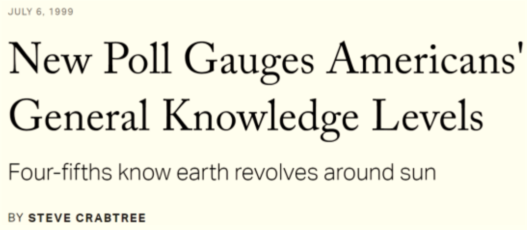
Statistical Inference

Foundations for Inference

Behnam Bahrak

Spring 2020

A Survey



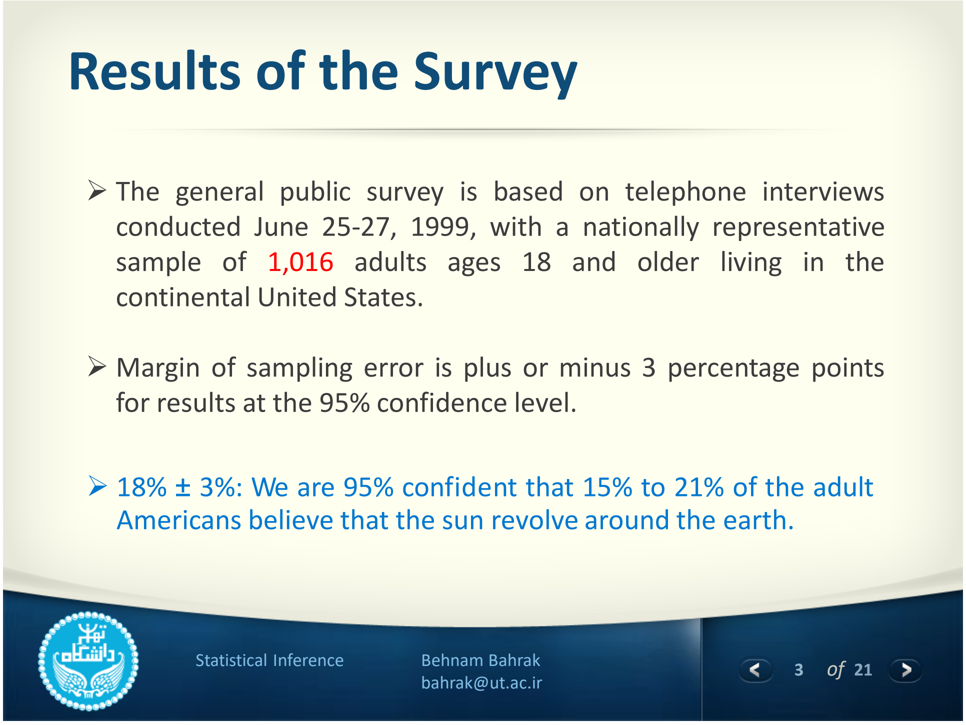
JULY 6, 1999

New Poll Gauges Americans' General Knowledge Levels

Four-fifths know earth revolves around sun

BY **STEVE CRABTREE**

- Probing a universal measure of knowledge, Gallup asked the basic science question: "As far as you know, does the earth revolve around the sun or does the sun revolve around the earth?"
  - 79% of Americans correctly respond that the earth revolves around the sun, while 18% say it is the other way around.



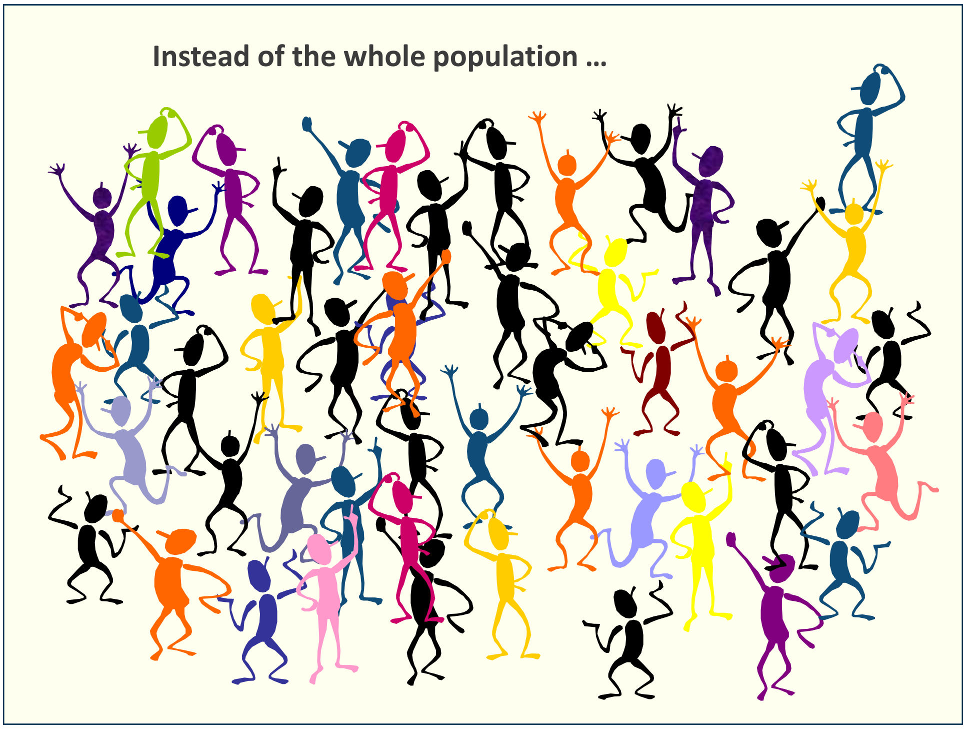
Results of the Survey

- The general public survey is based on telephone interviews conducted June 25-27, 1999, with a nationally representative sample of 1,016 adults ages 18 and older living in the continental United States.
- Margin of sampling error is plus or minus 3 percentage points for results at the 95% confidence level.
- 18% ± 3%: We are 95% confident that 15% to 21% of the adult Americans believe that the sun revolve around the earth.

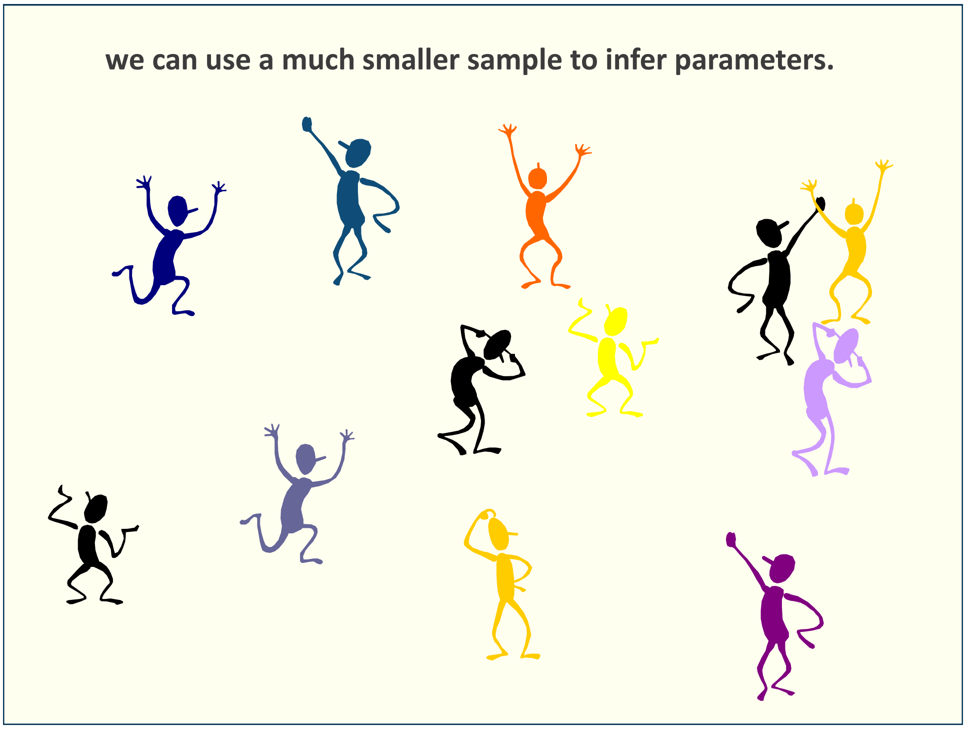
Statistical Inference

Behnam Bahrak
bahrak@ut.ac.ir

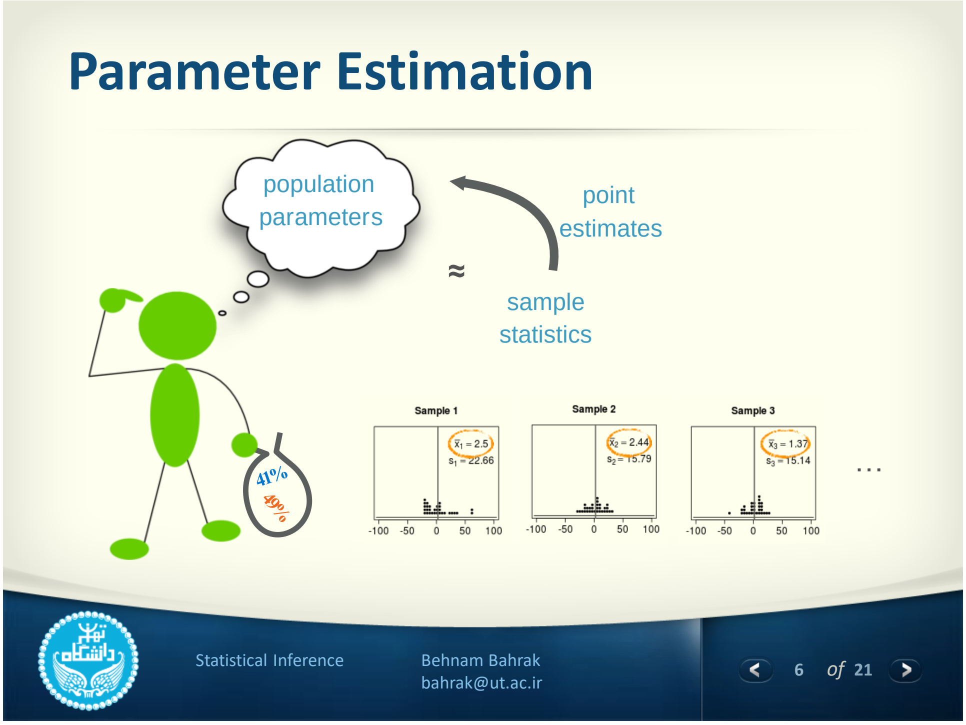
3 of 21



Instead of the whole population …



we can use a much smaller sample to infer parameters.



Parameter Estimation

population parameters

point estimates

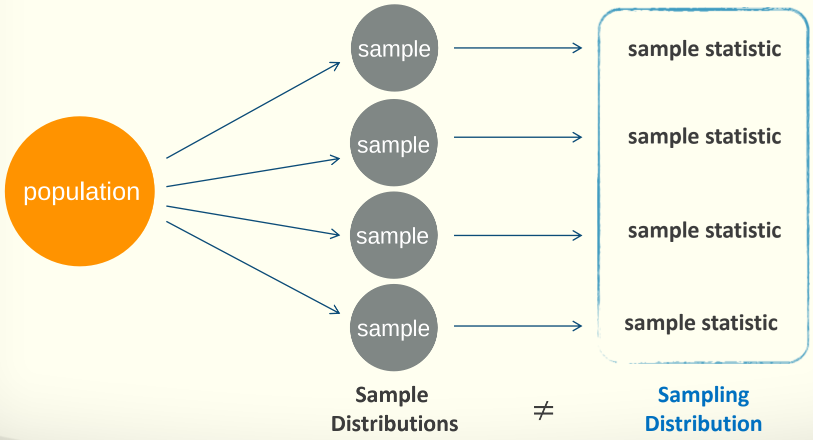
≈

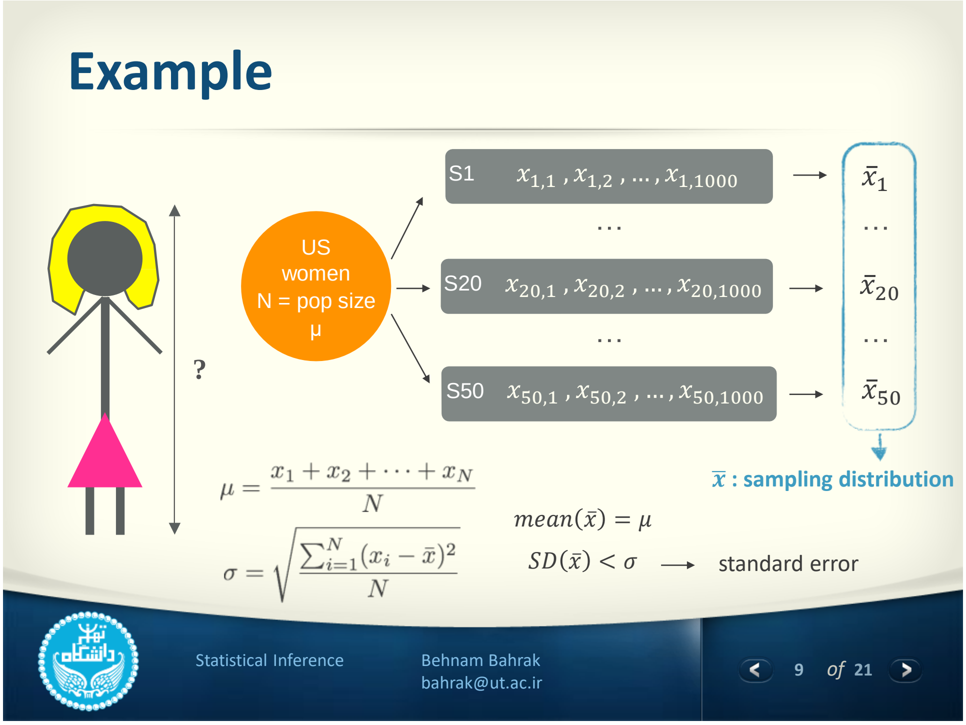
sample statistics

Parameter Estimation

- ➢ We are often interested in population parameters.
- ➢ Since complete populations are difficult (or impossible) to collect data on, we use sample statistics as point estimates for the unknown population parameters of interest.
- ➢ Sample statistics vary from sample to sample.
- ➢ Quantifying how sample statistics vary provides a way to estimate the margin of error associated with our point estimate.

Sample vs. Sampling Distribution





Example

US women
N = pop size
μ

S1 $x_{1,1}\,, x_{1,2}\,, \dots, x_{1,1000}$ ⟶ $\bar{x}_1$

…

S20 $x_{20,1}\,, x_{20,2}\,, \dots, x_{20,1000}$ ⟶ $\bar{x}_{20}$

…

S50 $x_{50,1}\,, x_{50,2}\,, \dots, x_{50,1000}$ ⟶ $\bar{x}_{50}$

$\bar{x}$: sampling distribution

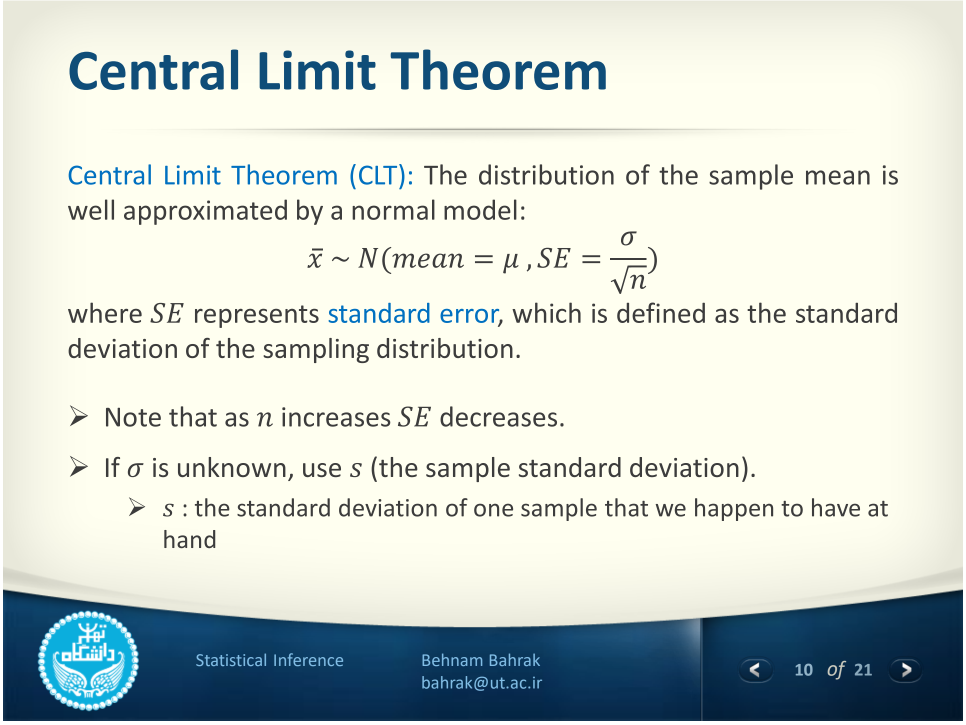
?

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

$$mean(\bar{x}) = \mu$$

$SD(\bar{x}) < \sigma$ ⟶ standard error



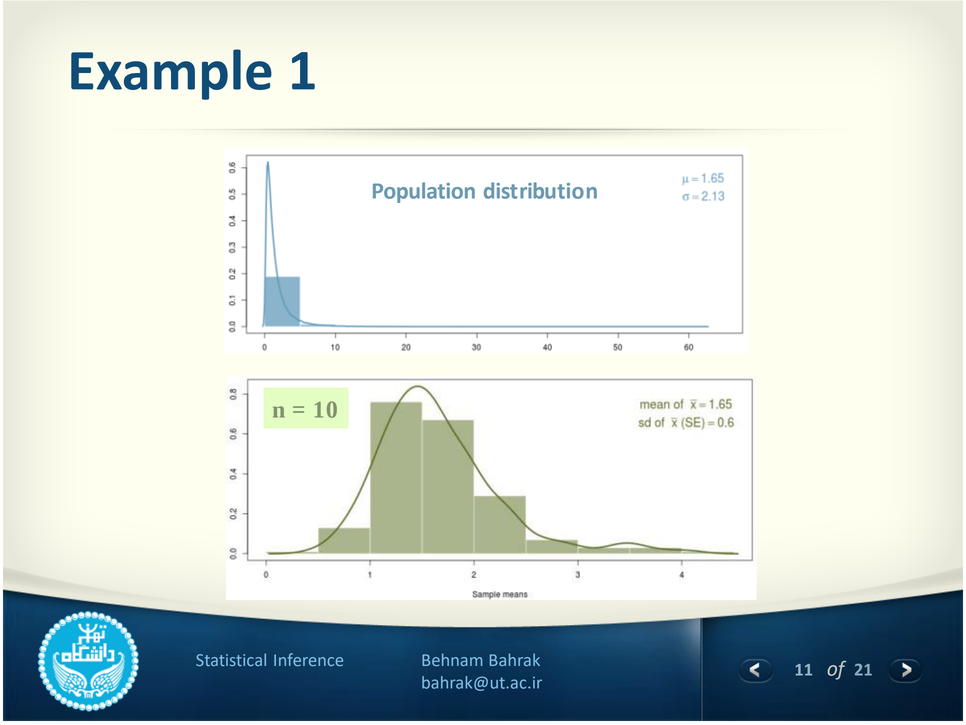
Central Limit Theorem

Central Limit Theorem (CLT): The distribution of the sample mean is well approximated by a normal model:

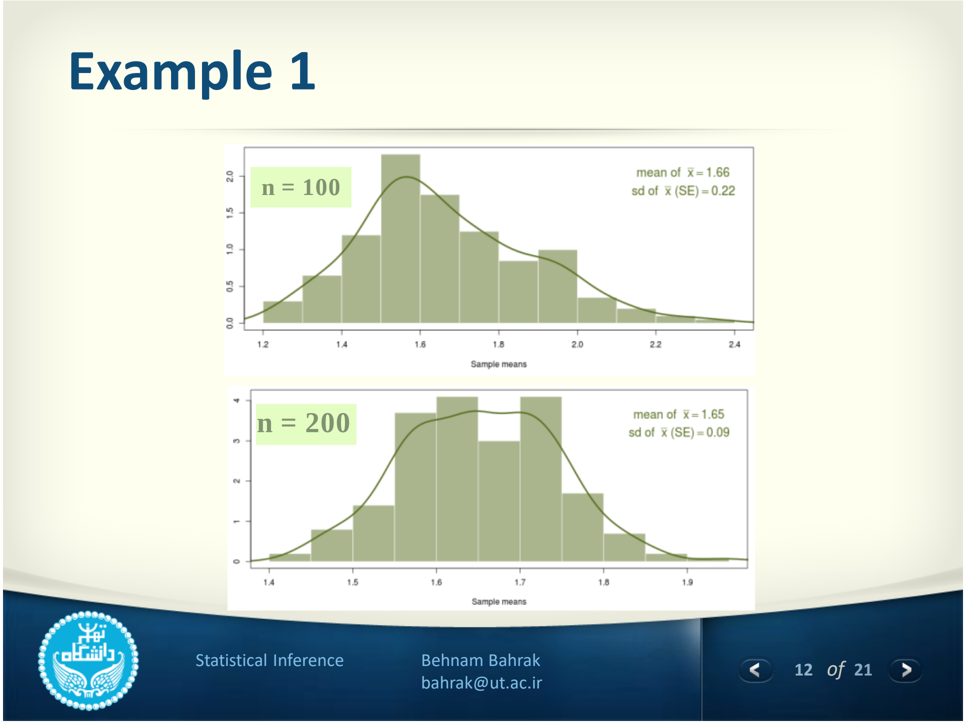
$$\bar{x} \sim N(mean = \mu\,, SE = \frac{\sigma}{\sqrt{n}})$$

where SE represents standard error, which is defined as the standard deviation of the sampling distribution.

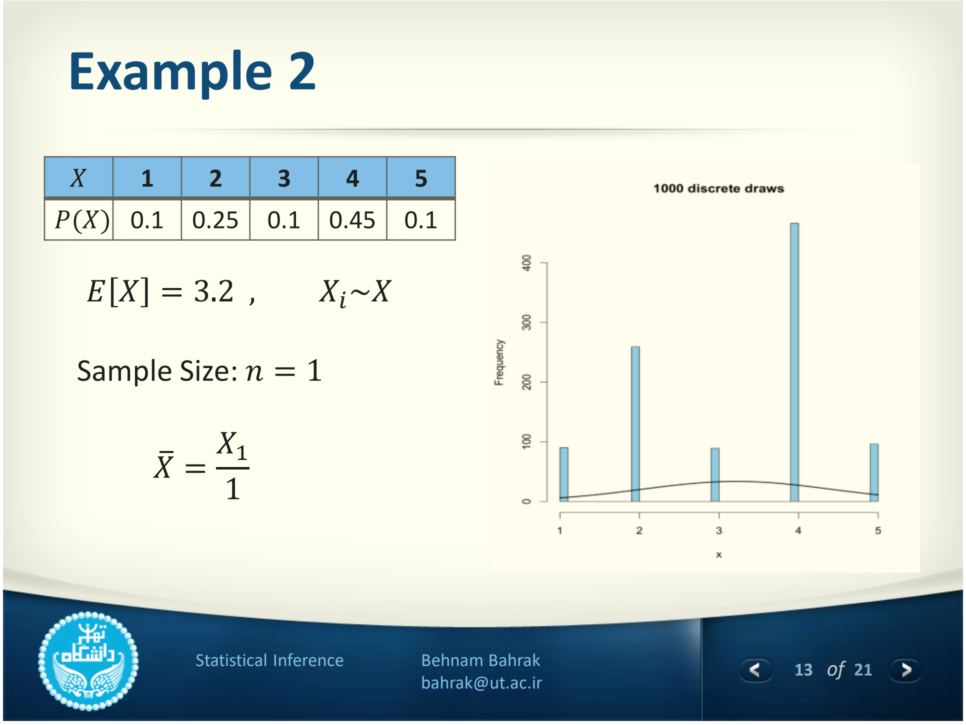
- Note that as n increases SE decreases.
- If σ is unknown, use s (the sample standard deviation).
  - s : the standard deviation of one sample that we happen to have at hand



Example 1



Example 1



Example 2

X	1	2	3	4	5
$P(X)$	0.1	0.25	0.1	0.45	0.1

$E[X] = 3.2$, $\quad X_i \sim X$

Sample Size: $n = 1$

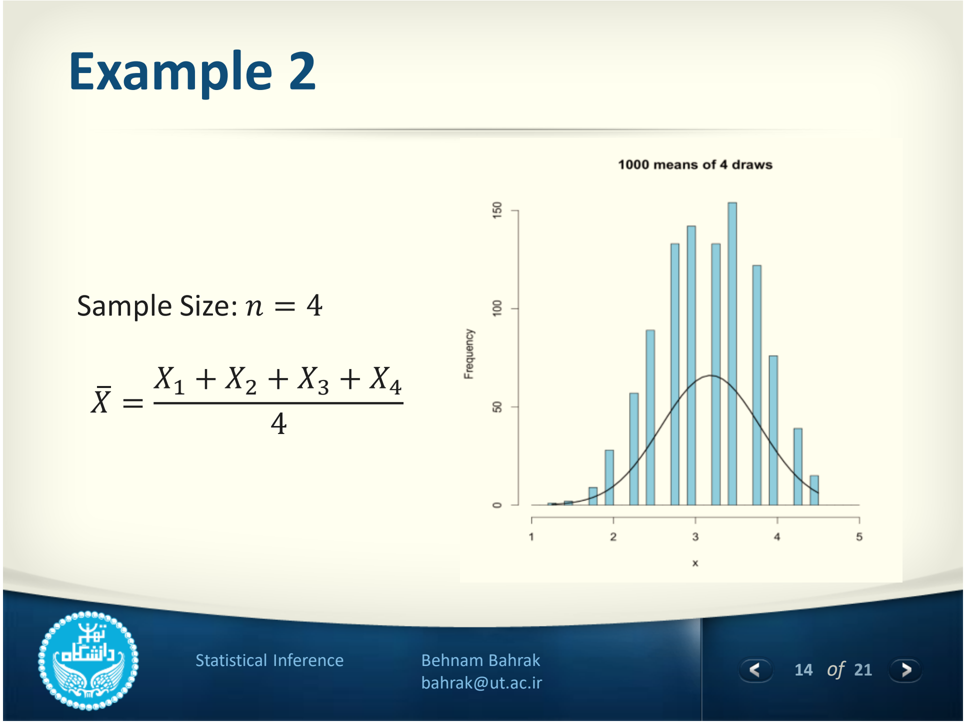
$$\bar{X} = \frac{X_1}{1}$$

1000 discrete draws

Statistical Inference

Behnam Bahrak
bahrak@ut.ac.ir

13 of 21



Example 2

Sample Size: $n = 4$

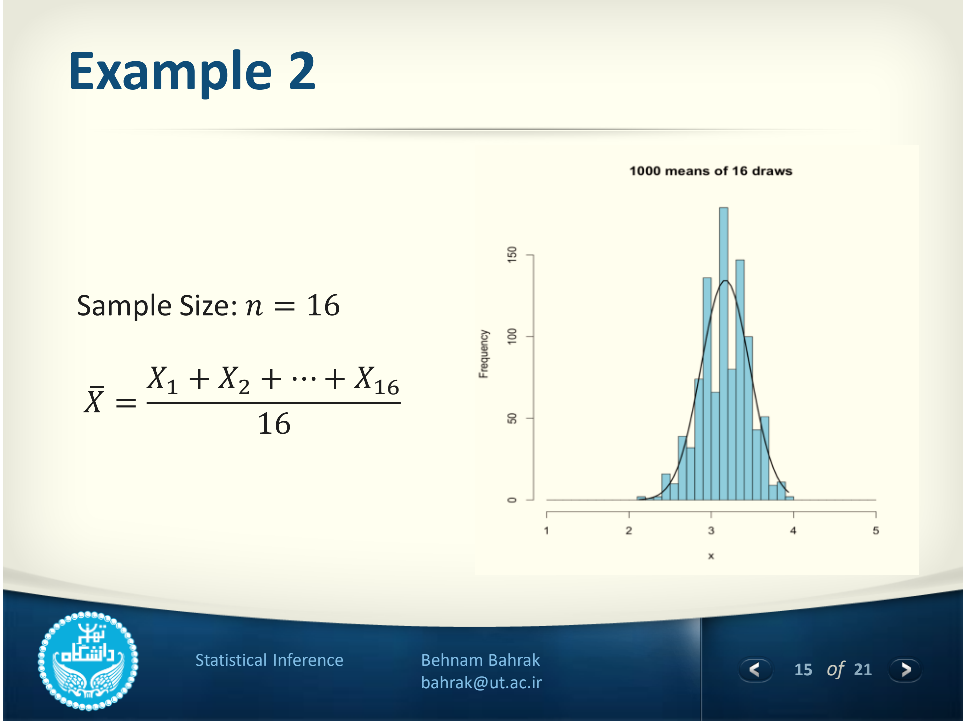
$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

1000 means of 4 draws

Statistical Inference

Behnam Bahrak
bahrak@ut.ac.ir

14 of 21



Example 2

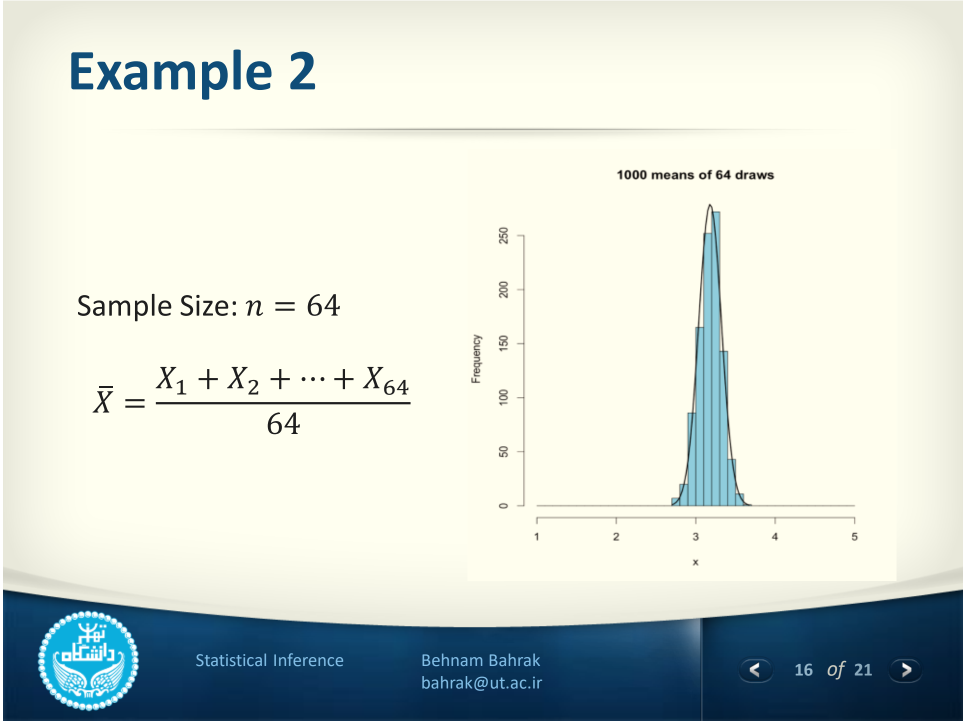
Sample Size: $n = 16$

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{16}}{16}$$

Statistical Inference

Behnam Bahrak
bahrak@ut.ac.ir

15 of 21



Example 2

Sample Size: $n = 64$

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_{64}}{64}$$

Statistical Inference

Behnam Bahrak
bahrak@ut.ac.ir

16 of 21

CLT Conditions

- Certain conditions must be met for the CLT to apply:

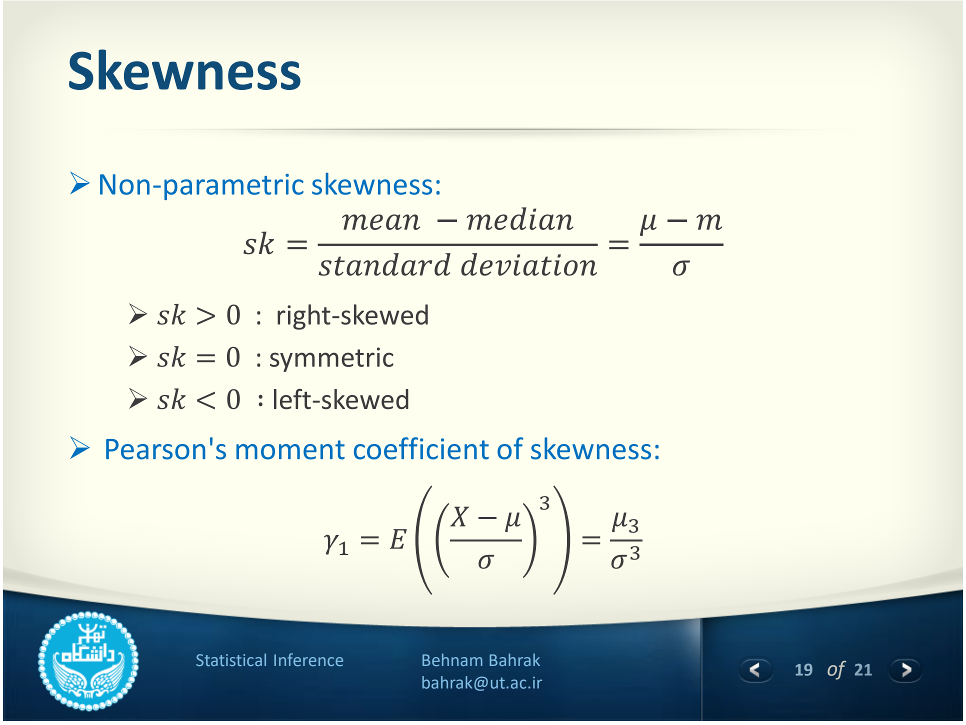
  1. Independence: Sampled observations must be independent. This is difficult to verify, but is more likely if:
     - random sampling/assignment is used, and
     - if sampling without replacement, $n < 10\%$ of the population.

  2. Sample size/skew: Either the population distribution is normal, or if the population distribution is skewed, the sample size is large.
     - the more skewed the population distribution, the larger sample size we need for the CLT to apply
     - for moderately skewed distributions $n > 30$ is a widely used rule of thumb

Justification for the Conditions

- If sampling without replacement, n needs to be less than 10% of the population. Why is this the case?
  - if we grab a very big portion of the population to be in our sample, its going to be very difficult to make sure that the sampled individuals are independent of each other.

- The more skewed the population distribution, the larger sample size we need for the CLT to apply. Why?
  - When the sample size is small, the sample means will be quite variable, and the shape of their distribution will mimic the population distribution.



Skewness

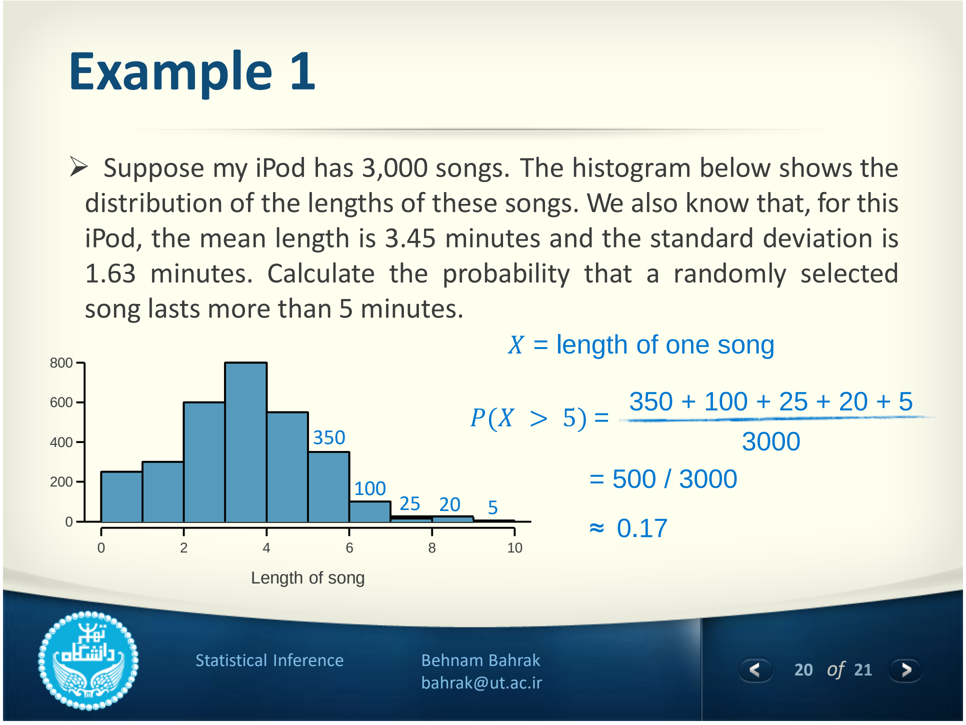
- Non-parametric skewness:

$$sk = \frac{mean\ - median}{standard\ deviation} = \frac{\mu - m}{\sigma}$$

  - $sk > 0$: right-skewed
  - $sk = 0$: symmetric
  - $sk < 0$: left-skewed

- Pearson's moment coefficient of skewness:

$$\gamma_1 = E\left(\left(\frac{X-\mu}{\sigma}\right)^3\right) = \frac{\mu_3}{\sigma^3}$$



Example 1

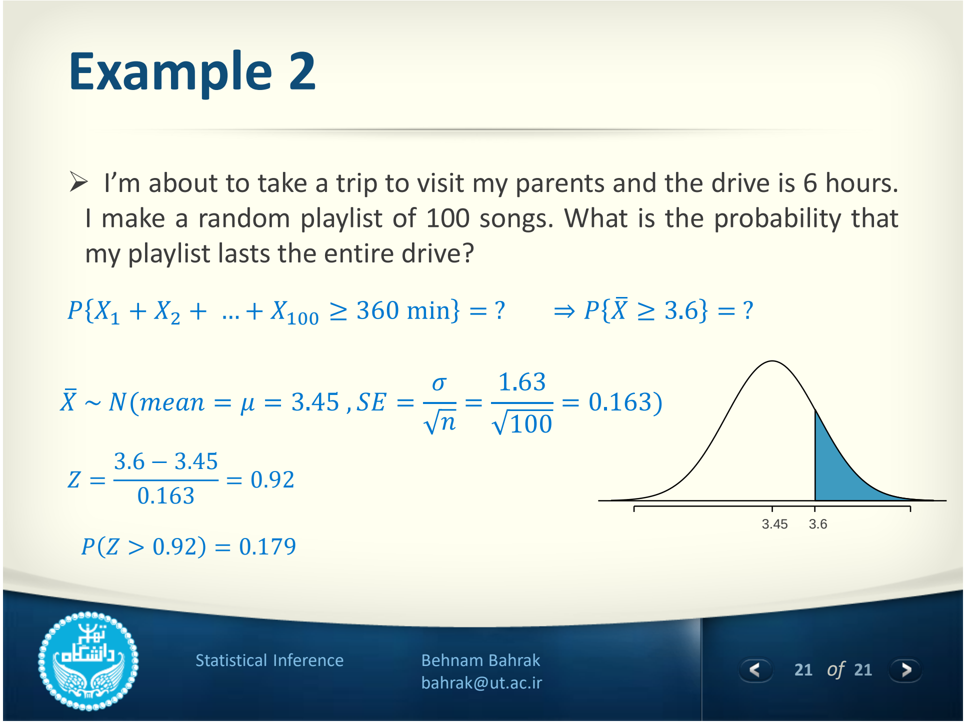
- Suppose my iPod has 3,000 songs. The histogram below shows the distribution of the lengths of these songs. We also know that, for this iPod, the mean length is 3.45 minutes and the standard deviation is 1.63 minutes. Calculate the probability that a randomly selected song lasts more than 5 minutes.

X = length of one song

$$P(X > 5) = \frac{350 + 100 + 25 + 20 + 5}{3000}$$

= 500 / 3000

≈ 0.17



Example 2

- I'm about to take a trip to visit my parents and the drive is 6 hours. I make a random playlist of 100 songs. What is the probability that my playlist lasts the entire drive?

$P\{X_1 + X_2 + \ldots + X_{100} \geq 360 \text{ min}\} = ? \qquad \Rightarrow P\{\bar{X} \geq 3.6\} = ?$

$\bar{X} \sim N(mean = \mu = 3.45\,, SE = \frac{\sigma}{\sqrt{n}} = \frac{1.63}{\sqrt{100}} = 0.163)$

$Z = \frac{3.6 - 3.45}{0.163} = 0.92$

$P(Z > 0.92) = 0.179$