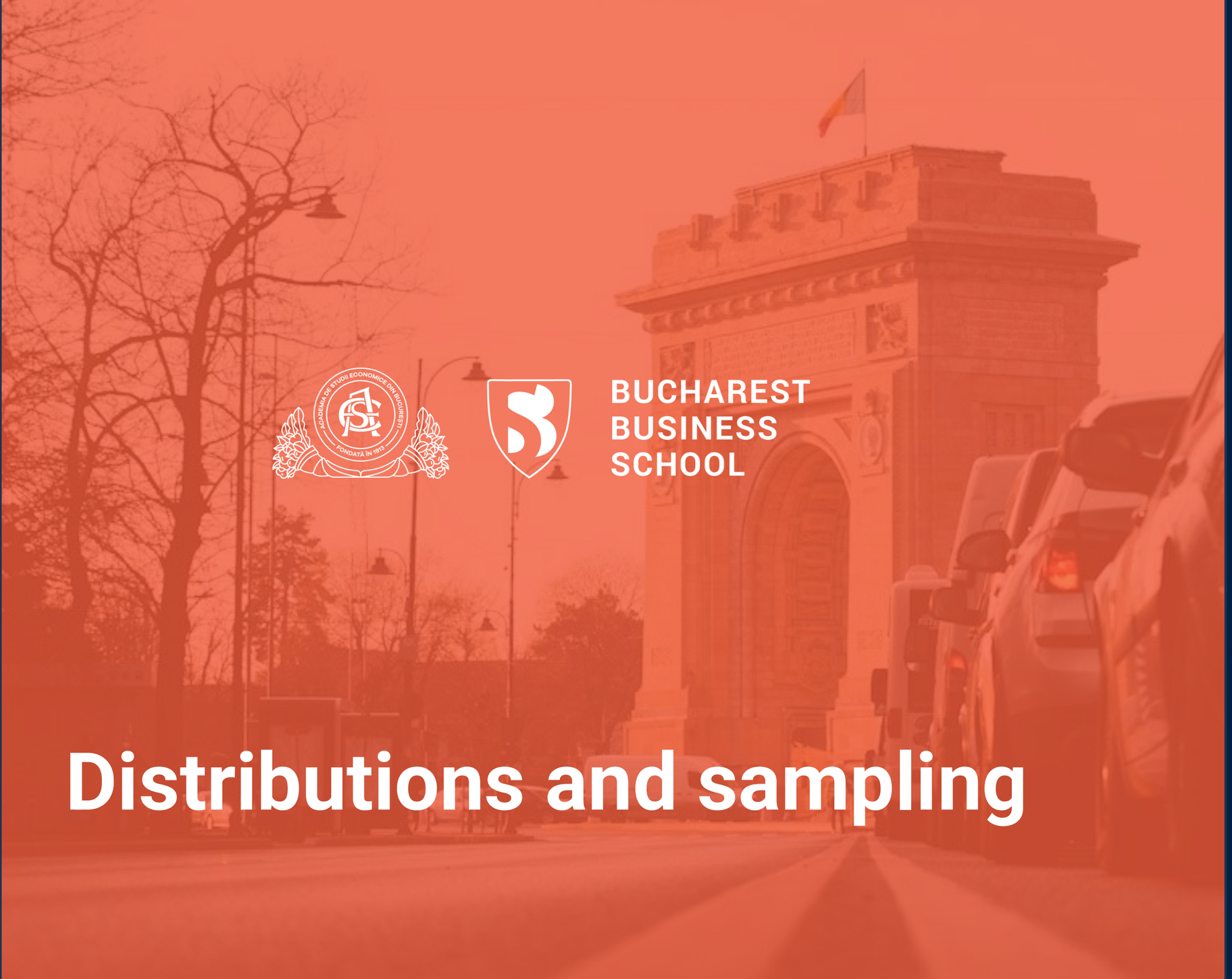


BUCHAREST BUSINESS SCHOOL

ASSOCIATION OF MBAS ACCREDITED

Leadership Through Value

Distributions and sampling

DATA MINING AND BIG DATA

CATALINA CHINIE
PH.D. ASSOCIATE PROFESSOR

SUMMARY

Data Mining and Big Data

Block 3

- Probability distributions and the normal distribution
- Sampling and confidence interval estimation

Variables

- In Statistics one operates with discrete and continuous variables.
- A probability distribution includes 2 elements.
- Element 1: alternative values of the variable
- Element 2: probability of occurrence for each alternative value.
- Discrete variables: x and p(x)
- Continuous variables: x and f(x)

Variables

- Discrete variables

$$P(X) \geq 0 \qquad \sum_x P(x) = 1$$

- If the two conditions are met one says that (x, p(x)) is a probability distribution.

Expected value (mean)

$$\mu = E(x) = \sum_x xP(x)$$

Variance

(how far away the data points fall from the center)

$$\sigma^2 = E(X-\mu)^2 = \sum_x (x-\mu)^2 P(x)$$

High variance – higher variability – dissimilar values with likelihood of extreme values

Low variance – lower variability – consistent values in the dataset

Standard deviation

(how tightly the data is clustered around the mean)

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_x (x-\mu)^2 P(x)}$$

Variables

Continuous variables

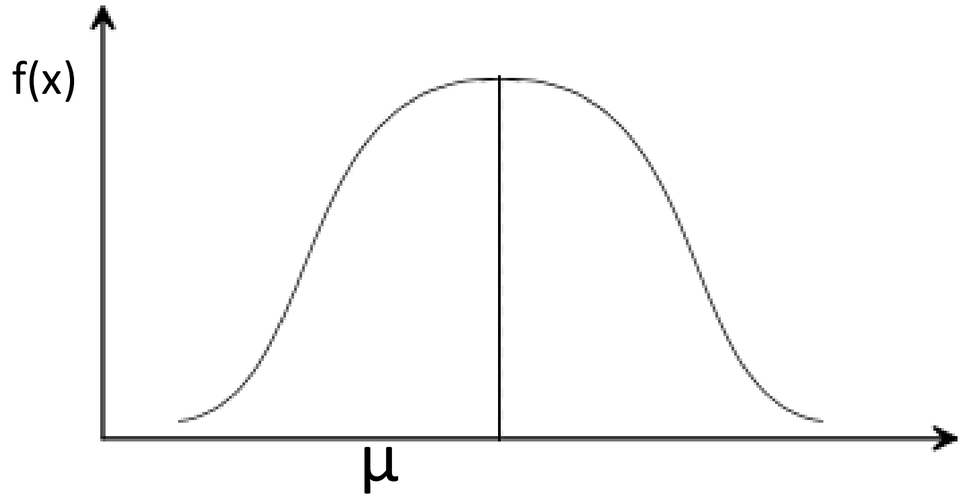
a) f(x) > 0 for all values of x

b) The area under the probability density function f(x) over all values of the random variable X is equal to 1.0

- If the two conditions are met one says that (x, f(x)) is a probability distribution.

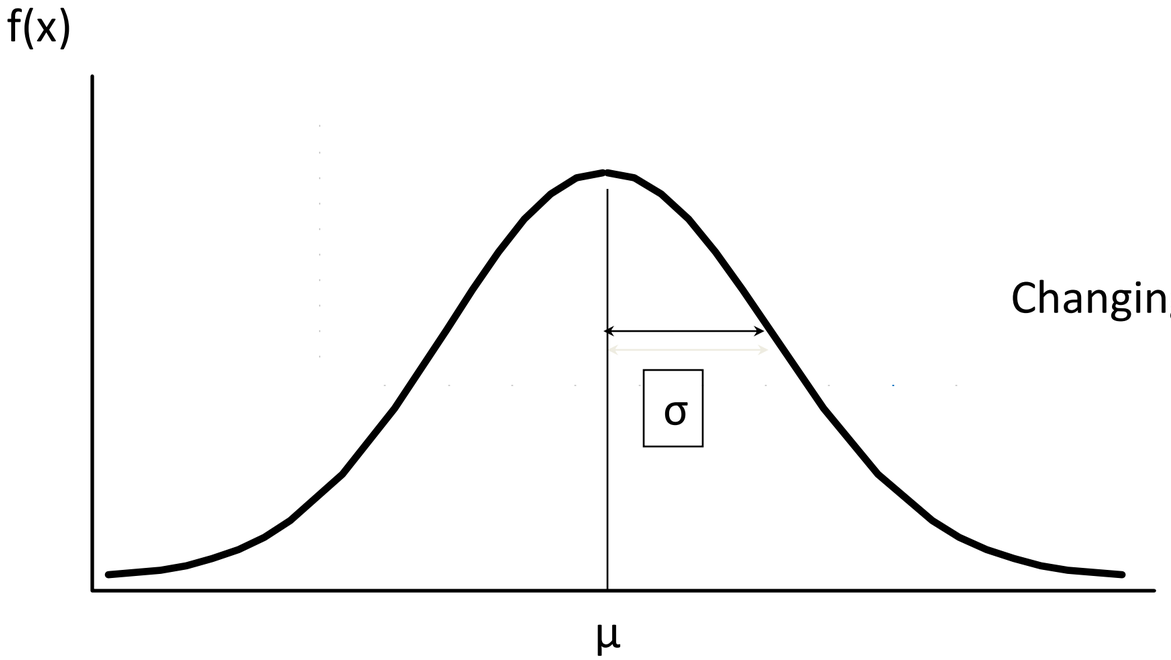
The Normal Distribution

- **Bell Shaped**
- **Symmetrical**
- **Central tendency indicators are equal**
- **Location is determined by the mean, μ**
- **Spread is determined by the standard deviation, σ**
- **The random variable has an infinite theoretical range: + ∞ to – ∞**



The Normal Distribution

Changing μ shifts the distribution left or right.



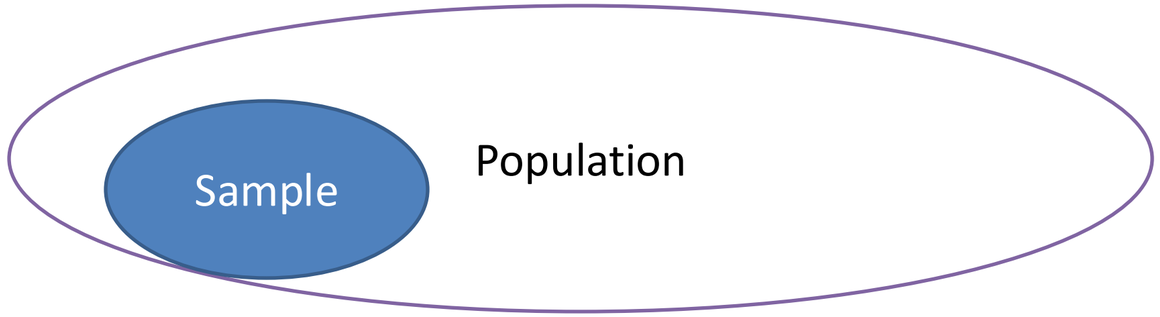
Changing σ increases or decreases the spread.

Mean μ and variance σ characterize the normal distribution $X \sim N(\mu, \sigma^2)$

Sampling

Making statements about a population using the results of a sample

Estimator		Parameter
(sample)	**Inference**	(population)
computable	→	not-computable



Sampling

A sampling distribution is a distribution of all of the possible values of a statistic for a given size sample selected from a population.

If a population is normally distributed with the mean μ and standard deviation σ, the sampling distribution of the mean is also normally distributed with the mean μ(x) and the variance σ(x)

$$\mu_{\bar{x}} = \mu$$ and $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The Central Limit Theorem states that the distribution of a variable X is approximately normal if n (sample size) is large enough, with mean miu and standard deviation sigma.

Confidence intervals

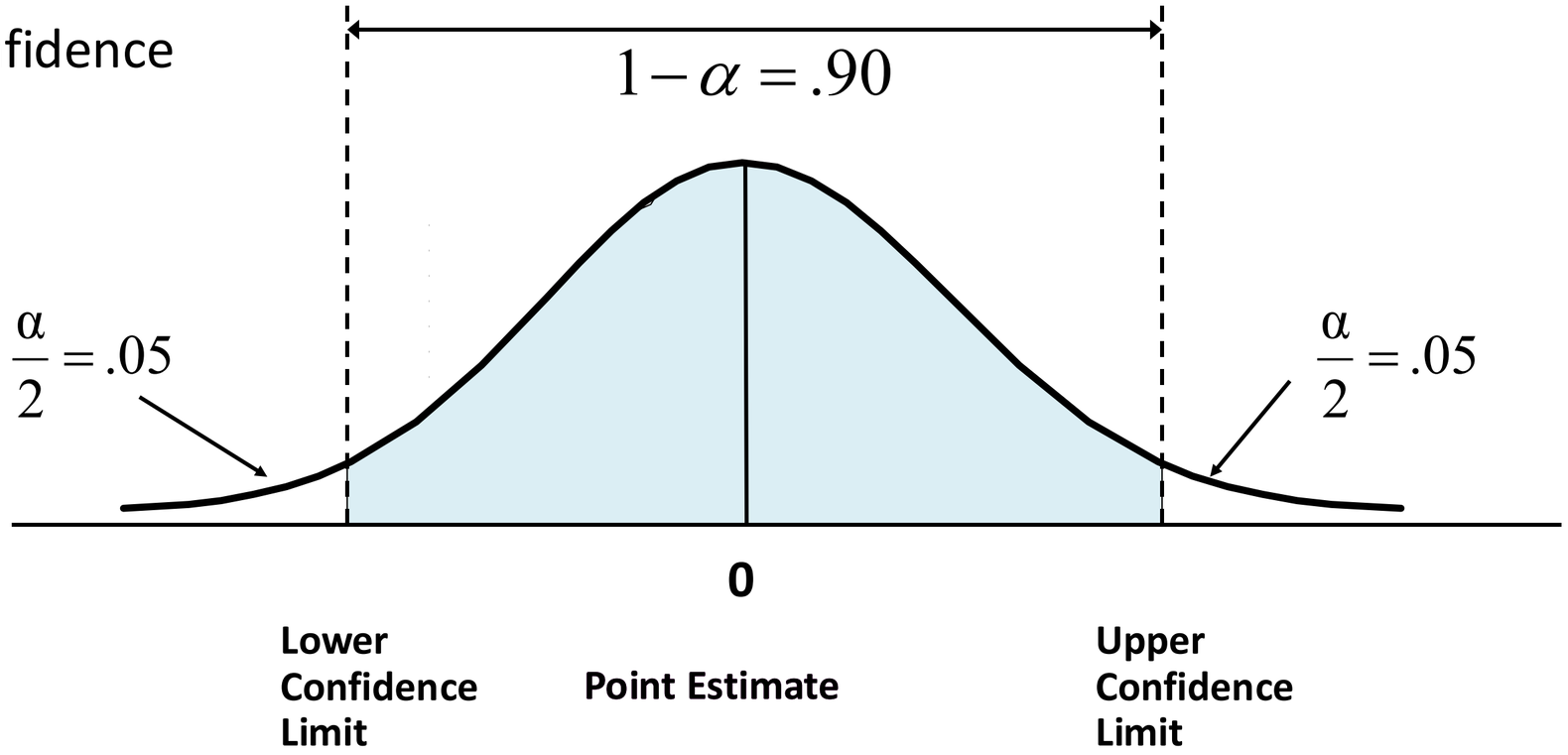
Confidence intervals incorporate the uncertainty and sample error to create a range of values the actual population value is likely to fall within

The general formula for all confidence intervals (for the mean – miu) is:

Point Estimate ± (Confidence)(Standard Error)

Confidence intervals

Consider a 90% confidence interval:



More resources

For more details please watch the following videos:

https://www.youtube.com/watch?v=cqK3uRoPtk0

https://www.youtube.com/watch?v=YXLVjCKVP7U

https://www.youtube.com/watch?v=CfZa1daLjwo

https://www.youtube.com/watch?v=iYiOVISWXS4

https://www.youtube.com/watch?v=rzFX5NWojp0

https://www.youtube.com/watch?v=FXZ2O1Lv-KE

https://www.youtube.com/watch?v=FXZ2O1Lv-KE

https://www.youtube.com/watch?v=hlM7zdf7zwU

https://www.youtube.com/watch?v=czdwHU27OqA

BUCHAREST
BUSINESS
SCHOOL