STAT 234 Lecture 16B
Overview of Confidence Intervals
Section 8.1

Yibi Huang
Department of Statistics
University of Chicago

Section 8.1 Overview of Confidence Intervals

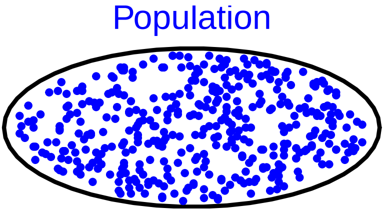
- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a sample statistic (point estimate) to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

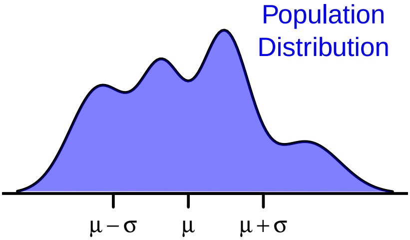
We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny (http://www.flickr.com/photos/clearlydived/7029109617) on Flickr.

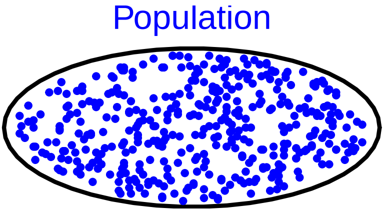
Variability in Estimation (Review)

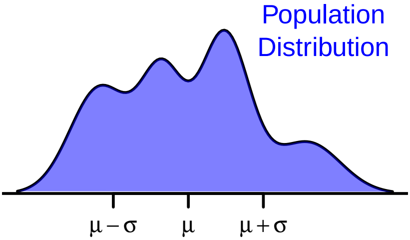




- Suppose we are interested in some numerical characteristic X about individuals in a certain population.

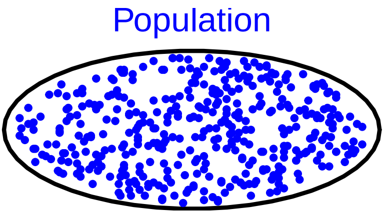
Variability in Estimation (Review)

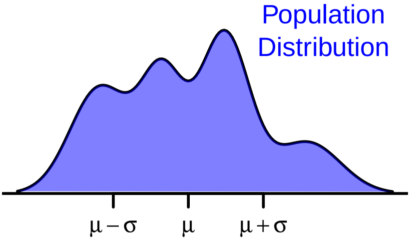




- Suppose we are interested in some numerical characteristic X about individuals in a certain population.
- If it's possible to interview each individual in the population and record his/her X value, we can then make a histogram for the recorded X-values and that's the *population distribution*.

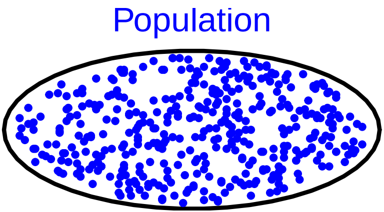
Variability in Estimation (Review)

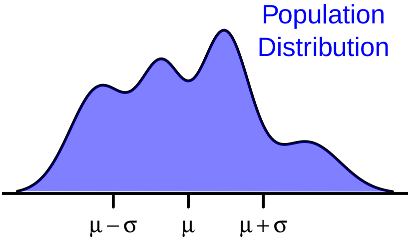




- Suppose we are interested in some numerical characteristic X about individuals in a certain population.
- If it's possible to interview each individual in the population and record his/her X value, we can then make a histogram for the recorded X-values and that's the *population distribution*.
- The population distribution is arbitrary (not necessarily normal), with a *population mean* μ and a *population SD* σ.

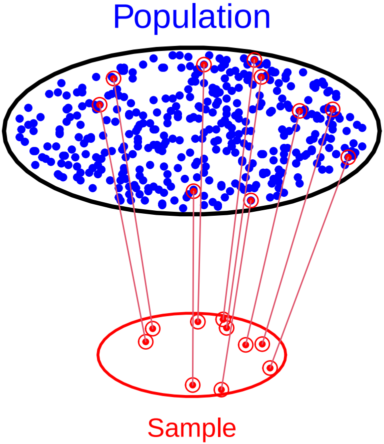
Variability in Estimation (Review)

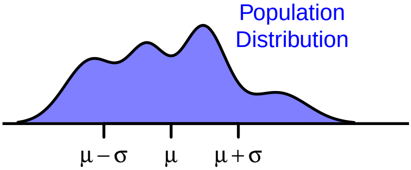




- Suppose we are interested in some numerical characteristic X about individuals in a certain population.
- If it's possible to interview each individual in the population and record his/her X value, we can then make a histogram for the recorded X-values and that's the *population distribution*.
- The population distribution is arbitrary (not necessarily normal), with a *population mean* μ and a *population SD* σ.
- The goal is to estimate the population mean μ

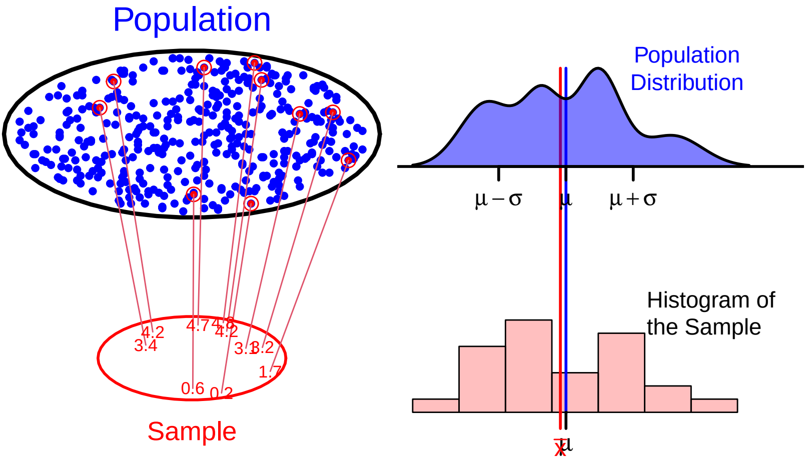
Variability in Estimation (Review)





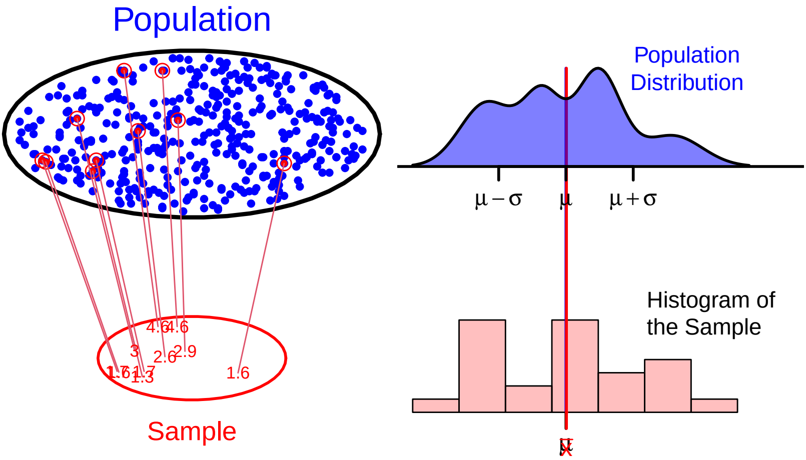
A (simple) random sample is taken from the population.

Variability in Estimation (Review)



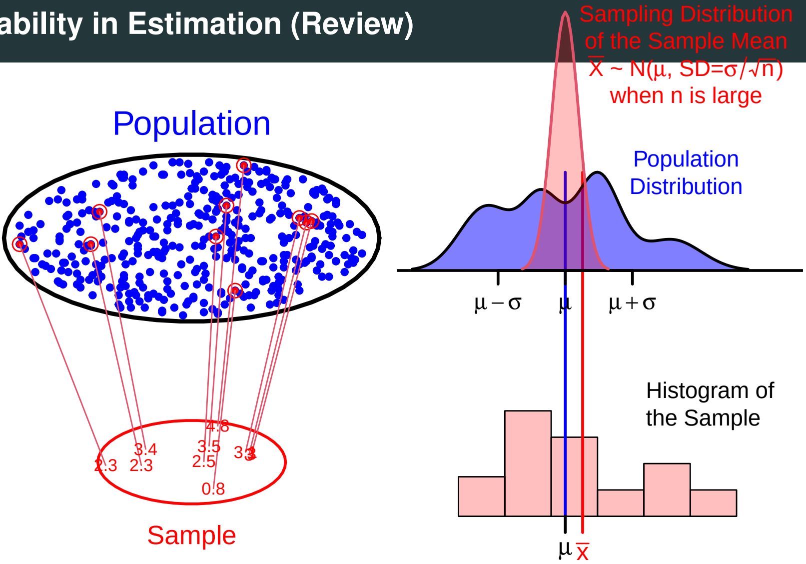
The X-value for each individual in the sample is recorded. One can make a histogram for the recorded X-values.

Variability in Estimation (Review)



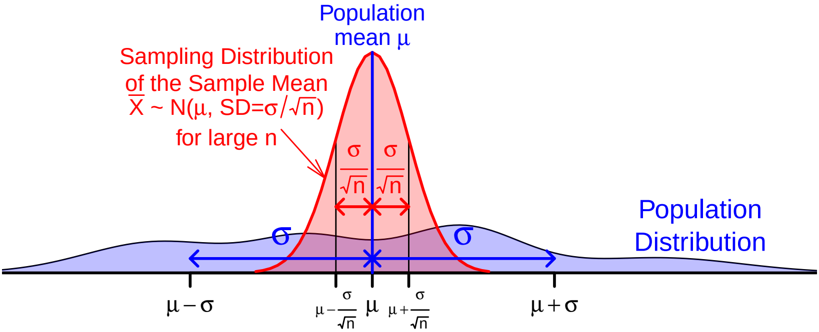
The population mean μ is estimated by the sample mean $\overline{X}$, which will change from sample to sample.

Variability in Estimation (Review)



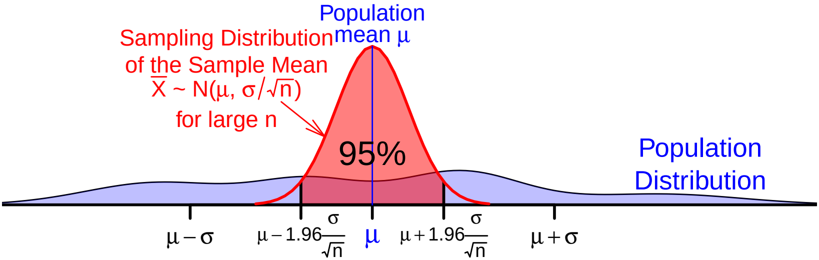
The distribution of the sample mean $\overline{X}$ is approx. normal w/ mean μ and SD $= \sigma/\sqrt{n}$ when n is large by CLT.

σ v.s. $\sigma/\sqrt{n}$

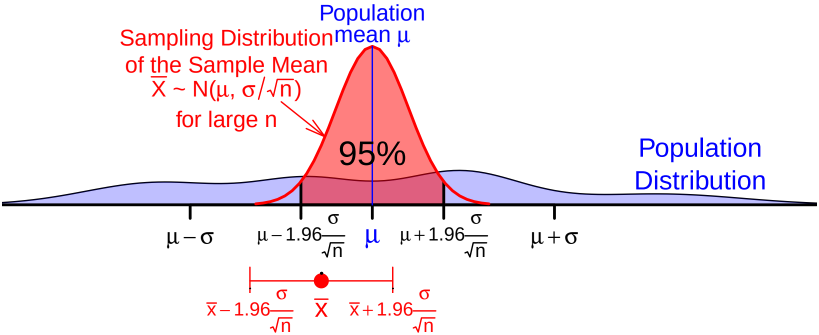


- σ is the SD of the population
- $\frac{\sigma}{\sqrt{n}}$ is the SD of the sampling distribution of $\overline{X}$
    - $\frac{\sigma}{\sqrt{n}}$ is usually called the *standard error (SE)*, to differentiate it from the population SD σ

As a normal random variable will fall within 1.96 SDs from the center 95% of the time, $\overline{X}$ **will fall within** $1.96\frac{\sigma}{\sqrt{n}}$ **from** μ **95% of the time** since $\overline{X}$ is approx. $N(\mu, \sigma/\sqrt{n})$ for large n by CLT.



As a normal random variable will fall within 1.96 SDs from the center 95% of the time, $\overline{X}$ **will fall within** $1.96\frac{\sigma}{\sqrt{n}}$ **from** μ **95% of the time** since $\overline{X}$ is approx. $N(\mu, \sigma/\sqrt{n})$ for large n by CLT.



Or equivalently, μ **will be within** $1.96\frac{\sigma}{\sqrt{n}}$ **from** $\overline{X}$ **95% of the time**.

A **95% confidence interval for** μ is hence defined to be

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}} = \left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Procedures to Construct a 95% Confidence Interval for μ

1. Take a simple random sample (or i.i.d. sample) of some large enough size n and find the sample mean $\overline{X}$.
2. If n is large, the 95% confidence interval for μ is given by

$$\overline{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

But σ is usually unknown ...

But σ is usually unknown ...

The unknown population SD σ is replaced by our best guess — *the sample SD* s. So an approximate 95% confidence interval for μ is

$$\overline{X} \pm 1.96 \frac{s}{\sqrt{n}}$$

- However, this replacement is hazardous because
  - s is a poor estimate of σ when the sample size n is small and
  - s is very **sensitive to outliers**
- So we require $n \geq 30$ and sample shouldn't have any outlier nor be too skewed ⇒ Need to check histogram of the data
- We will discuss working with samples where $n < 30$ in the next chapter

Example: Average Number of Exclusive Relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\overline{X} = 3.2 \qquad s = 1.74$$

The approximate 95% confidence interval is about

$$\begin{aligned} \overline{X} \pm 1.96 \times \text{SE} &= \overline{X} \pm 1.96 \times \frac{s}{\sqrt{n}} \\ &= 3.2 \pm 1.96 \times \frac{1.74}{\sqrt{50}} \\ &\approx 3.2 \pm 0.5 = (2.7, 3.7) \end{aligned}$$

True or False

True or False and explain: We are 95% confident that the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

True or False

True or False and explain: We are 95% confident that the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

False. The confidence interval $\overline{X} \pm$ 1.96 SE definitely (100%) contains the *sample mean* $\overline{X}$, not just with probability 95%.

True or False

True or False and explain: We are 95% confident that the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

False. The confidence interval $\overline{X} \pm$ 1.96 SE definitely (100%) contains the *sample mean* $\overline{X}$, not just with probability 95%.

True or False and explain: 95% of college students have been in 2.7 to 3.7 exclusive relationships.

True or False and explain: We are 95% confident that the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

False. The confidence interval $\overline{X} \pm$ 1.96 SE definitely (100%) contains the *sample mean* $\overline{X}$, not just with probability 95%.

True or False and explain: 95% of college students have been in 2.7 to 3.7 exclusive relationships.

False. The confidence interval is for covering the population mean μ, not for covering 95% of the entire population. If 95% of college students have been in 2.7 to 3.7 exclusive relationships, the SD won't be as large as 1.74.

True or False

True or False and explain: There is 0.95 probability that the true mean number of exclusive relationships of college students falls in the interval $(2.7, 3.7)$

True or False

True or False and explain: There is 0.95 probability that the true mean number of exclusive relationships of college students falls in the interval $(2.7, 3.7)$

True or False and explain: The interval $(2.7, 3.7)$ has probability of 0.95 of enclosing the true mean number of exclusive relationships of college students.

True or False

True or False and explain: There is 0.95 probability that the true mean number of exclusive relationships of college students falls in the interval $(2.7, 3.7)$

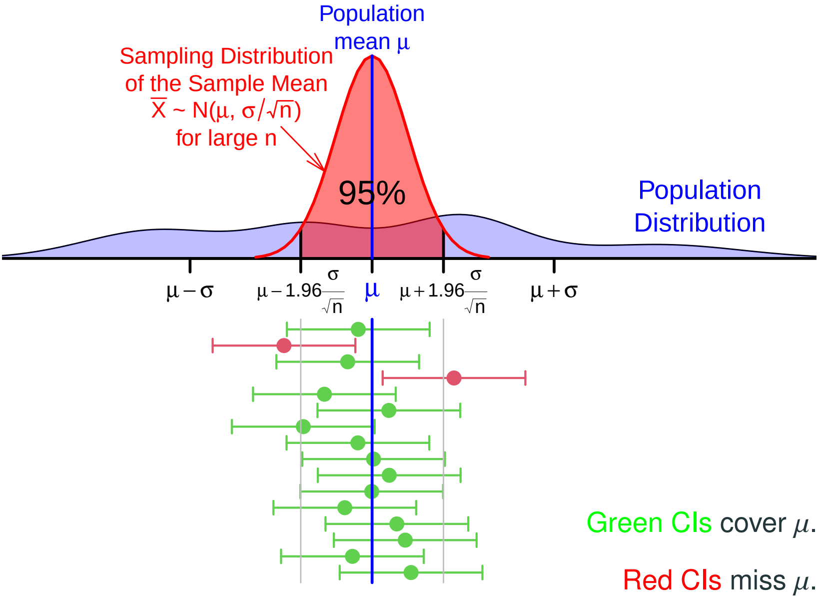
True or False and explain: The interval $(2.7, 3.7)$ has probability of 0.95 of enclosing the true mean number of exclusive relationships of college students.

Both are False. The population mean μ is a fixed number, not random. It is either in the interval (2.7, 3.7), or not in the interval. There is no uncertainty involved.

What does “95% confidence” mean?

What is the thing that has a 95% chance to happen?

- It is the **procedure to construct the 95% interval**.
- About 95% of the intervals constructed following the procedure (taking a SRS and then calculating $\overline{X} \pm 1.96\, s/\sqrt{n}$) will cover the true population mean μ.
- After taking the sample and an interval is constructed, the constructed interval either covers μ or it doesn’t. We don’t know. Only God knows.
- Just like lottery, before you pick the numbers and buy a lottery ticket, you have some chance to win the prize. After you get the ticket, you either win or lose.


Population
mean μ
Sampling Distribution
of the Sample Mean
$\overline{X} \sim N(\mu, \sigma/\sqrt{n})$
for large n
95%
Population
Distribution
$\mu - \sigma$
$\mu - 1.96\frac{\sigma}{\sqrt{n}}$
μ
$\mu + 1.96\frac{\sigma}{\sqrt{n}}$
$\mu + \sigma$
Green CIs cover μ.
Red CIs miss μ.

True or False

True or False and explain: If a new random sample of size 50 is taken, we are 95% confident that the new sample mean will be between 2.7 and 3.7.

True or False

True or False and explain: If a new random sample of size 50 is taken, we are 95% confident that the new sample mean will be between 2.7 and 3.7.

False. The confidence interval is for covering the population mean μ, not for covering the mean of another sample. The SE $\sigma/\sqrt{n}$ or $s/\sqrt{n}$ is a typical distance between the sample mean and population mean, not a typical distance between two sample means.

True or False

True or False and explain: This confidence interval $\overline{X} \pm 1.96\, s/\sqrt{n}$ is not valid since the number of exclusive relationships is integer-valued. Neither the population nor sample is normally distributed.

True or False

True or False and explain: This confidence interval $\overline{X} \pm 1.96\, s/\sqrt{n}$ is not valid since the number of exclusive relationships is integer-valued. Neither the population nor sample is normally distributed.

False. The construction of the CI $\overline{X} \pm 1.96\, s/\sqrt{n}$ only uses the normality of the sampling distribution of the sample mean. Neither the population nor the sample is required to be normally distributed. By the central limit theorem, with a large enough sample size we can assume that the sampling distribution is nearly normal and calculate a confidence interval.

Confidence Intervals at Other Confidence Levels

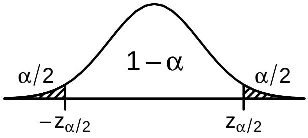
A confidence interval for a population mean μ at confidence level $(1 - \alpha)$ is

$$\text{sample mean} \pm z_{\alpha/2}\text{SE}$$

where $z_{\alpha/2}$ is a number such that

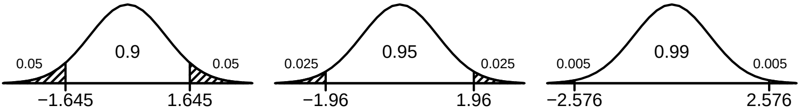
$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha \quad \text{or}$$

where $Z \sim N(0, 1)$.



Commonly used confidence levels:

90% CI, $\alpha = 0.1$	95% CI, $\alpha = 0.05$	99% CI, $\alpha = 0.01$
$z_{0.1/2} \approx 1.645$	$z_{0.05/2} \approx 1.960$	$z_{0.01/2} \approx 2.576$



Example

For the “number of exclusive relationships” example, recall

$$\overline{X} = 3.2, \quad s = 1.74, \quad \text{SE} = \frac{s}{\sqrt{n}} = \frac{1.74}{\sqrt{50}} \approx 0.246$$

- 90% CI: $\overline{X} \pm \mathit{1.645} \times \text{SE} = 3.2 \pm 1.645 \times 0.246 \approx 3.2 \pm 0.40$
- 95% CI: $\overline{X} \pm \mathit{1.96} \times \text{SE} = 3.2 \pm 1.96 \times 0.246 \approx 3.2 \pm 0.48$
- 99% CI: $\overline{X} \pm \mathit{2.576} \times \text{SE} = 3.2 \pm 2.576 \times 0.246 \approx 3.2 \pm 0.63$

How to Choose the Confidence Level?

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a shorter interval?

How to Choose the Confidence Level?

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a shorter interval?

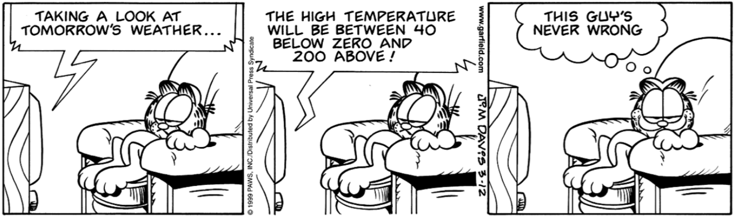
A wider interval.

How to Choose the Confidence Level?

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a shorter interval?

A wider interval.

Can you see any drawbacks to using a wider interval?

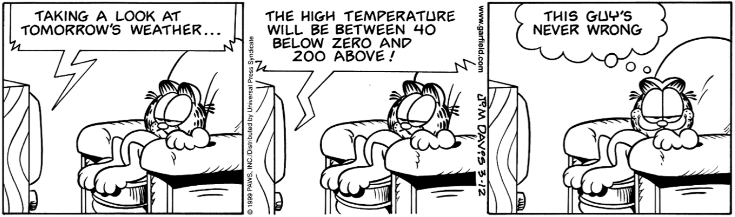


How to Choose the Confidence Level?

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a shorter interval?

A wider interval.

Can you see any drawbacks to using a wider interval?



A wide interval may not be informative.

Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif