Statistics Bootcamp using R

DAY 2 DATA VISUALIZATION & UNDERSTANDING PATTERN

2.2 DESCRIPTIVE STATISTICS & SAMPLING

GU Zhan (Sam)
Institute of Systems Science
National University of Singapore

issgz@nus.edu.sg

Agenda

Day 2 : Data Visualization & Understanding Pattern

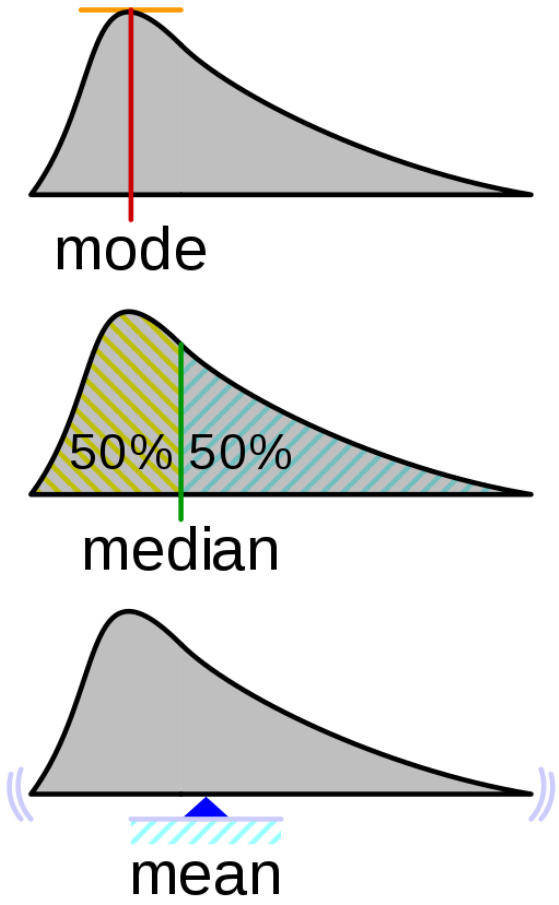
- Data Visualization
- **Descriptive Statistics & Sampling**
- Introduction to Normal Distribution

Learning objectives

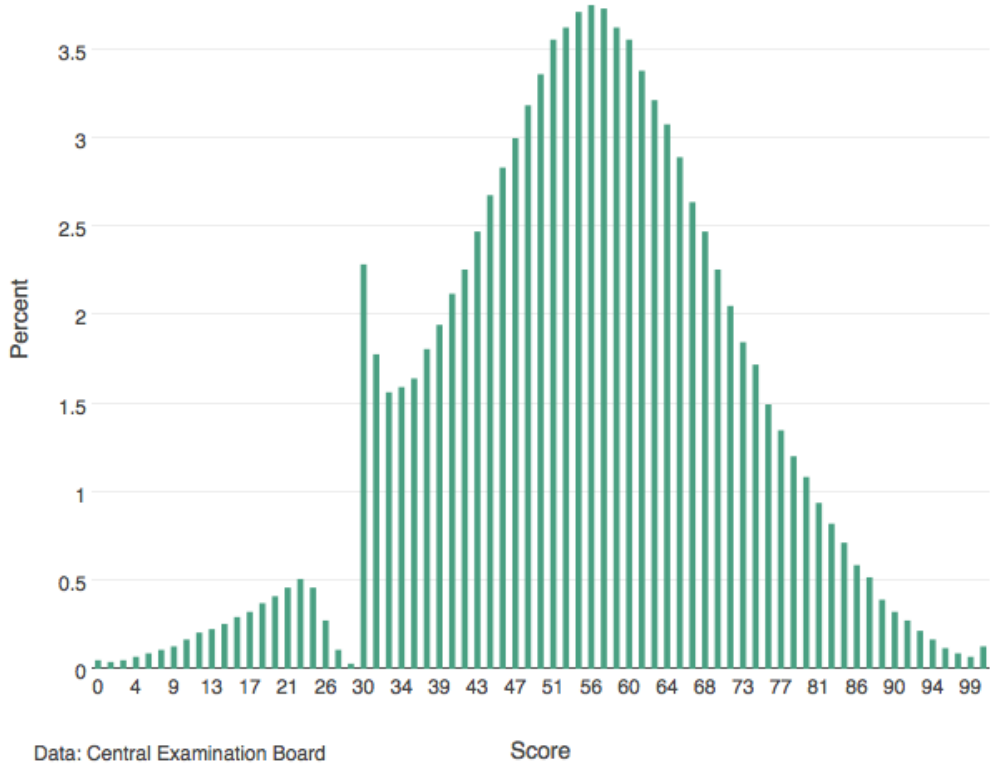
- Understand the measures of central tendency & measures of variability for statistical inference
- Understand sampling techniques & sampling distribution
- Understand concepts of estimation

Measures of central tendency (revisit)

- Three common measures:
  - Mode: the most commonly occurring number
  - Median: the number that sits in the middle of an ordered group of numbers
  - Mean: the arithmetic average



Distribution of the results of the Poland's High School Exit Exam

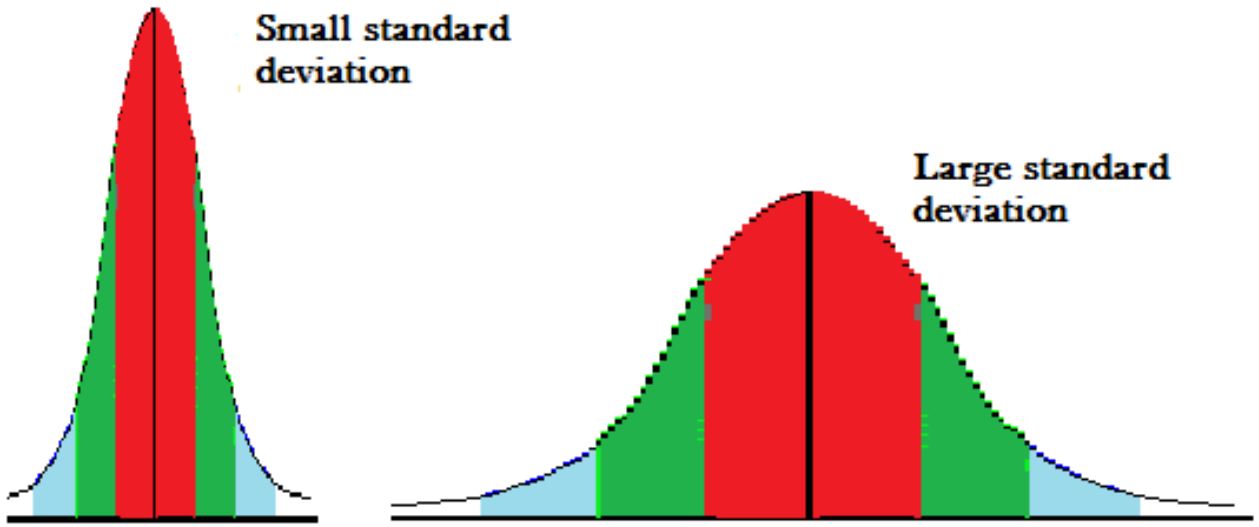


Source: https://imgur.com/gallery/4nnbyru

Summary: measuring central tendency

Average	How to calculate	When to use it
Mean	Add all the numbers in a data set together, and then divide by how many there are.	The data is fairly **symmetric** and shows just the one trend.
Median	Line up all the values in ascending order. If there are an odd number of values, the median is the one in the middle. If there are an even number of values, add the two middle ones together, and divide by two.	When the data is **skewed** because of outliers.
Mode	Choose the value(s) with the highest frequency. If the data is showing two clusters of data, report a mode for each group.	When you're working with **categorical** data. When the data shows two or more clusters.

Measures of variability



Source: https://www.statisticshowto.datasciencecentral.com/average-deviation/

- Variability (dispersion): the degree to which individual data points are distributed around the mean

- Some common measures:
  - Range: the distance from the lowest to the highest value
  - Variance: the average of the squared deviations from the arithmetic mean for a set of numbers.
  - Standard deviation: the square root of the variance

Revisit Data sources (Raw material for calculating Measures of Central Tendency, Measures of Dispersion and all other analysis)

- Data sources are created in one of the four ways:
  1. Data distributed by an organization or individual
  2. Outcomes of a designed experiment
  3. Responses from a survey
  4. Result of an observational study

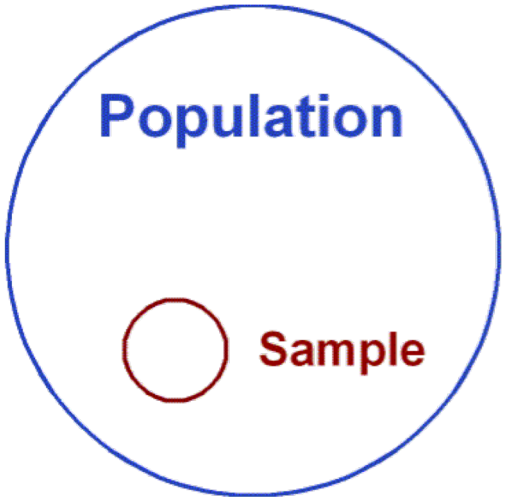
After collecting data we take representative samples to perform various calculations

But in general we are interested in the characteristics of a large population

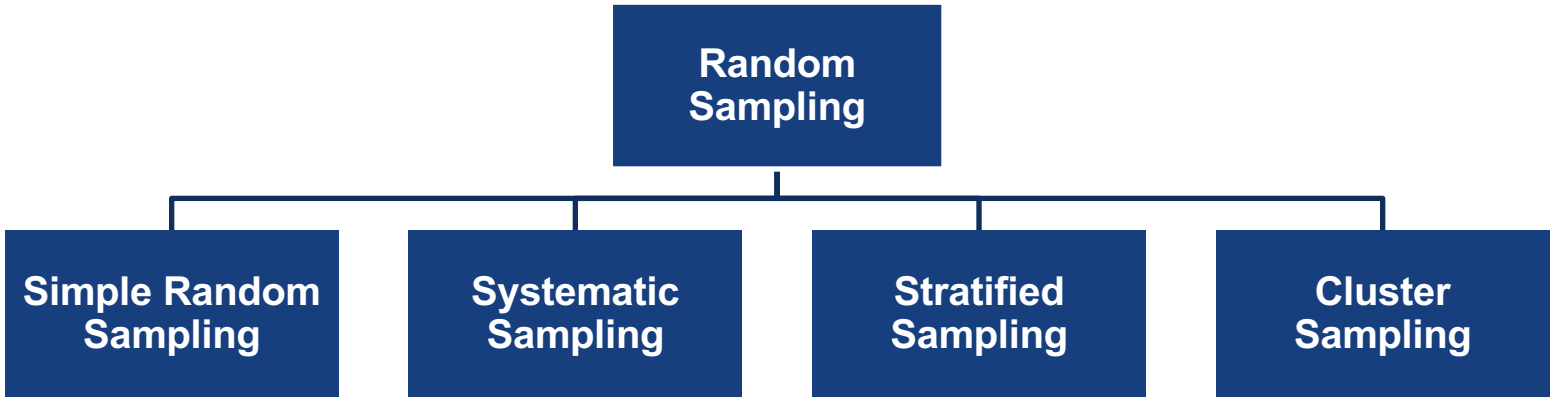
- Could be humans
- Could be animals
- Could be machines etc.

As calculation from the entire population could be expensive we wish to make statements about the population based on the sample we draw from the population

This is performed by drawing different **representative** samples from a population (generally random sample)

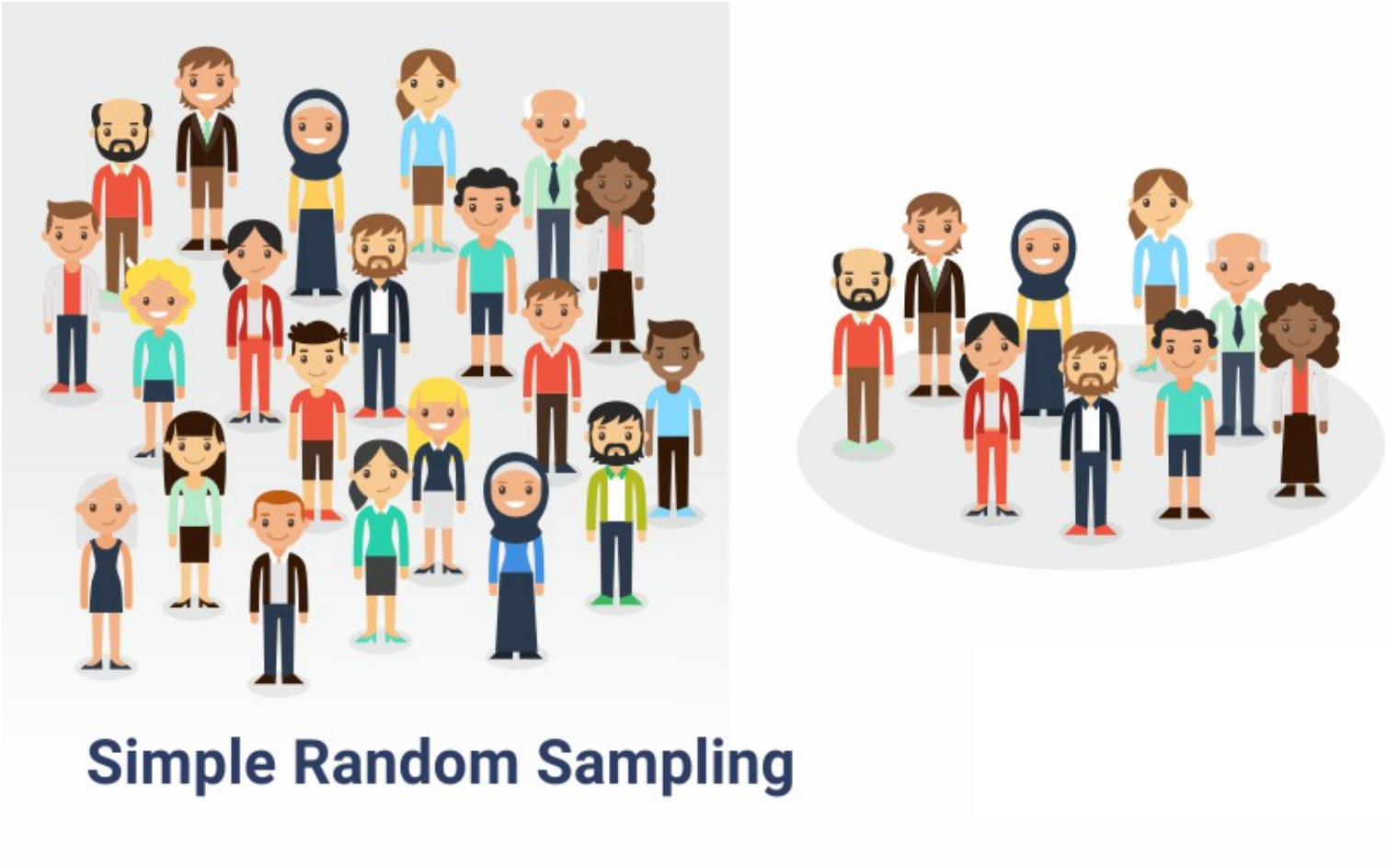


Random sampling techniques



Simple random sampling

Just random selection



Source: https://www.questionpro.com/blog/simple-random-sampling/

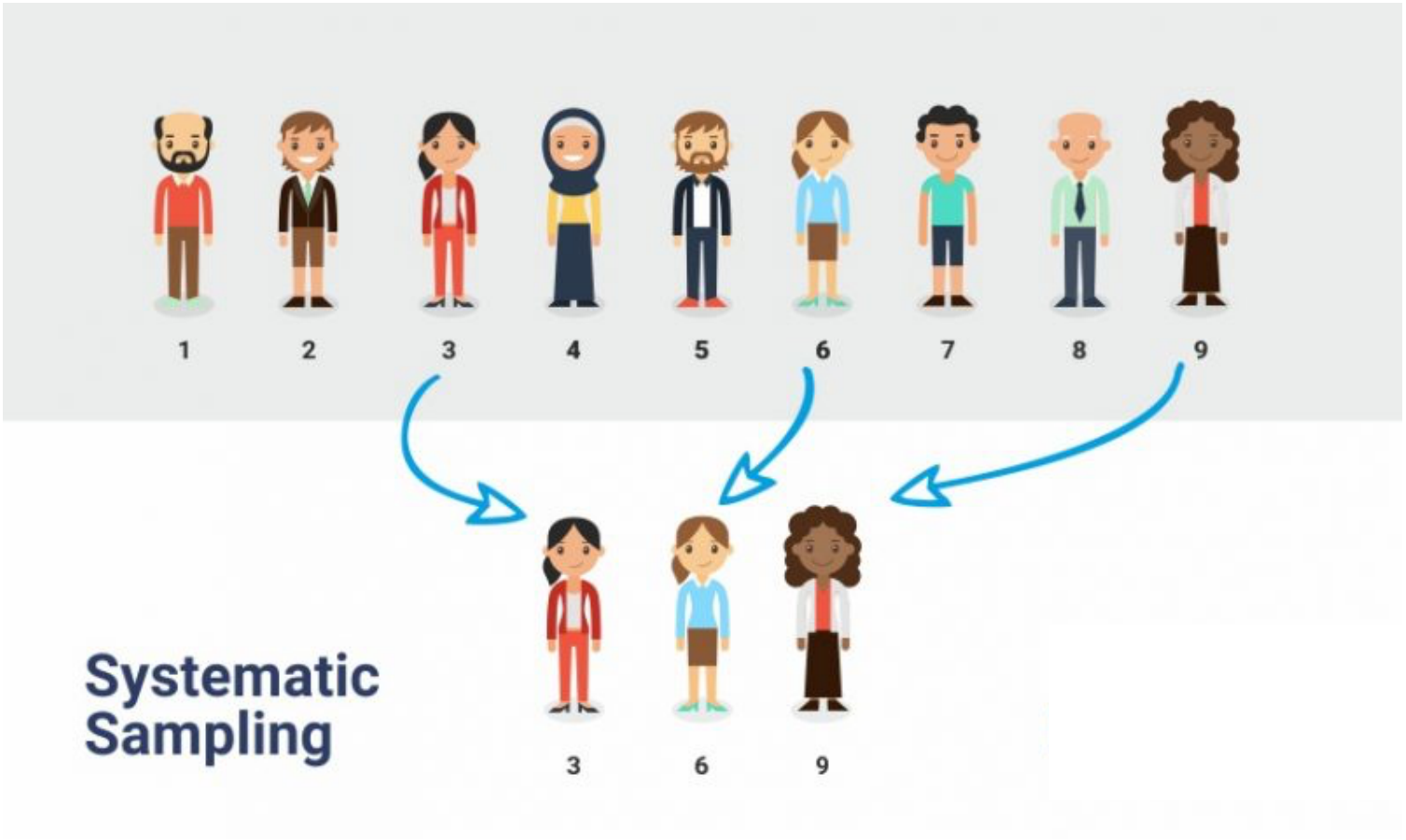
Cluster sampling

Used in situations like wars and natural calamities/disasters where cities form clusters

Source: https://www.questionpro.com/blog/cluster-sampling/

Systematic sampling

Simple yet systematic



Source: https://www.questionpro.com/blog/systematic-sampling/

Stratified random sampling

Higher accuracy than simple random sampling

Source: https://www.questionpro.com/blog/stratified-random-sampling/

Random sampling
in R

- Assign every item a number
- Use a table of random numbers (or a random generator) to select numbers
- The items assigned with the selected numbers can form a sample

```
# Randomly select 20 numbers from 1 to 50
> sample(1:50, 20)
 [1] 24 10 26 41 34  4 44 22 30 25 20 16 12  5 47  6 32 40 17

[20] 28
# Create a vector called "teacher"
> teacher = c("Sam", "Charles", "Rita", "Prakash", "Barry")


# Randomly select 4 individuals from the "teacher" vector.
> sample(teacher, 4, replace = TRUE)
[1] "Sam" "Rita" "Sam" "Prakash"      Why duplication appear?
```

The concept of estimation

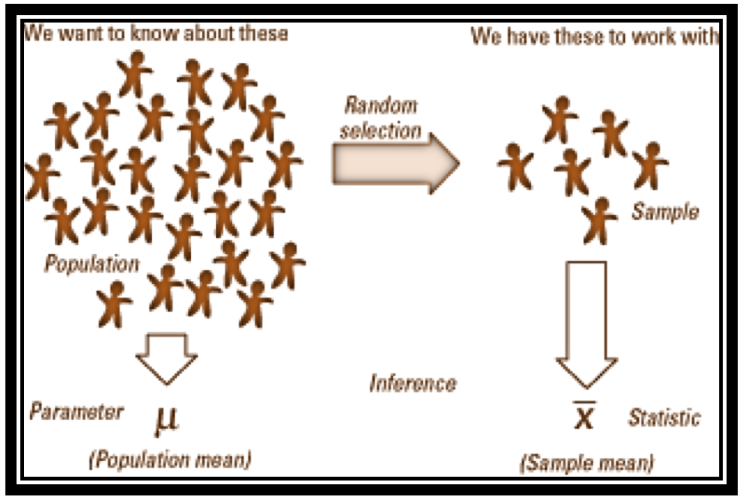
We are interested in the characteristics of a large population

These characteristics are usually described by some parameter , β say, that we do not know.

β could be

- The mean age of the population
- The proportion of the population who smoke
- The range of ages of the population
- The mean age of active FACEBOOK users

We wish to estimate this parameter based on a sample we draw from the population



What Estimators shall we create using sample values?

- We have a set of measurements (a sample), e.g. ages, x1, x2,….xn from a population.

Estimate the population mean & standard deviation, .e.g. average age

- We would like to estimate the population mean $\boldsymbol{\mu}$ by calculating the sample mean $\bar{x}$ where: $\hat{\mu} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

- Similarly we estimate the standard deviation of the population $\boldsymbol{\sigma}$ from the sample standard deviation s

$$\hat{\sigma} = s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$$

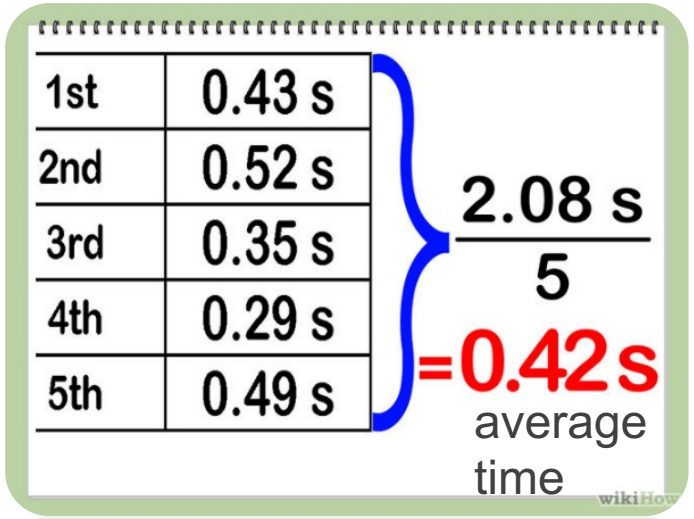
Estimate the population proportion (ratio), e.g. percentage of people elder than 50 years

- Suppose we know that a certain proportion of a population has a certain characteristic.
- We may want to estimate the proportion of the population $\boldsymbol{\pi}$ that has this characteristic based on a sample we have drawn from the population.
- The estimator we use is p

$$\hat{\pi} = p = \frac{\text{No. in sample with this characteristic}}{\text{n (total size of sample)}}$$

- Note : if we are estimating a certain parameter $\boldsymbol{\beta}$ (not through directly measurement), by some combination of x1, x2,….xn (independent variables or individuals from a sample) we can call it $\hat{\boldsymbol{\beta}}$

Uncertainty in estimates



REMEMBER

All of these statistics are based on a (usually) small sample of data

If we collected a bigger sample of data, then the value of these statistics may be different

So with each statistic we calculate there is some *uncertainty*

<u>But</u> we can also estimate these uncertainties

Point and Interval Estimates

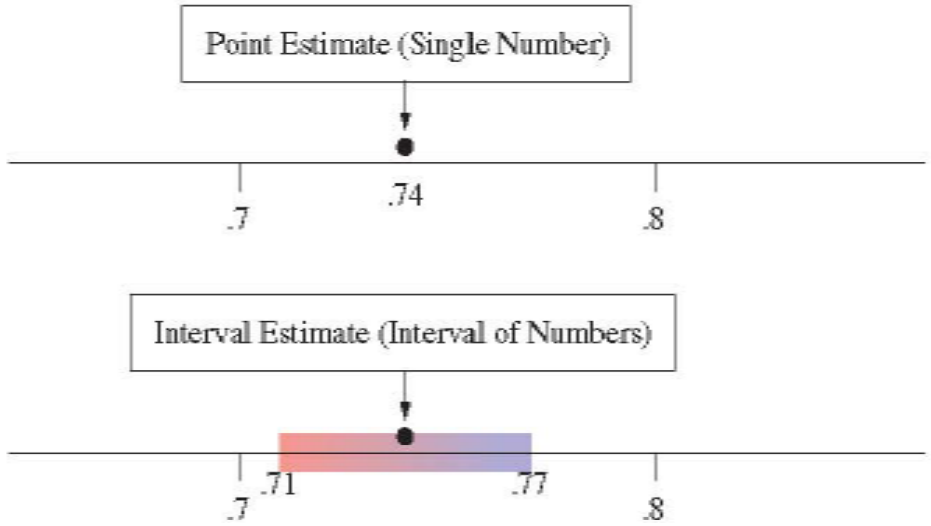
A **point estimate** is a **single number** (e.g., sample mean) that is our “best guess” for the population pparameter, e.g. average height of human beings

- The most common form of inference reported by the mass media

An **interval estimate (confidence interval)** is an **range of numbers** within which the population parameter value is believed to fall into (be within this range).

- It is useful to think of confidence intervals as a range of "plausible" values for the parameter

Question:
What is the fundamental reason for using interval estimate?



Calculating Errors in Estimates

We have collected a sample data $x_1, x_2, x_3, ...x_N$, from a larger population

We calculate the sample mean as $$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

We hope this is also a good estimate of the mean of the overall population

But it could be subject to error (as discussed previously)

Calculating Errors in Estimates- Example

For a class of 20 people their ages are: 21, 23, 24, 19, 25. 35, 29, 31, 34, 26, 23, 25, 29, 27, 28, 23, 33, 31, 21, 20

- The actual mean is 26.35

But if we took a sample of **four** numbers from the above group: 21, 26, 28, 20

- then the mean of these is 23.75

if we took a sample of **eight** numbers from the above group: 23, 25, 29, 26, 25, 28, 33, 20

- then the mean of these is 26.125

Intuitively, which mean would you **trust** more? 23.75 or 26.125?
So although we can make estimates from samples, they may not be completely accurate

- We must also make estimates of errors that accompany the estimates

Calculating Errors in Estimates (Continued)

- We calculate the sample mean of a population by taking a sample of n individuals

  and calculate the sample mean= $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

- We also calculated the population standard deviation as $\sigma = \sqrt{\frac{\sum_{k=1}^{n} (x_k - \mu)^2}{n}}$

  and sample standard deviation as $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

- We calculate the standard error of the mean = $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

- This gives some idea of how much variation there could be in the sample mean

Calculating Errors in Estimates (Continued)

In the earlier example we saw the following:

We took a sample of four numbers from the previous group: 21, 26, 28, 20

- The mean of these was 23.75 (years)
- The sample's standard error SE is 1.931 (years)

When we took a sample of eight numbers from the previous group: 23, 25, 29, 26, 25, 28, 33, 20

- The mean of these was 26.125 (years)
- The sample's standard error SE is 1.394 (years)

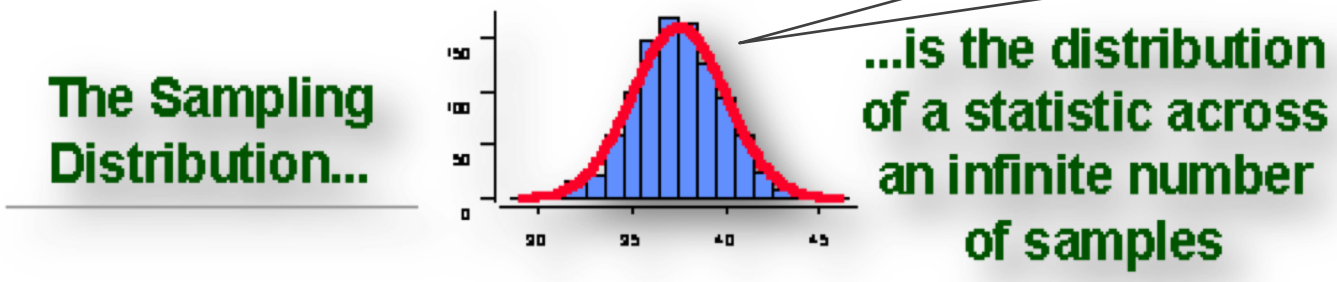
Intuitively, 8 samples are more trustworthy than 4 samples; Mathematically, SE 1.394 < 1.931.

Multiple Samples Creates A Sampling Distribution

Every sample has 10 people, sample size: n = 10

Here are three samples. Thus, the size of sample mean: k = 3

Question:
How many individual average values/heights are there, which forms the sample distribution/histogram?



Source https://slideplayer.com/slide/3299504/

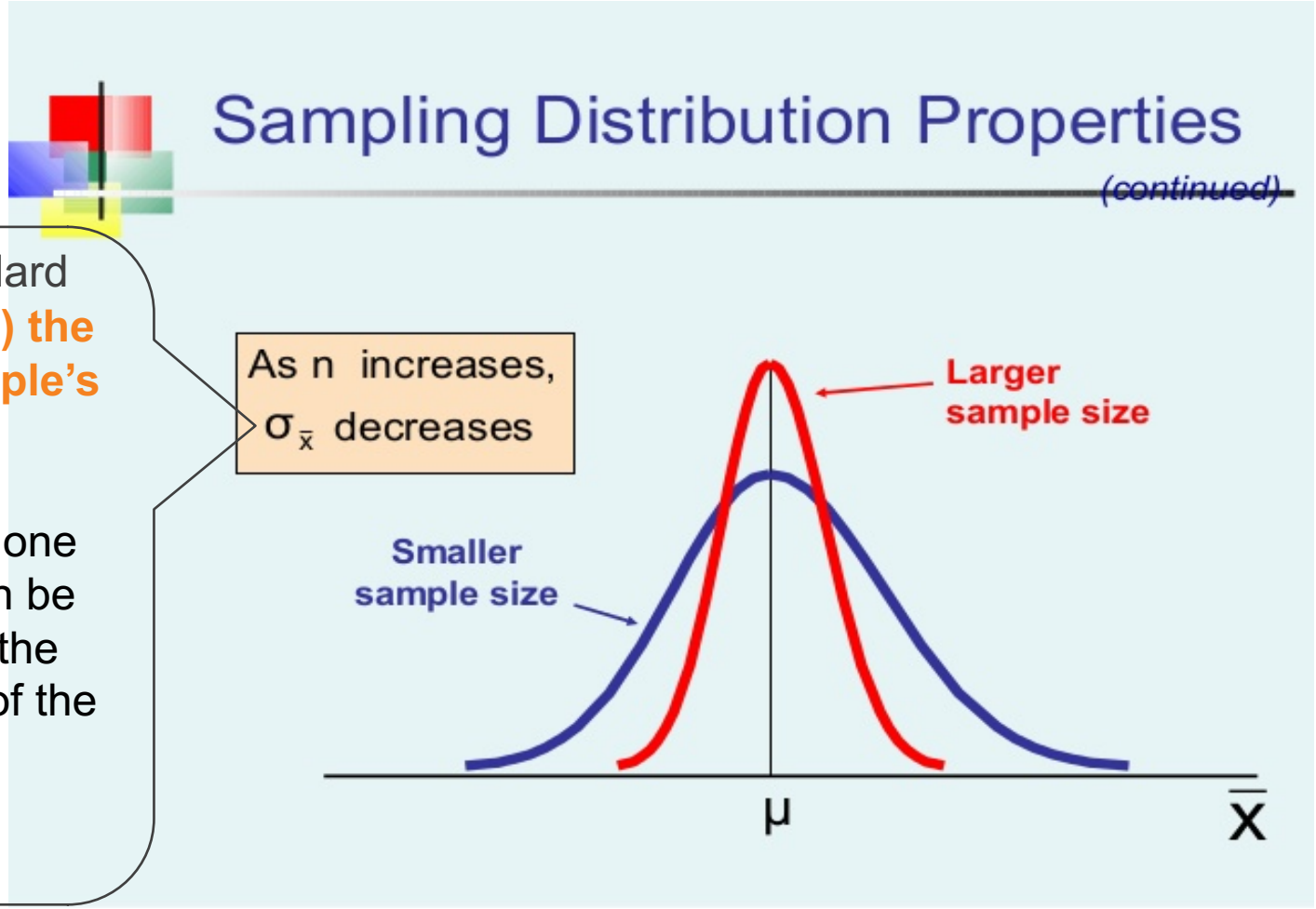
Errors Reduces As Sample Size Increases

The error of estimating the population mean μ using sample's means

σ_x is the standard deviation of **(all) the individual sample's means**

If there is only one sample, σ_x can be estimated as the standard error of the mean:

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$



Source https://slideplayer.com/slide/1396692/

End of Lecture Notes