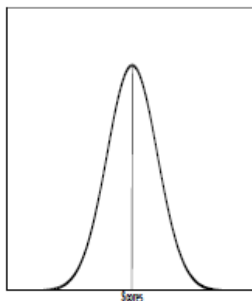




**Population distribution**  
Typically unknown...



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

**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...

Point Estimate  
**Point Estimate**

*Hi, my name is Sample Mean & I am the Point Estimate of the average weight of every puppy in*



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



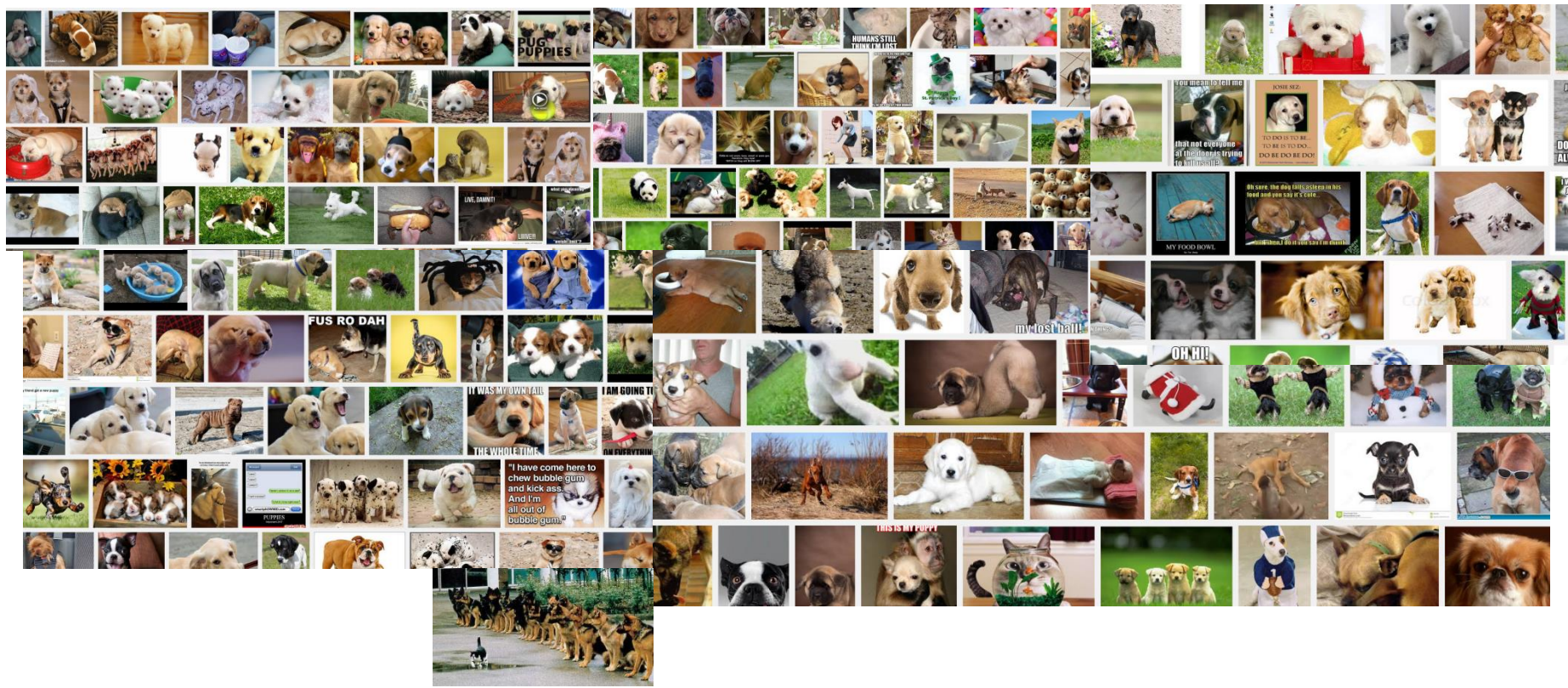
Is this point estimate based on a very tiny sample?

hence; very inefficient.. Imprecise???



Is this point estimate based on a large “representative” sample?

Hence: high 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 is somewhere between 2.2 and 2.8 pounds  
... with a given level of “probability” e.g 95% of the time!!

***Provide us some sense as to the accuracy of statistics.***

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

**NOTE:**

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

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

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

where c.i. = confidence interval

$\bar{X}$  = 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

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

3. Working with proportions

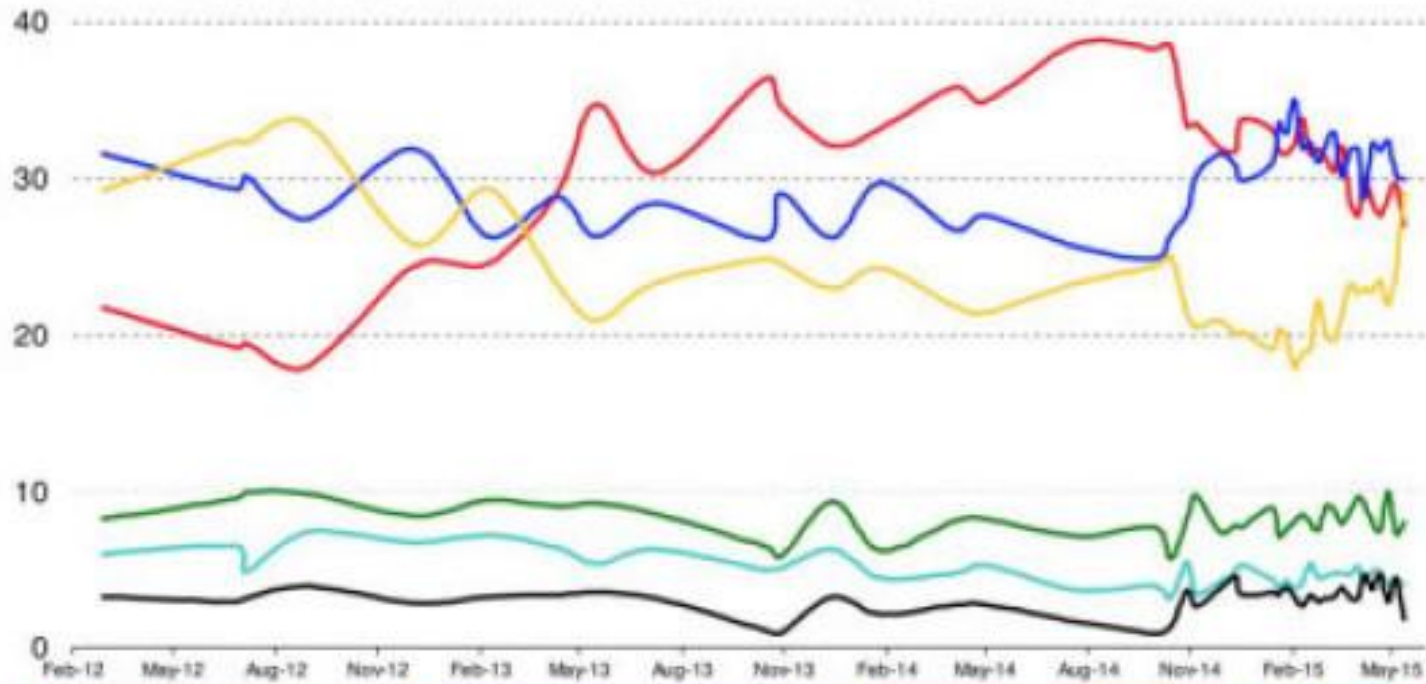
FORMULA 6.3

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

# Tracking federal vote intention



Q. If a federal election were held tomorrow, which party would you vote for?



— Liberal — Conservative — NDP — green — BLOC — Other

Note: The data on federal vote intention are based on decided and leaving voters only.

Copyright 2015

No reproduction without permission

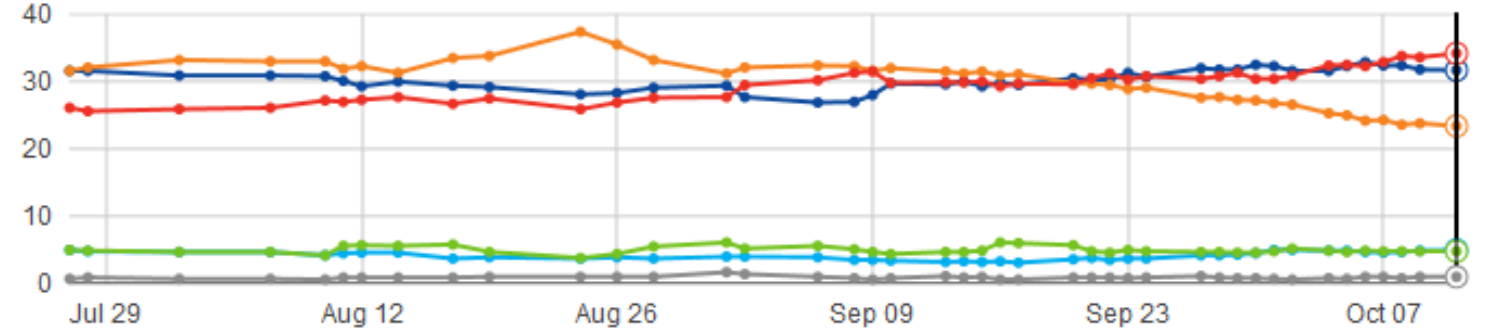
BASE: Canadians; May 6-12, 2015 (n=2,177), MOE +/- 2.1%, 19 times out of 20



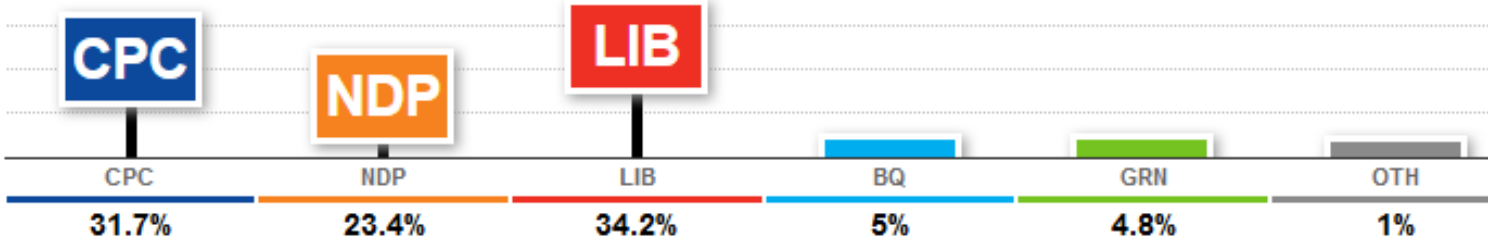
# Current election?

Oct 11, 2015

Region:



Click signs to show or hide party data



The above are point estimates. But what of the confidence intervals?

+/- 2.5% 19 times out of 20??? CPC and LIB are in a statistical dead heat

## Example of a "Confidence Interval":

In May 2015, with a sample of 1900 voters, the pollster estimated that:

30.0 % of Canadians will vote "conservative", +/- 2.25%

19 times out of 20 (or 95% of the time)... we estimate that the true Population parameter falls between 27.75% and 32.25%

Using the exact same method, yet with a sample of only 500 voters, the pollster estimated that:

30% of Canadians will vote "conservative", +/- 4.38%

19 times out of 20 (or 95% of the time), population parameter falls Between 25.62% and 34.38%

Which is preferable?

i.e. the larger sample, smaller range, more precision...



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!!!



# 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”.
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..

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)



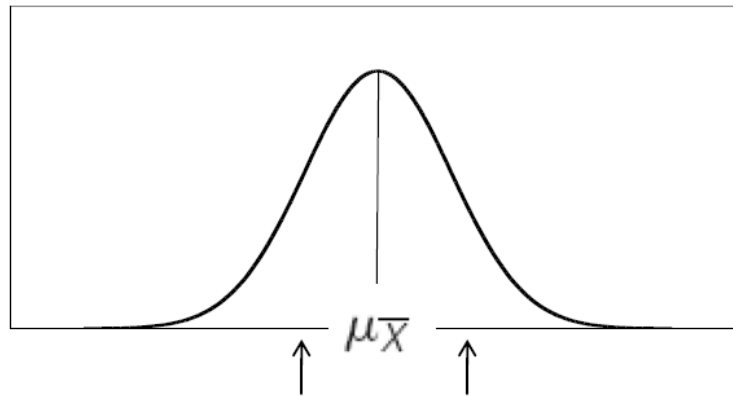
# Constructing Confidence Intervals

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

## The Sampling Distribution

Normal in shape

# of samples

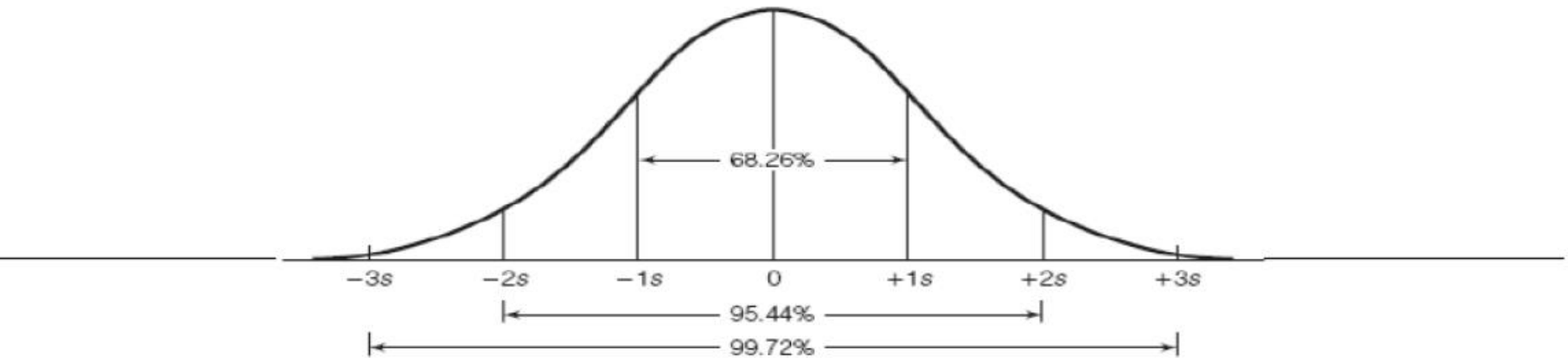


Innumerable different samples have many means

# Theoretical Normal Curve

- In Statistics we work with a “Theoretical distribution” meant to represent a normal distribution

FIGURE 5.3 AREAS UNDER THE THEORETICAL NORMAL CURVE



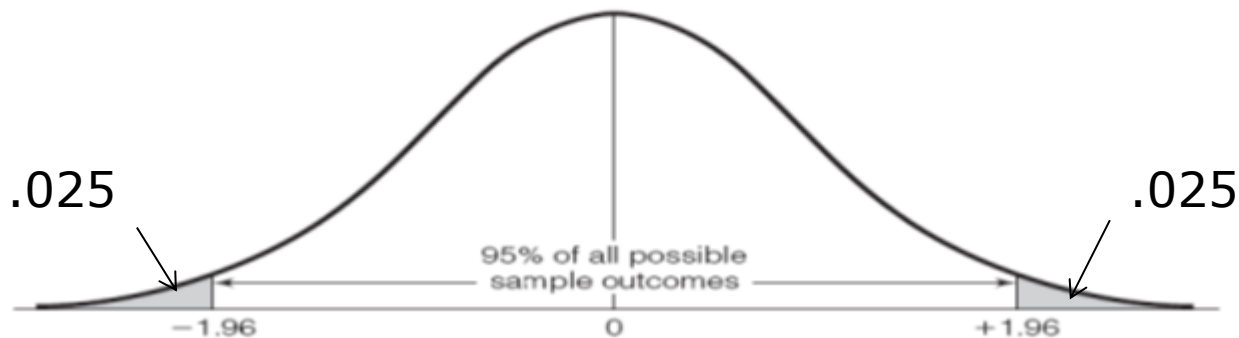
- The mean is assigned a value of 0
- The standard deviation is assigned a value of 1
- Describe this distribution in terms of Z scores (Standard score)
- A Z score of 1 is 1 standard deviation above the mean,..
- A Z score of -1 is 1 standard deviation below the mean,.. etc.

| Between                     | Lies               |
|-----------------------------|--------------------|
| $\pm 1$ standard deviation  | 68.26% of the area |
| $\pm 2$ standard deviations | 95.44% of the area |
| $\pm 3$ standard deviations | 99.72% of the area |

In a sampling distribution, about 95.44 % of all samples would fall within  $\pm 2$  standard errors (recall from last week, a Standard error is the standard deviation of sampling distribution)

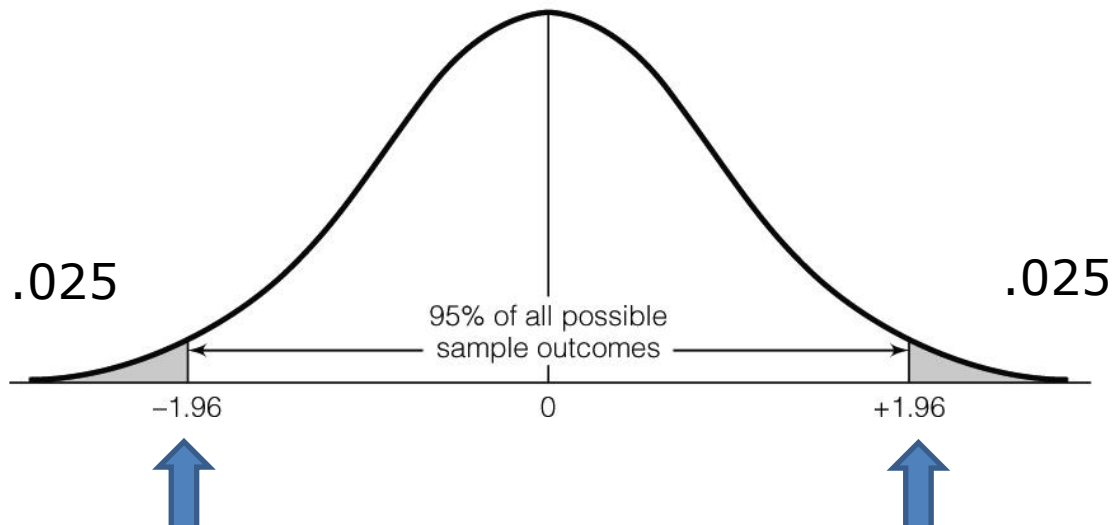
In a normal curve, what Z score would give us 95 percent of all sample outcomes? A situation whereby our sample outcome would fall within a range, 19 times out of 20???

FINDING THE Z SCORE THAT CORRESPONDS TO AN ALPHA ( $\alpha$ ) OF 0.05



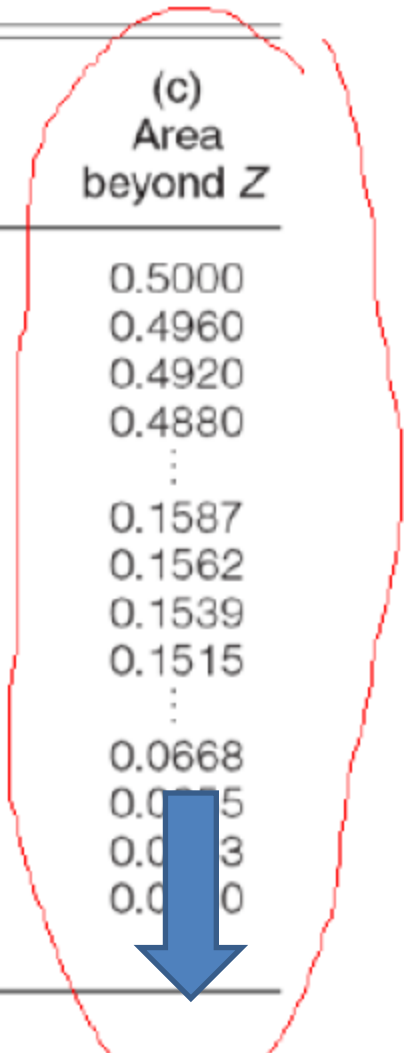
What is the appropriate Z score? Look to Appendix A, but start with Column C, find .025 and identify the corresponding Z score...

FIGURE 6.5 FINDING THE z SCORE THAT CORRESPONDS TO AN ALPHA ( $\alpha$ ) OF 0.05



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

| (a)<br>Z | (b)<br>Area between<br>Mean and Z | (c)<br>Area<br>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                            | 0.0645                  |
| 1.52     | 0.4357                            | 0.0623                  |
| 1.53     | 0.4370                            | 0.0600                  |
| ⋮        | ⋮                                 | ⋮                       |
| 1.96     |                                   | 0.0250                  |



# To be precise: 95% confidence intervals include +/- 1.96 Z scores (find it in your table)

FIGURE 7.4 THE SAMPLING DISTRIBUTION WITH ALPHA ( $\alpha$ ) EQUAL TO 0.05

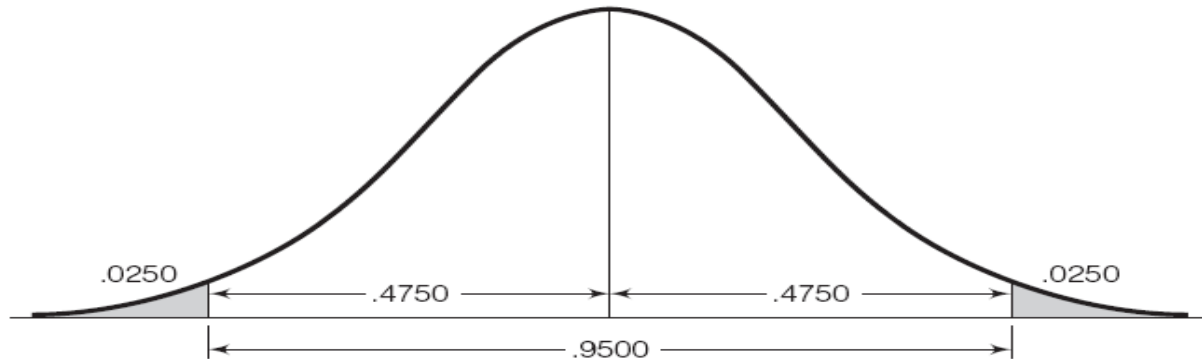
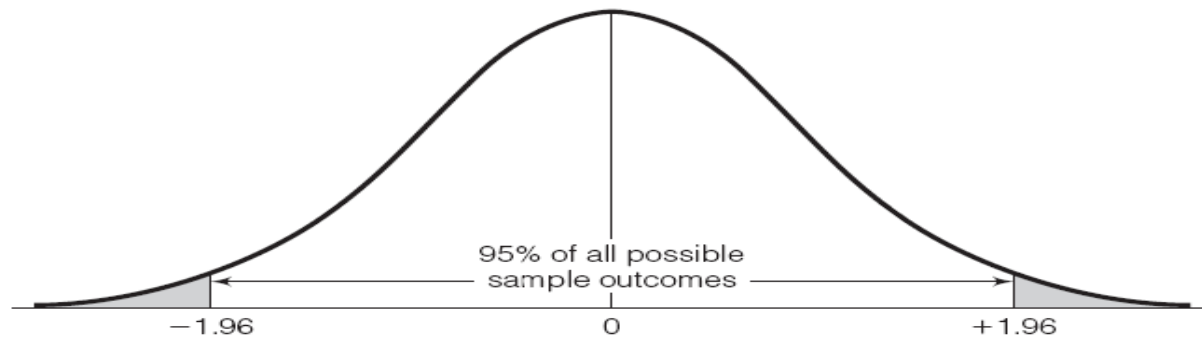


FIGURE 7.5 FINDING THE Z SCORE THAT CORRESPONDS TO AN ALPHA ( $\alpha$ ) OF 0.05



Theoretically, 95% of the **sampling distribution** falls with +/- 1.96 standard errors from the mean

## Z-values for Various Alpha Levels

| <u>Confidence Level</u> | <u><math>\alpha</math></u> | <u><math>\alpha/2</math></u> | <u>Z-score</u> |
|-------------------------|----------------------------|------------------------------|----------------|
| 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  $\alpha/2$ )



## **First set of calculation:**

### ***Constructing Confidence Intervals for Means (Population standard deviation known)***

First, set the alpha,  $\alpha$  (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  $\sigma$  known  
Formula 6.1

FORMULA 6.1

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

where c.i. = confidence interval

$\bar{X}$  = 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 ( $\sigma = 3$ ).

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

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

FORMULA 6.1

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

where c.i. = confidence interval

$\bar{X}$  = 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:

$$\text{c.i.} = 6.0 \pm 1.96(3/\sqrt{178})$$

$$\text{c.i.} = 6.0 \pm 1.96(3/13.34)$$

$$\text{c.i.} = 6.0 \pm 1.96(.22)$$

$$\text{c.i.} = 6.0 \pm .44$$

We can estimate that households in Canada average  $6.0 \pm .44$  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  $\mu$ .

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

First, set the alpha,  $\alpha$  (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  $\sigma$  known  
Formula 6.1

FORMULA 6.1

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

where c.i. = confidence interval

$\bar{X}$  = 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

**If  $\sigma$   
unknown  
Formula  
6.2**

FORMULA 6.2

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

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

**IMPORTANT POINT!**

In Formula 6.2  $\sigma$  is replaced by  $s$ . Further,  $n$  is replaced by  $n-1$  to correct for the fact that  $s$  is a biased estimator of  $\sigma$ .

# 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 ( $s=3$ ).

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

$$\text{c.i.} = 2.5 \pm 1.96(3/\sqrt{500-1})$$

$$\text{c.i.} = 2.5 \pm 1.96(3/22.34)$$

$$\text{c.i.} = 2.5 \pm 1.96(.1343)$$

$$\text{c.i.} = 2.5 \pm .26$$

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



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  $\mu$ .

# 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  $P_u$**



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

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

$$\text{c.i.} = .22 \pm 1.96 \sqrt{.5(1-.5)/764}$$

$$\text{c.i.} = .22 \pm 1.96 (\sqrt{.25/764})$$

$$\text{c.i.} = .22 \pm 1.96 (\sqrt{.00033})$$

$$\text{c.i.} = .22 \pm 1.96 (.018)$$

$$\text{c.i.} = .22 \pm .04$$

Changing back to %'s, we can estimate that  $22\% \pm 4\%$  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


| <u>Confidence Level</u> | <u><math>\alpha</math></u> | <u><math>\alpha/2</math></u> | <u>Z-score</u> |
|-------------------------|----------------------------|------------------------------|----------------|
| 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  $\alpha/2$ )

# 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%              | $\$35,000 \pm 14.77$ | \$29.54        |
| .05   | 95%              | $\$35,000 \pm 17.55$ | \$35.10        |
| .01   | 99%              | $\$35,000 \pm 23.09$ | \$46.18        |
| .001  | 99.9%            | $\$35,000 \pm 29.45$ | \$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%              | $\$35,000 \pm 14.77$ | \$29.54        |
| .05   | 95%              | $\$35,000 \pm 17.55$ | \$35.10        |
| .01   | 99%              | $\$35,000 \pm 23.09$ | \$46.18        |
| .001  | 99.9%            | $\$35,000 \pm 29.45$ | \$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. = $\$35,000 \pm 1.96(200/\sqrt{99})$  |        | c.i. = $\$35,000 \pm 1.96(200/\sqrt{499})$   |  |
| c.i. = $\$35,000 \pm 39.40$                |        | c.i. = $\$35,000 \pm 17.55$                  |  |
| Sample 3 ( $N = 1,000$ )                   |        | Sample 4 ( $N = 10,000$ )                    |  |
| c.i. = $\$35,000 \pm 1.96(200/\sqrt{999})$ |        | c.i. = $\$35,000 \pm 1.96(200/\sqrt{9,999})$ |  |
| c.i. = $\$35,000 \pm 12.40$                |        | c.i. = $\$35,000 \pm 3.92$                   |  |
| 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. =  $.32 \pm 1.96 (\sqrt{.25/1000})$
  - c.i. =  $.32 \pm 1.96 (.0158)$
  - c.i. =  $.32 \pm .03$
  - 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 residents will vote in this way?
  - c.i. =  $.32 \pm 1.96 \sqrt{.25/250}$
  - c.i. =  $.32 \pm .062$
  - 95% chance, between .258 and .382 or 25.8% and 38.2%
  - What if a second party had 30%?
  - C.i. =  $.30 \pm .062$
  - 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 95% CI on this estimate..