# STA258H5
# University of Toronto Mississauga

Al Nosedal and Omid Jazi

Winter 2023

# STA258H5
# University of Toronto Mississauga

Al Nosedal and Omid Jazi

Winter 2023

## Population

A group of interest (typically large)

## Sample

A subset of a population

## Parameter (of population)

A numerical characteristic of a population

$\mu$ : population mean
$\sigma^2$ : " variance
($\sigma$ : " st. dev)

} in a real life setting are usually unknown

[Note: Different to parameter of a distribution]

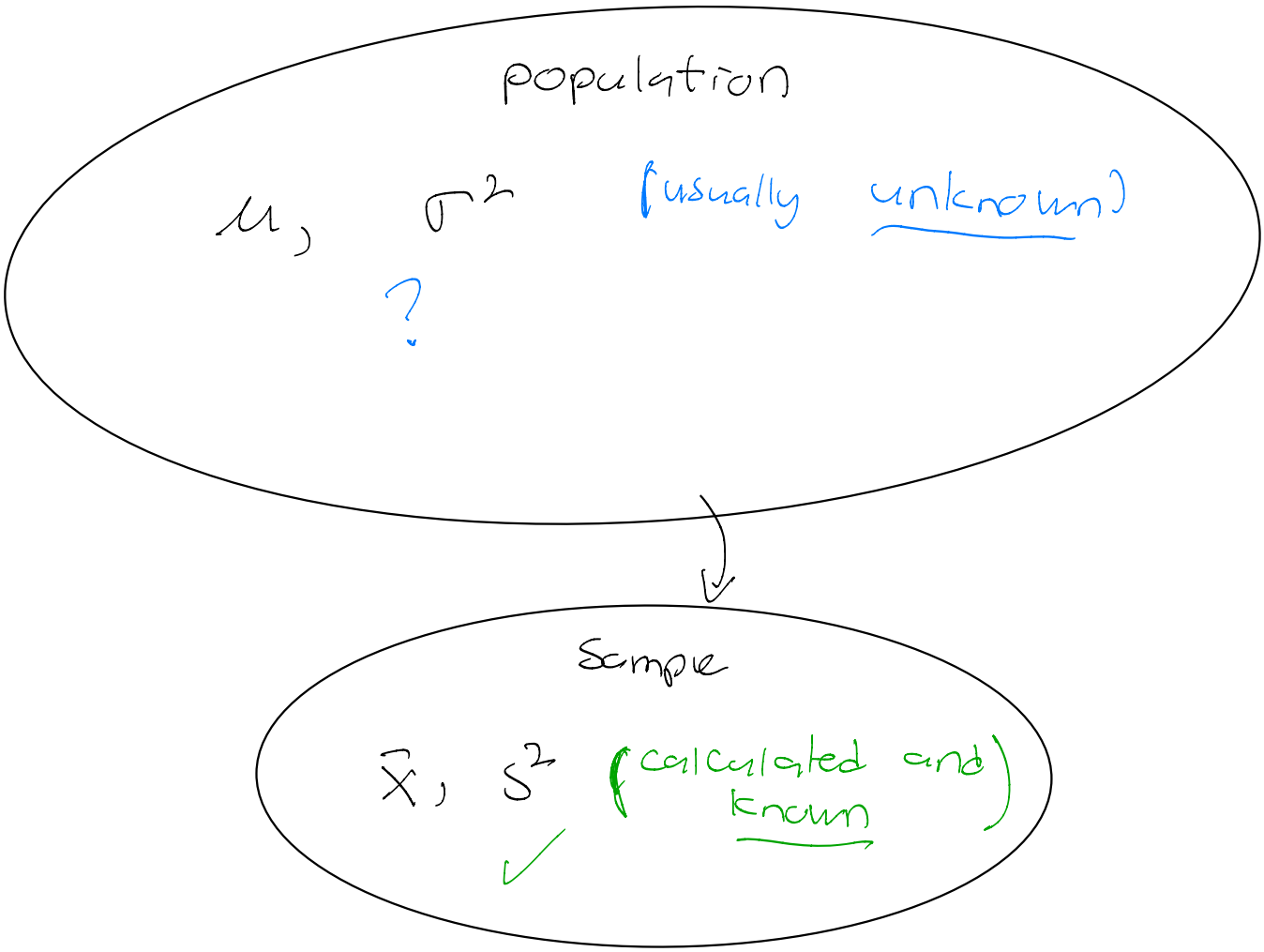
## Statistic

A numerical characteristic of a Sample

$\bar{x}$ : Sample mean
$s^2$ : " variance
($s$ : " st. dev)

} calculated and known (function of data)

High Level Picture




## Statistical Inference

use statistics (known) to make conclusions on parameters (unknown) and quantify the degree of certainty of statements made.

# ONE SAMPLE CONFIDENCE INTERVALS ON A MEAN WHEN THE POPULATION VARIANCE IS KNOWN

# Point and Interval Estimates

The sample mean, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, is a number we use to estimate the population mean, $\mu$. This is called a **point estimate**.

But, we know it's not equal to $\mu$. Then, we'd rather estimate the population mean using an **interval estimate** that gives a range of real numbers that we hope contains the population mean, $\mu$.

## Examples

$\bar{X}$ is a point estimate of $\mu$

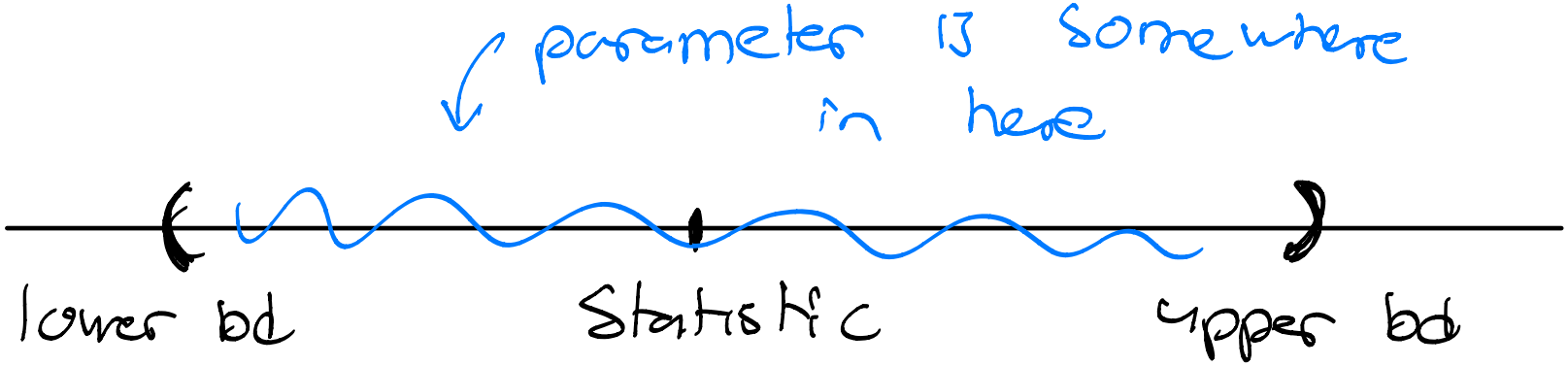
$s^2$ " " " " " $\sigma^2$

$s$ " " " " " $\sigma$

↑ calculated with data from a sample

Due to nature of randomness and calculating based on a subset, statistics are not guaranteed to be exactly equal to parameters.

Therefore we create intervals around statistics which we believe captures the parameter




## Confidence Interval

A range of values we believe captures a parameter with a certain level of confidence.

## Skeleton

z or t

st. dev of the sampling dist

Estimator (statistic) $\pm$ ( Value from a reference distrib ) $\times$ ( Standard error of estimate )

Margin of error

The exact form depends on parameter of interest and information available

## Interpretation of CI's

eg 95%

Suppose we construct a C% confidence interval

## Intuitive Interpretation

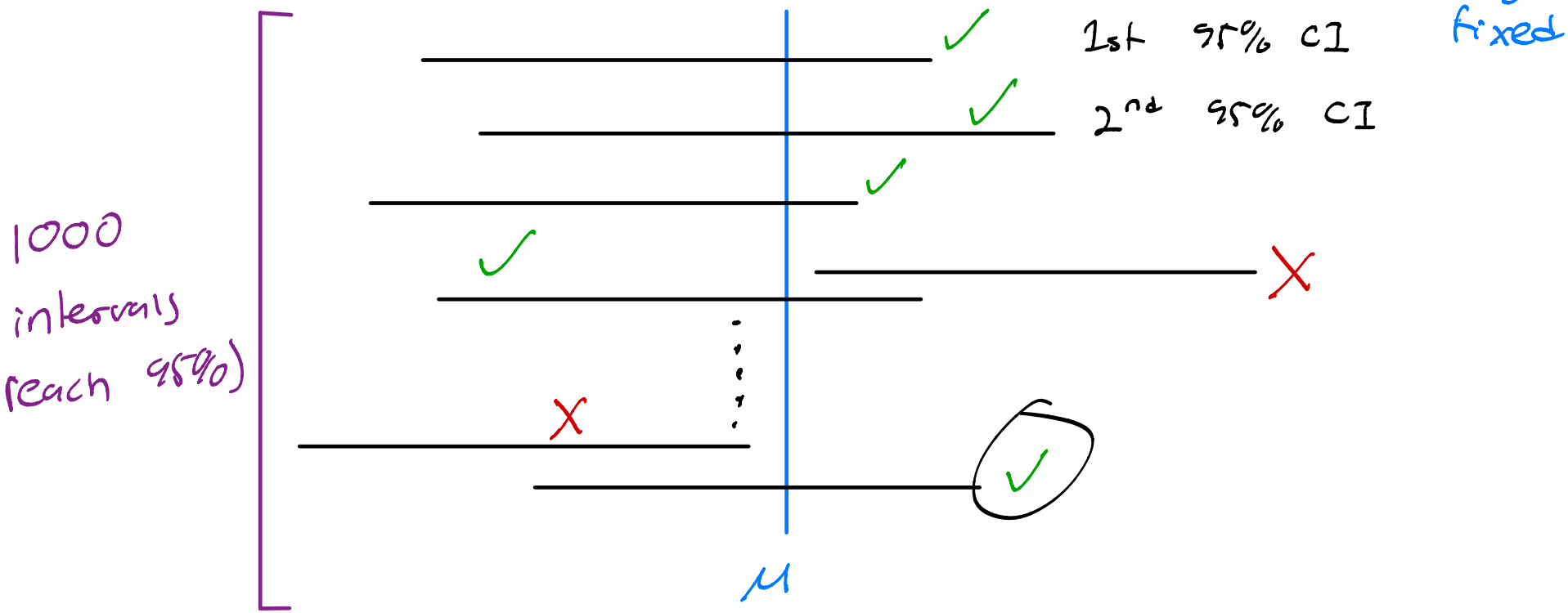
we are C% confident the target parameter is inside the CI constructed.

## Formal Definition

In repeated sampling, approximately C% out of all the C% CI's constructed captures the parameter

(See slide 11)

Suppose we construct 1000 95% CI's for $\mu$ (fixed)




Approximately 95% of the 1000 intervals (i.e approx 950) capture $\mu$.

# Confidence Interval for $\mu$

CONFIDENCE INTERVAL FOR THE MEAN (NORMAL SAMPLE, VARIANCE KNOWN).

Let $X_1, X_2, ..., X_n$ be iid $N(\mu, \sigma^2)$, where $\mu$ is unknown and $\sigma$ is known.

We know that $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

We also know that $P(-1.96 < Z < 1.96) = 0.95$.

$$\Rightarrow \quad P(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96) = 0.95.$$

$$\Rightarrow \quad P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- This is a random interval $\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$
- The interval is random since $\bar{X}$ is random due to sampling.
- The population mean $\mu$ is a fixed, but unknown, number.
- The probability $\mu$ is inside the random interval is 0.95. You can think about it as the success rate for the method.
- 95% of all samples give an interval that captures $\mu$, and 5% of all samples give an interval that does not capture $\mu$.

Once we observe our sample,

- This is NOT a random interval $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$
- The probability $\mu$ is inside this interval is either 1 or 0.

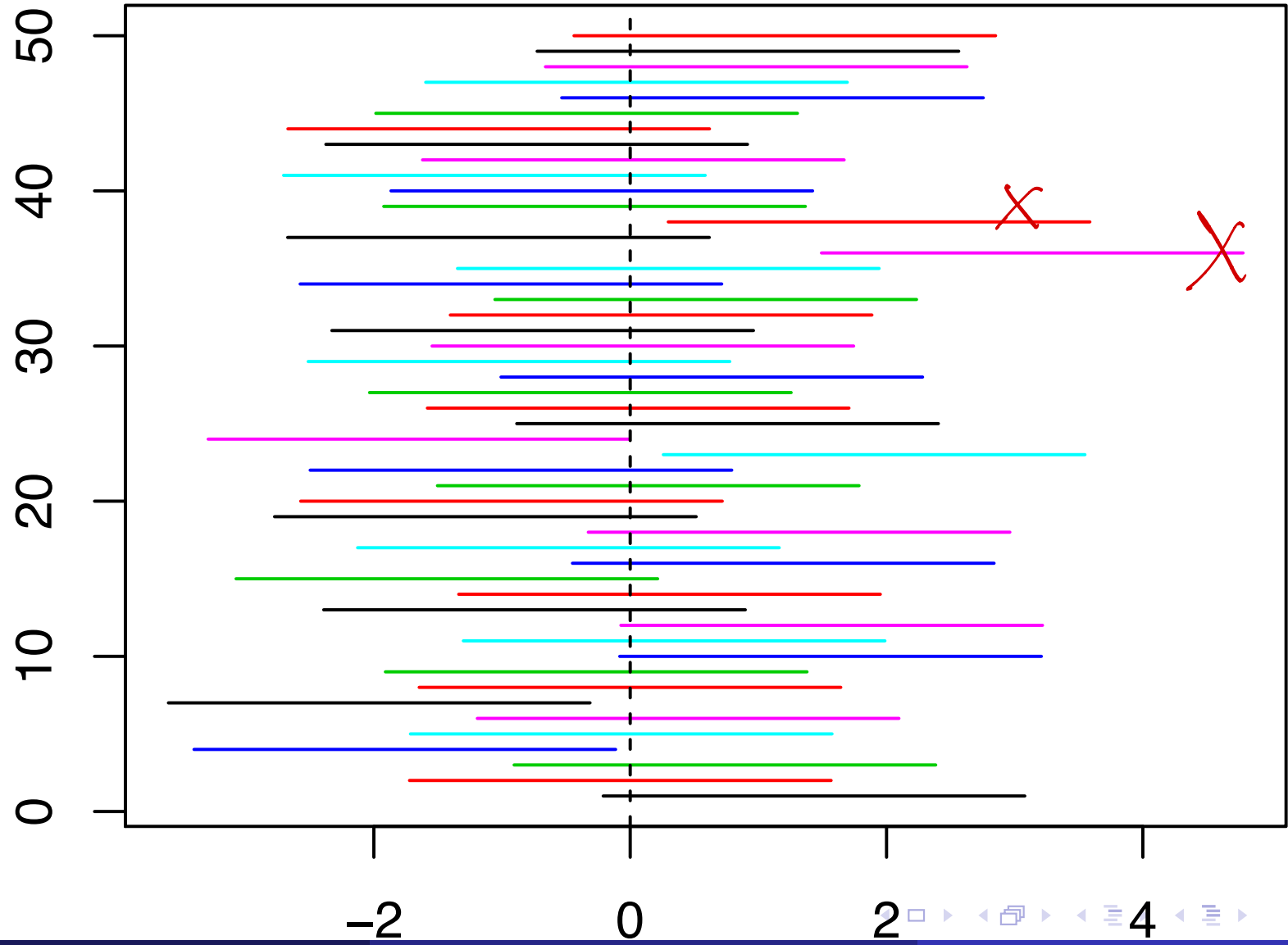
# Confidence interval isn't always right

The fact that not all confidence intervals contain the true value of the parameter is often illustrated by plotting a number of random confidence intervals at once and observing. Let's do it!

```r
## Step 1. Generate random samples;

set.seed(2017)
m=50;
# m= number of samples;
n=25;
# n= number of obs in sample;
mu.i=0;
# mu.i = mean of obs;
sigma.i=5;
# sigma.i =std. dev. of obs;
mu.total=n*mu.i;
# mean of Total;
sigma.total=sqrt(n)*sigma.i;
# std. dev. of Total;
```

```
## Step 2. Construct CIs;

xbar=rnorm(m,mu.total,sigma.total)/n;

SE=sigma.i/sqrt(n);

alpha=0.10;

z.star=qnorm(1-alpha/2);
```

```
## Step 3. Graph CIs;

matplot(rbind(xbar-z.star*SE,xbar+z.star*SE),rbind(1:m,1:m),
type="l",lty=1, xlab=" ",ylab=" ");

abline(v=0,lty=2);
```

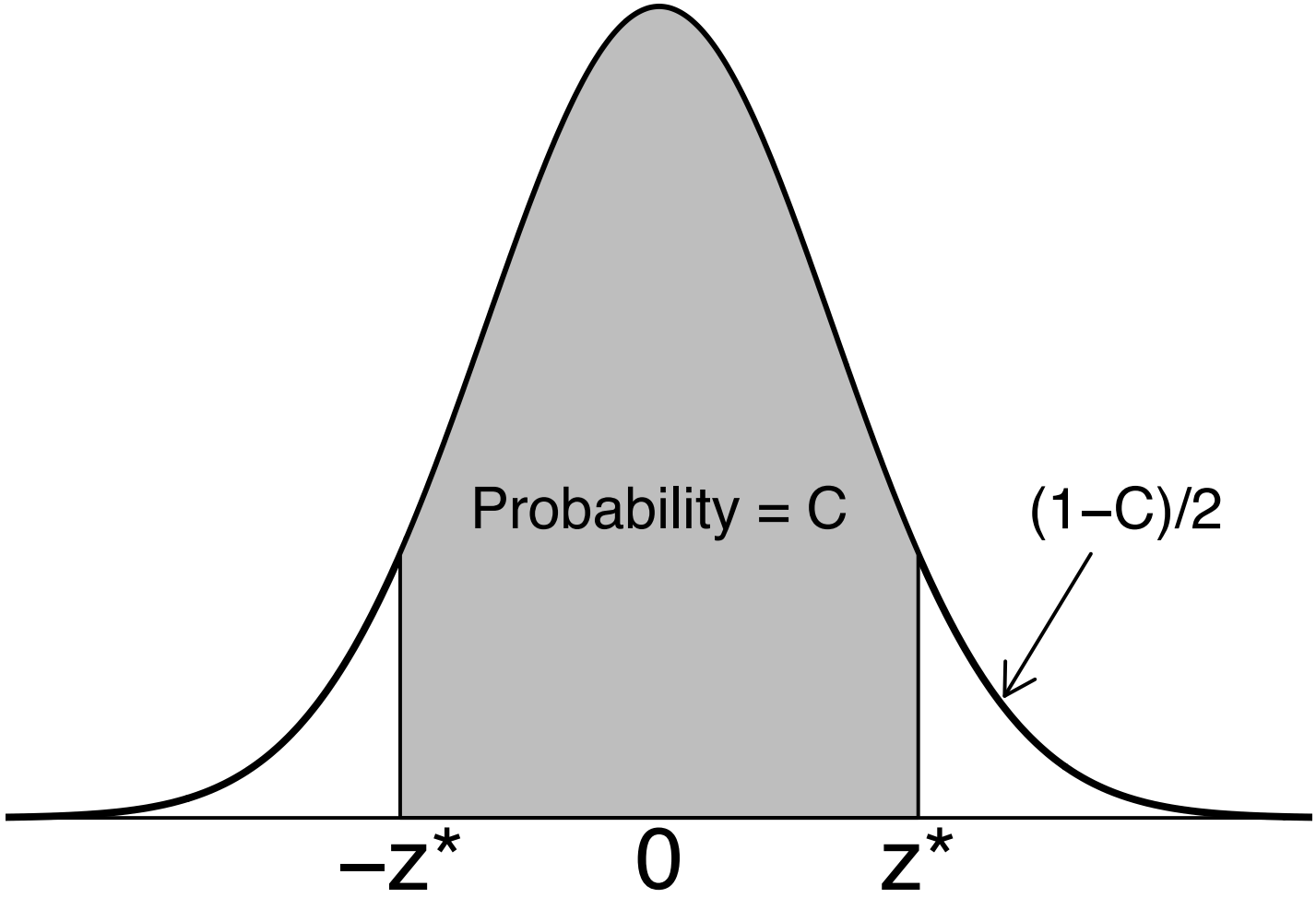

0
10
20
30
40
50
−2
0
2
4

# Confidence Interval for the mean of a Normal population

Draw an SRS (Simple Random Sample) of size $n$ form a Normal population having unknown mean $\mu$ and **known** standard deviation $\sigma$. A level $C$ confidence interval for $\mu$ is

$$\bar{x} \pm z_* \frac{\sigma}{\sqrt{n}}$$

The critical value $z_*$ is illustrated in a Figure below and depends on $C$.


Probability = C
(1–C)/2
–z*
0
z*

# Large Sample CI for $\mu$ (Normal data)

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Valid if

- n large
- random sample from a Normal distribution
- independent observations

Some definitions

- $1 - \alpha$ is the confidence coefficient
- $100(1 - \alpha)\%$ is the confidence level

# One Sample CI on the population mean $\mu$

In questions, this info is based on wording

## 1/ when pop. st. dev $\sigma$ is KNOWN (unrealistic but most approachable to start)

from standard normal

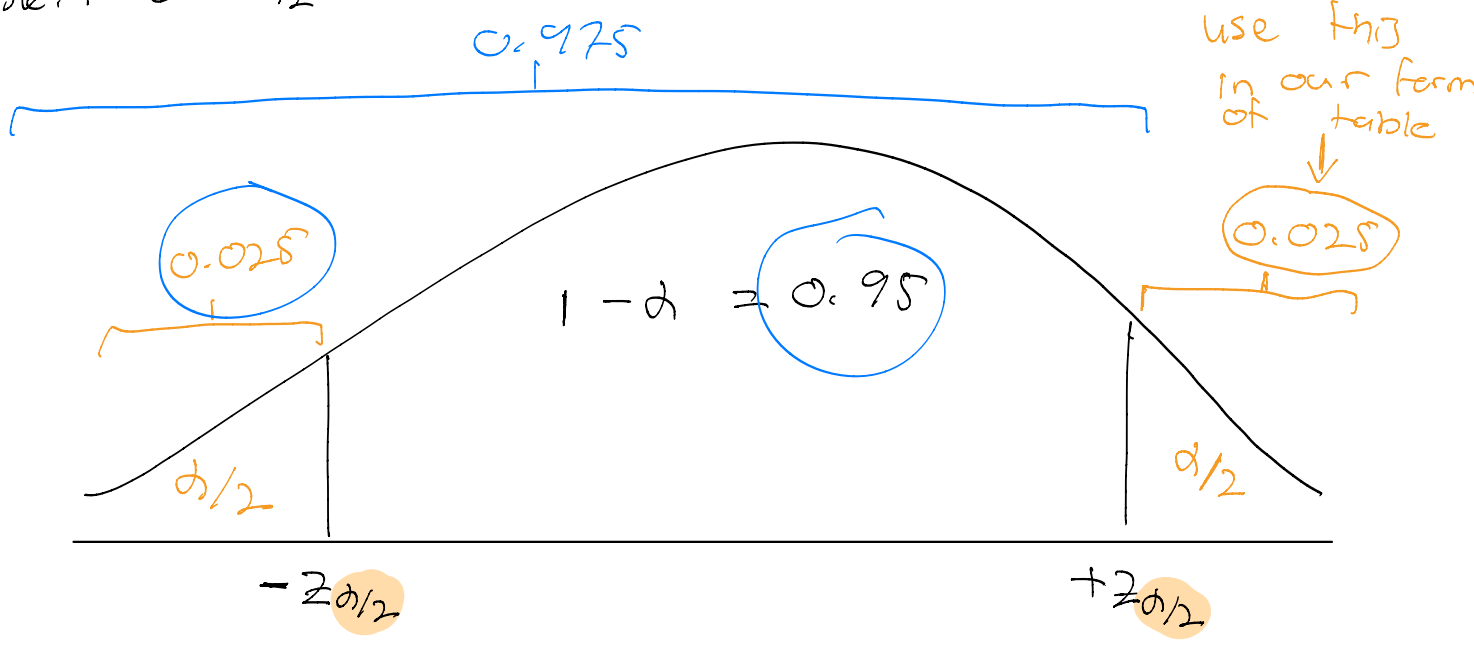
standard error

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

margin of error

How to find $z_{\alpha/2}$?

Example: Find $z_{\alpha/2}$ for a 95% CI on $\mu$




$-1.96$

$+1.96$

$1-\alpha = 0.95$

$\alpha = 0.05$

$\alpha/2 = 0.025$

table

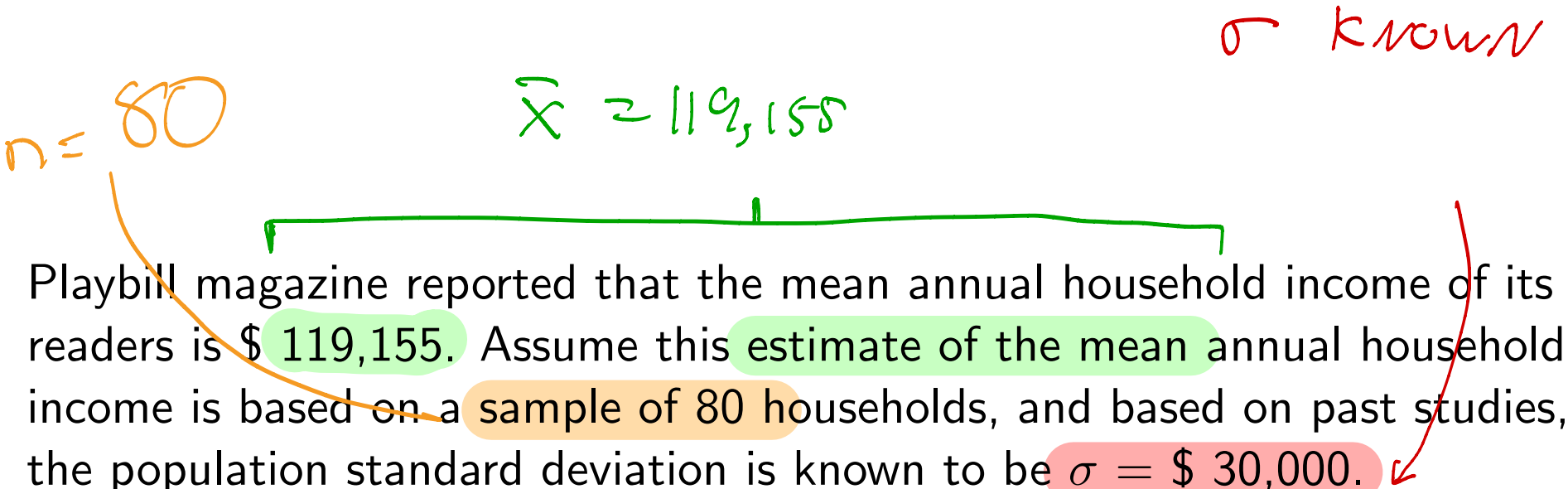
R

```
> qnorm(1-α/2)
> qnorm(1-0.025)
> qnorm(0.975)
```

| Confidence coefficient | Confidence level | $z$ |
|---|---|---|
| 0.90 | 90% | 1.645 |
| 0.95 | 95% | 1.96 |
| 0.99 | 99% | 2.576 |

Exercise: use table to obtain $z_{\alpha/2}$ for 90%, 99%

# Example




Playbill magazine reported that the mean annual household income of its readers is \$ 119,155. Assume this estimate of the mean annual household income is based on a sample of 80 households, and based on past studies, the population standard deviation is known to be $\sigma = \$\ 30{,}000$.

a. Develop a 90 % confidence interval estimate of the population mean.
b. Develop a 95 % confidence interval estimate of the population mean.
c. Develop a 99 % confidence interval estimate of the population mean.

Example (Slide 16)

$\bar{X} = 119155$, $n = 60$, $\sigma = 30000$ ($\sigma$ known)

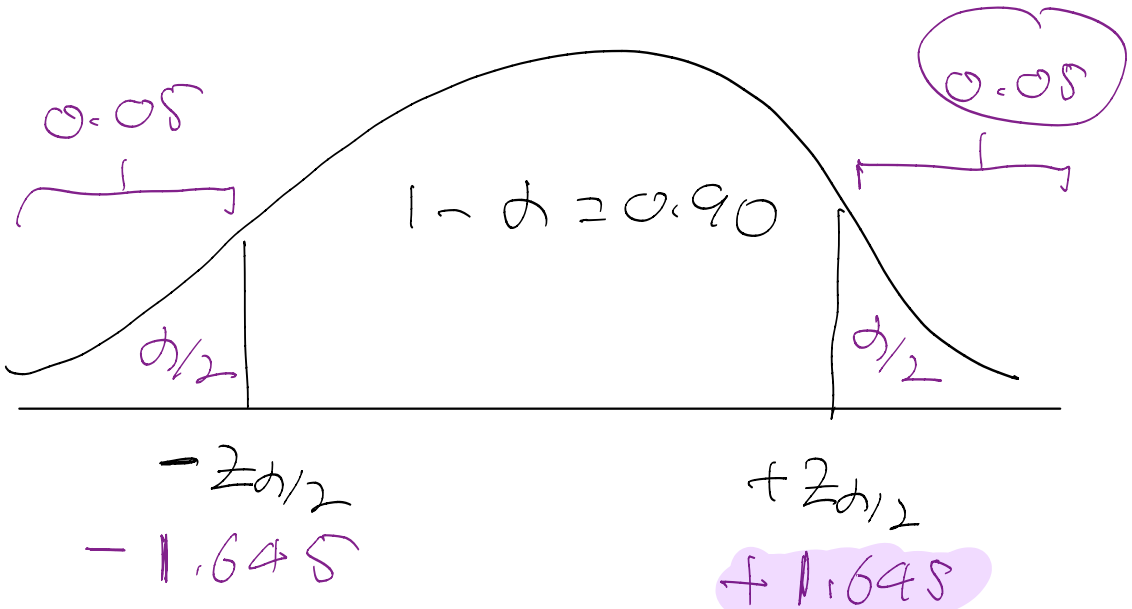
90% CI for $\mu$. Since $\sigma$ known, CI is

$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$z_{\alpha/2}$ for 90% CI

our form.

of table




$1-\alpha = 0.90$

$\alpha = 0.10$

$\alpha/2 = 0.05$

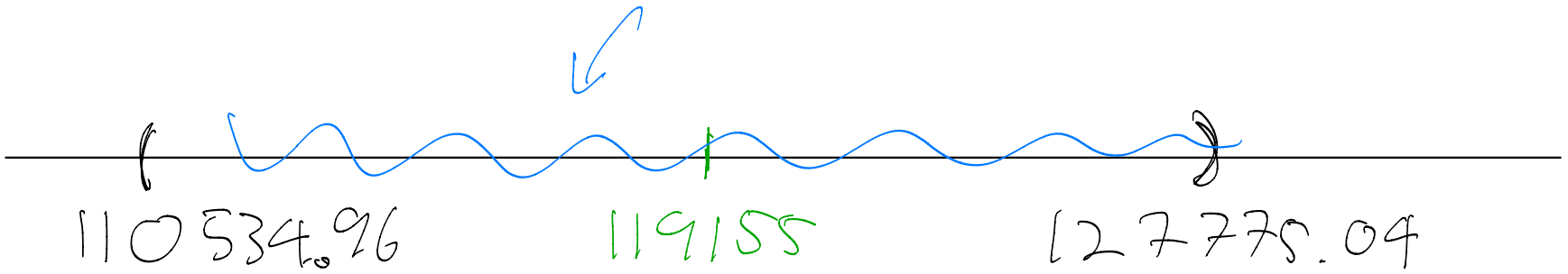
$$\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 119155 \pm 1.645 \left( \frac{30000}{\sqrt{80}} \right)$$

$$= 119155 \pm 8620.04$$

$$= (119155 - 8620.04,\ 119155 + 8620.04)$$

$$= (110534.96,\ 127775.04)$$

90% confident $\mu$ inside



Interpretation

We are 90% confident the mean household income of magazine readers is between $110,534.96 and $127,775.04.

# Solution

In this case, $\mu =$ mean annual household income of **all** its readers.
$\sigma = 30,000$, $n = 80$, $1 - \alpha = 0.90$, $\alpha = 0.10$.
a) margin of error $= Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (1.64)\left(\frac{30000}{\sqrt{80}}\right) = 5500.727 \approx 5500.73$
Confidence Interval is given by: estimate $\pm$ margin of error. That is:
$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
$119,155 \pm 5500.73$

$(113,654.27,\ 124,655.73)$

# Solution

In this case, $\mu =$ mean annual household income of **all** its readers.
$\sigma = 30,000$, $n = 80$, $1 - \alpha = 0.95$, $\alpha = 0.05$.
b) margin of error $= Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (1.96)\left(\frac{30000}{\sqrt{80}}\right) = 6574.04$
Confidence Interval is given by: estimate $\pm$ margin of error. That is:
$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
$119,155 \pm 6574.04$

$(112,580.96,\ 125,729.04)$

# Solution

In this case, $\mu =$ mean annual household income of **all** its readers.
$\sigma = 30,000$, $n = 80$, $1 - \alpha = 0.99$, $\alpha = 0.01$.
c) margin of error $= Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (2.57)\left(\frac{30000}{\sqrt{80}}\right) = 8620.042 \approx 8620.04$
Confidence Interval is given by: estimate $\pm$ margin of error. That is:
$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
$119,155 \pm 8620.04$

$(110,534.96,\ 127,775.04)$

# Example

Normal

$\sigma = 15$

$n = 15$ (small)

The number of cars sold annually by used car salespeople is Normally distributed with a standard deviation of 15. A random sample of 15 salespeople was taken, and the number of cars each sold is listed here. Find the 95% confidence interval estimate of the population mean. Interpret the interval estimate.

$\bar{x} = ?$

# Example

Raw data

| | | | | |
|---|---|---|---|---|
| 79 | 43 | 58 | 66 | 101 |
| 63 | 79 | 33 | 58 | 71 |
| 60 | 101 | 74 | 55 | 88 |

$$\bar{X} = \frac{79 + 43 + \cdots + 55 + 88}{15} = 68.6$$

Exercise: Complete

## R function

```
# R Code;

simple.z.test = function(x,sigma,conf.level=0.95) {
n = length(x);
xbar=mean(x);
alpha = 1 - conf.level;
zstar = qnorm(1-alpha/2);
SE = sigma/sqrt(n);
xbar + c(-zstar*SE,zstar*SE);
}
```

```r
# Step 1. Entering data;

cars=c(79, 43, 58, 66, 101, 63, 79,
33, 58, 71, 60, 101, 74, 55, 88);

# Step 2. Finding CI;

simple.z.test(cars,15);

## [1] 61.00909 76.19091
```

Ans

# Interpretation

We estimate that the mean number of cars sold annually by all used car salespersons lies between 61 and 76, approximately. This type of estimate is correct 95% of the time.

# Example

Suppose a student measuring the boiling temperature of a certain liquid observes the readings (in degrees Celsius) 102.5, 101.7, 103.1, 100.9, 100.5, and 102.2 on 6 different samples of the liquid. He calculates the sample mean to be 101.82. If he knows that the distribution of boiling points is Normal, with standard deviation 1.2 degrees, what is the confidence interval for the population mean at a 95% confidence level?

A **confidence interval** uses sample data to estimate an unknown population parameter with an indication of how accurate the estimate is and of how confident we are that the result is correct.

The **interval** often has the form
estimate $\pm$ margin of error

The **confidence level** is the success rate of the method that produces the interval.

A level C **confidence interval for the mean** $\mu$ of a Normal population with **known** standard deviation $\sigma$, based on an SRS of size $n$, is given by

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

The **critical value** $z^*$ is chosen so that the standard Normal curve has area $C$ between $-z^*$ and $z^*$.

Other things being equal, the **margin of error** of a confidence interval gets smaller as

- the confidence level $C$ decreases,
- the population standard deviation $\sigma$ decreases, and
- the sample size $n$ increases.

# CASES WHERE VALID

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Valid for

1. large samples where population is normal
2. " " " " " not normal
   (By CLT)
3. small " " " " normal

large : $n \geq 30$

# APPENDIX

# Confidence Intervals

Interval estimators are commonly called **confidence intervals**. The upper and lower endpoints of a confidence interval are called the **upper** and **lower confidence limits**, respectively. The probability that a (random) confidence interval will enclose $\theta$ (a fixed quantity) is called the **confidence coefficient**.

# Confidence Intervals (cont.)

Suppose that $\hat{\theta}_L$ and $\hat{\theta}_U$ are the (random) lower and upper confidence limits, respectively, for a parameter $\theta$. Then, if

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha,$$

the probability $(1 - \alpha)$ is the **confidence coefficient**.

# Pivotal quantities

One very useful method for finding confidence intervals is called the **pivotal method**. This method depends on finding a pivotal quantity that possesses two characteristics:

- It is a function of the sample measurements and the unknown parameter $\theta$, where $\theta$ is the **only** unknown quantity.
- Its probability distribution does not depend on the parameter $\theta$.