# Lesson 3. Confidence Intervals – Part 1

## 1 Populations vs. samples

> **Example 1.** Suppose researchers are interested in estimating the average BMI (body mass index) of American men over the age of 18. The researchers obtain a random sample of 422 men and calculate a mean BMI of 28.0 kg/m$^2$. Further suppose the standard deviation of all American men BMIs is known to be 5.4 kg/m$^2$.

| | **Definition** | **For Example 1** |
|---|---|---|
| **population** | All individuals in the group of interest. | |
| **parameter** | Numerical characteristic of the population distribution. Fixed value. Does not depend on data. | |
| **sample** | A collection of independent, identically distributed (i.i.d) r.v.s. | |
| **data** | An observed sample; actual numbers. | |
| **statistic (estimator)** | A function of the random variables in the sample. Used to estimate a parameter. | |
| **observed statistic (estimate)** | The numerical value of a statistic, calculated with data values. | |

- A parameter describes a population, while a statistic describes a sample

- Examples of statistics:

| | statistic (estimator) | observed statistic (estimate) |
|---|---|---|
| **sample mean** | | |
| **sample standard deviation** | | |

- A **simple random sample (SRS)** is a subset of the population, chosen randomly such that
  - each individual has the same chance of being chosen, and
  - all subsets of the same size have the same chance of being chosen

## 2 Confidence intervals

- We can use the sample mean to get a **point estimate** of the population mean, but...
- Usually, we would also like to get an **interval estimate**, a range of plausible values that includes a margin of error
- The most general form of an interval estimate is

| |
|---|
| |

- The margin of error is composed of the **critical value** times a **standard error (SE)** term:

| |
|---|
| |

### 2.1 CI for population mean when population variance ($\sigma^2$) is known

- Suppose $x_1, \ldots, x_n$ is data from a simple random sample from a population with unknown mean $\mu$ and known variance $\sigma^2$
- The $(1-\alpha)100\%$ **CI for the population mean** is

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

where

$\alpha$ = **significance level** $\qquad$ $1-\alpha$ = **confidence level**

$\bar{x}$ = sample mean (estimate) $\qquad$ $\sigma$ = population standard deviation

$z_{\alpha/2}$ = $(1-\alpha/2)$-quantile of $N(0,1)$ $\qquad$ $n$ = sample size

- Interpretation – if the $(1-\alpha)100\%$ CI for the population mean is $(\ell, u)$:

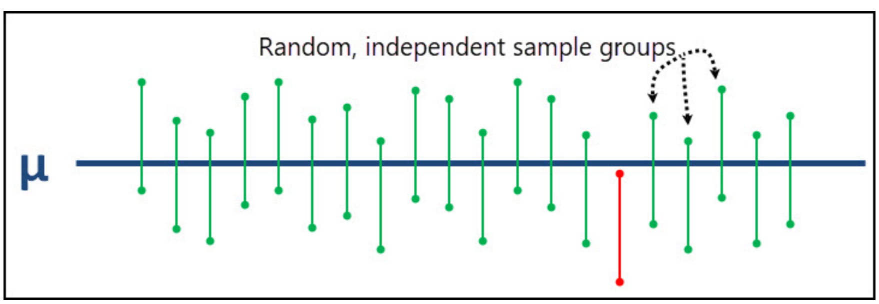
  We are $\underline{100(1-\alpha)}\%$ confident that the <u>population mean</u> is between $\underline{\ell}$ and $\underline{u}$.

  The underlined parts should be rephrased to correspond to the specific CI and data

**Example 2.**

a. Based on Example 1, calculate a 95% confidence interval for the mean BMI of all American men. Note that $z_{0.025} = 1.96$.
b. Interpret your interval from part a.

## 2.2 What does it mean to be "95% confident"?




- "95% confidence" means that if we were to <u>repeatedly take samples of size $n$</u> and construct the corresponding confidence intervals, 95% of the intervals would contain the true population mean $\mu$
- The probability that the <u>process</u> of forming a CI <u>will capture</u> $\mu$ is 0.95
- It is NOT correct to say that the probability we captured $\mu$ with our particular CI estimate is 0.95

### 2.3 CI for population mean when population variance ($\sigma^2$) is unknown

- Now suppose $x_1, \ldots, x_n$ is data from a simple random sample from a population with unknown mean $\mu$ and unknown variance $\sigma^2$
- The $(1-\alpha)100$% **CI for the population mean** is

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

where

| | |
|---|---|
| $\alpha$ = significance level | $1-\alpha$ = confidence level |
| $\bar{x}$ = sample mean (estimate) | $s$ = sample standard deviation (estimate) |
| $t_{\alpha/2,n-1} = (1-\alpha/2)$-quantile of $t(n-1)$ | $n$ = sample size |

- ○ Interpretation – if the $(1-\alpha)100$% CI for the population mean is $(\ell, u)$:

  We are $100(1-\alpha)$% confident that the population mean is between $\ell$ and $u$.

  The underlined parts should be rephrased to correspond to the specific CI and data

### 2.4 Why is the $t$-distribution used when the population variance is unknown?

When $\sigma^2$ is known:

- Central Limit Theorem says $\bar{X} \sim N(\mu, \sigma^2)$

$$\Rightarrow \quad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$\Rightarrow$ Critical value from $N(0,1)$

When $\sigma^2$ is unknown:

- It can be shown that

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)$$

$\Rightarrow$ Critical value from $t(n-1)$

### 2.5 Technical conditions to check

- Two things must be met for the above CI formulas to be appropriate:

1. Data must be from a ________
2. Either ________ or ________