INTRAS

Co-funded by the
Erasmus+ Programme
of the European Union

“Advanced Statistics and Data Analysis”

Syllabus

Prof. Kyandoghere Kyamakya

Prof. Jean Chamberlain Chedjou

UNIVERSITÄT
KLAGENFURT

Intelligent Transport Systems: New ICT – based Master’s Curricula in Uzbekistan (INTRAS)
Agreement number: 2017-3516/001-001
Project reference number: 586292-EPP-1-2017-1-PL-EPPKA2-CBHE-JP

CONTENTS

- ❑ Objectives of the Lecture/Course
- ❑ Basic Instruments/Concepts used in the Course/Lecture
- ❑ Important scientific fronts covered by the Course/Lecture
- ❑ Important knowledge to be provided per chapter
- ❑ A Didactic example: **“Statistical analysis of stochastic phenomena”**

Course Goals and general issues/aspects covered in transportation

(Course Goals: Learning Objectives and Learning Outcomes)

Course Goals and general issues/aspects covered → (1)

Mathematics in traffic and transport ...

- ❑ Understanding what is a deterministic system
- ❑ Understanding what is a stochastic system
- ❑ Understanding the importance of statistics in the handling of randomness
- ❑ Understanding the analytical forms of the probability density functions and mastering of their application to model random scenarios with different complexities
- ❑ Learning how to construct a probability distribution for a random variable
- ❑ Learn how to calculate the average, the mode, the mean, the variance, and expected value for a discrete random variable.

Course Goals and general issues/aspects covered → (2)

- ❑ Learn how to plot the probability- and the cumulative-distribution functions of stochastic variables/scenarios
- ❑ Learn how to identify the properties of the normal distribution.
- ❑ Learn how to find probabilities for a normally distributed variable by transforming it into a standard normal variable.
- ❑ Learn how to find specific data values for given percentages, using the standard normal distribution: The Resident- and Traveling- Times Calculation.
- ❑ Learn how to apply the "Central Limit Theorem" to solve problems involving "sample means" for large samples
- ❑ Understanding the concept of „QUEUING“

Mathematics in traffic and transport ...

Course Goals and general issues/aspects covered → (3)

❑ Understanding the interest of „QUEUING“ in the processing of stochastic scenarios.

❑ Learn how to use a software/toolbox for the analysis of a „Queuing Process“.

❑ Learn how to find probabilities for a normally distributed variable by transforming it into a standard normal variable.

❑ Learn how to find specific data values for given percentages, using the standard normal distribution: The Resident- and Traveling- Times Calculation.

❑ Mastering techniques of “Approximation and fitting”: Norm approximation; Least-norm problems; Regularized approximation; Robust approximation; Function fitting and interpolation.

Mathematics in traffic and transport ...

Course Goals and general issues/aspects covered → (4)

- ❑ Learning selected advanced techniques for “Statistical estimation”: Parametric distribution estimation; Nonparametric distribution estimation; Optimal detector design and hypothesis testing; Chebyshev and Chernoff bounds ; Experiment design.

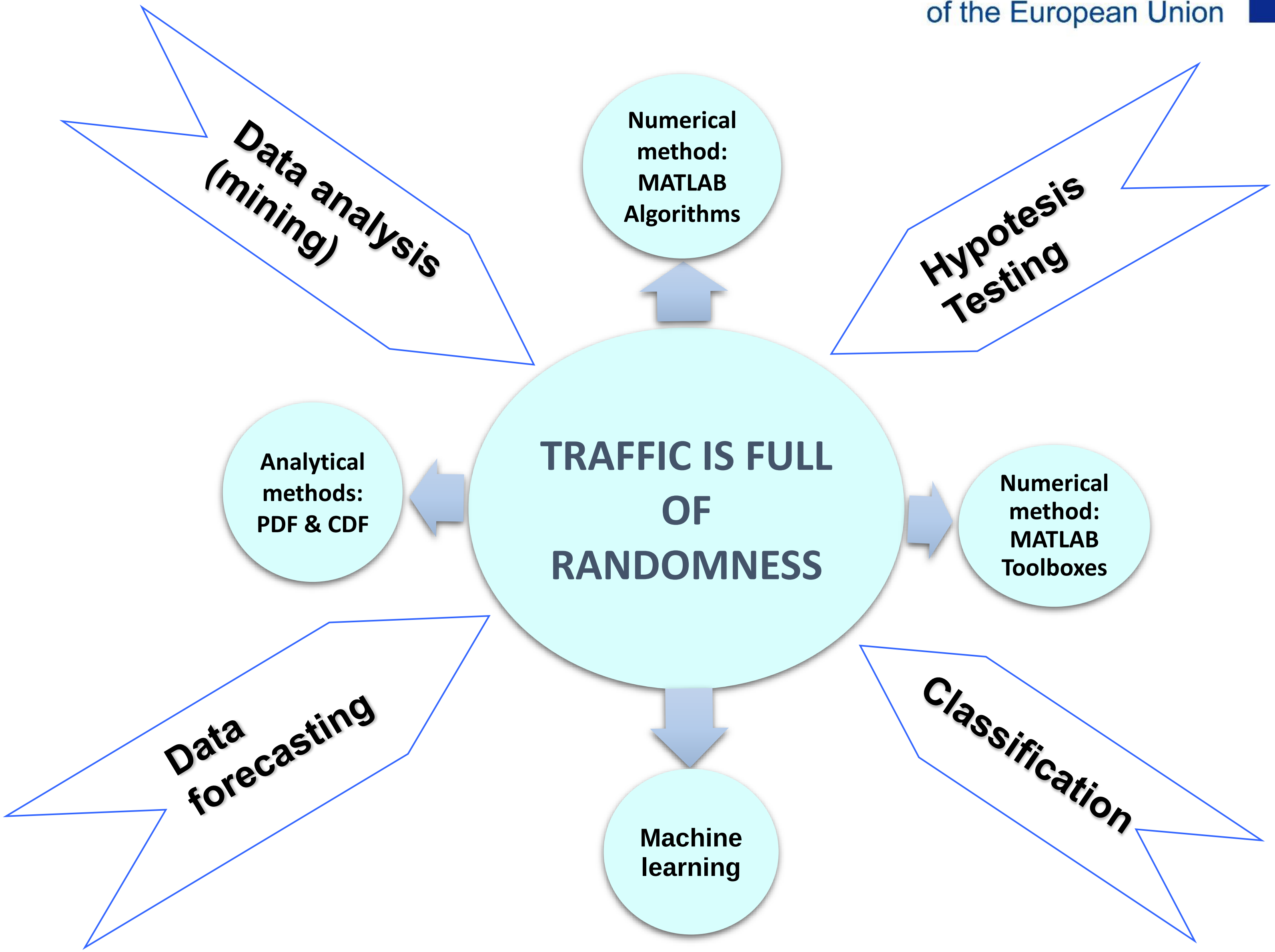
Mathematics in traffic and transport ...

- ❑ Understanding “Geometric problems” and mastering how to apply them for data classification. The following concepts and techniques must be well-understood: Projection on a set; Distance between; sets ; Euclidean distance and angle problems ; Extremal volume ellipsoids; Centering; Classification; Placement and location; Floor planning.

- ❑ Learn how to apply the Advanced techniques to the tackling (solving) of concrete (or Real-Life) problems selected in the field of Railway transportation, Road transportation, Supply chain Networks and Logistics. Application examples are given.

Basic Instruments/Concepts used in the Course/Lecture

Co-funded by the
Erasmus+ Programme
of the European Union



TRAFFIC IS FULL OF RANDOMNESS

Numerical method: MATLAB Algorithms

Analytical methods: PDF & CDF

Numerical method: MATLAB Toolboxes

Machine learning

Data analysis (mining)

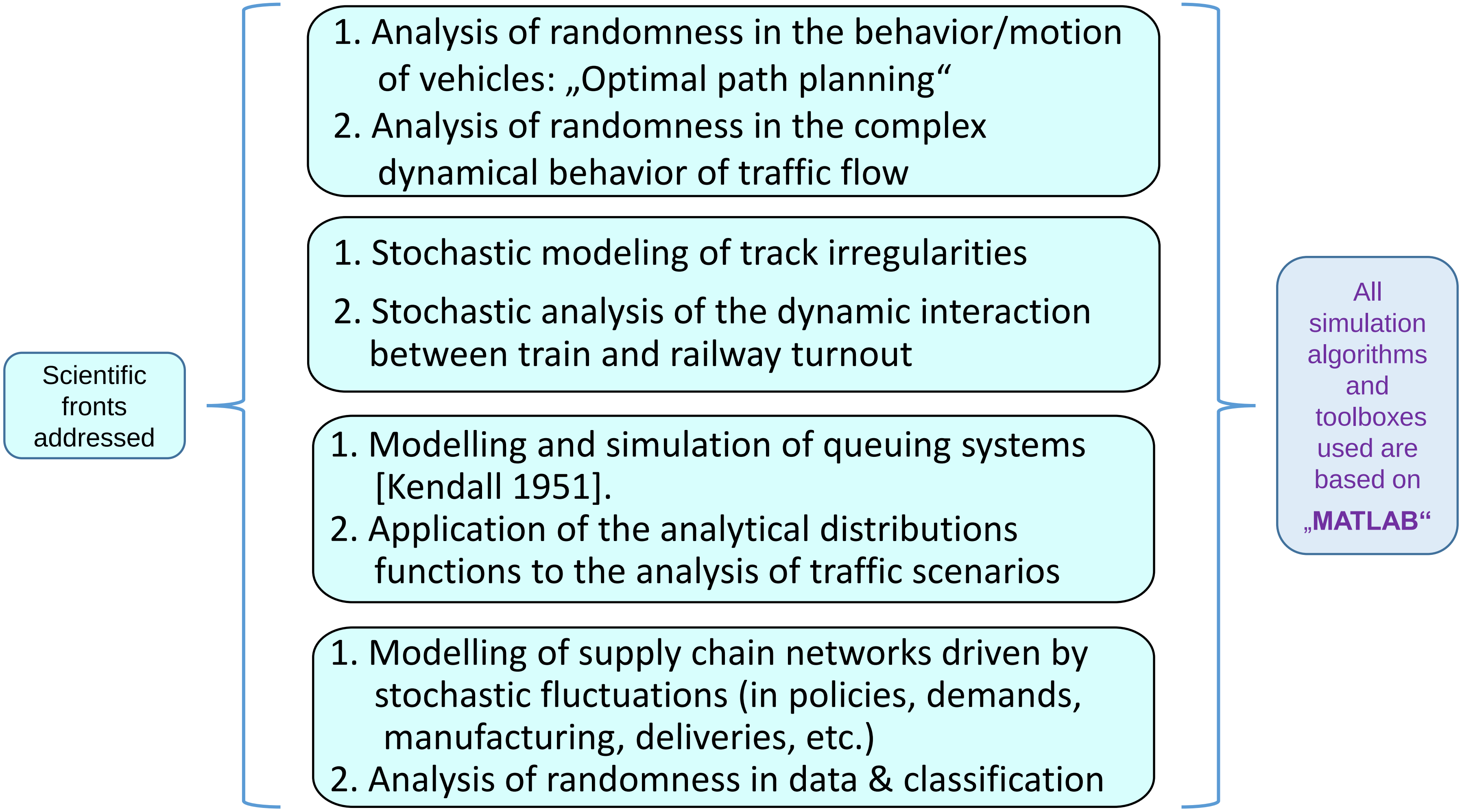
Hypotesis Testing

Data forecasting

Classification

Important scientific fronts covered by the Course/Lecture

Scientific fronts addressed in the Course



Scientific fronts addressed

1. Analysis of randomness in the behavior/motion of vehicles: „Optimal path planning“
2. Analysis of randomness in the complex dynamical behavior of traffic flow

1. Stochastic modeling of track irregularities
2. Stochastic analysis of the dynamic interaction between train and railway turnout

1. Modelling and simulation of queuing systems [Kendall 1951].
2. Application of the analytical distributions functions to the analysis of traffic scenarios

1. Modelling of supply chain networks driven by stochastic fluctuations (in policies, demands, manufacturing, deliveries, etc.)
2. Analysis of randomness in data & classification

All simulation algorithms and toolboxes used are based on

„MATLAB“

Important aspects to be covered in this Course/Lecture by each Lecturer

Important aspects to be investigated in each chapter

❑ Chapter 1. General introduction

- ✓ **Importance of statistics and data analysis** in road and railway transportation as well as in supply chain networks and Logistics.
- ✓ **Definition of the important keywords and concepts in statistics:** Statistics; Advanced statistics; Data; Data analysis; Deterministic system/scenario; Stochastic system/scenario; Distribution functions; Mean/Average; Variance; Moments of a distribution, Moments of order “k”, Standard deviation; Likelihood; Maximum Likelihood; Confidence interval; Time series, Data forecasting; etc. : Illustration through concrete examples selected in the field of transportation.
- ✓ **Analytical techniques of modelling.** Commonly used methods, concepts for data analysis: Mathematical expressions of distribution functions (exponential, shifted exponential, Markov, Gaussian, Poisson, etc.); Concepts: Bayesian estimation, Kalman filters, Polynomial Kalman filters, Polynomial chaos, etc.);
- ✓ **Time series forecasting.** Classical estimation techniques: Moving average (MA), Weighted moving average, Exponential smoothing, Autoregressive (AR), Autoregressive moving average (ARMA), Extrapolation, Linear prediction, Trend estimation (i.e., prediction of the variable as a linear or polynomial function of time).

Important aspects to be investigated in each chapter

❑ **Chapter 2. Statistical analysis of stochastic phenomena**

- ✓ **Introduction.** Definition of deterministic and stochastic systems; Deterministic versus Stochastic formalism; Importance and essence of the estimation process;
- ✓ **Fundamental parameters of stochastic processes/scenarios:** Mean/Average; Median; Mode; Variance; Standard deviation; Covariance; Percentile; Quantile; Quartile; Confidence level; Likelihood; Maximum Likelihood; Confidence interval.
- ✓ **Definition of the confidence interval** (CI)
- ✓ **Elements/Factors affecting the CI range**
- ✓ **Confidence limit for population mean**
- ✓ **Interval and level of confidence**
- ✓ **Confidence interval estimates**
- ✓ **Z-score Vs. Measurement scale**
- ✓ **Determination/Derivation of the CI**
- ✓ **Importance of the CI in Engineering**
- ✓ **Case studies.** C1: The Lecturer must choose 5 examples/exercises in Railway traffic; C2: Choose 5 examples/exercises in Road transportation; C3: Choose 5 examples/ exercises in Supply Chain Networks and Logistics. Solve all examples with students.

Remark. Full details of this chapter is presented (see below) in the frame of the **Didactic Lecture**.

Important aspects to be investigated in each chapter

- ❑ **Chapter 3. Basics of traffic theory: Fundamentals of queuing and simulation of queuing processes in stochastic scenarios/events selected in transportaion**
  - ✓ **Overview of traffic processes/phenomena.** Traffic; Main traffic problems; Traffic theory; The aim of traffic theory; Stochastic process; Random variable; Examples of random variables; Poisson process; Poisson distribution; Mathematical probability distribution functions; Main limitations of the stochastic theory; Methods used in traffic theory; Limits of the traffic theory methods;
  - ✓ **Probability distributions.** Exponential distribution (Markov process); Shifted exponential distribution; Poisson distribution; Relation between Poisson and exponential distribution; Erlang distribution.
  - ✓ **Motivation and overview of queuing.** Queuing system; Queuing theory; Selected applications of queuing in transportation; Fundamentals of queuing Systems.
  - ✓ **General queuing notation.** Five components of a queuing system (The Kendall notation 1951); Selected queuing models: M/M/1; M/M/2; M/M/n; D/D/1; D/D/2; D/D/n; M/D/1; M/D/2; M/D/n; D/M/1; D/M/2; D/M/n; G/G/1; G/G/2; G/G/n.
  - ✓ **State analysis of queue models/systems.** Poisson’s law (in: Physics, Operational research, communications; Transportation); Single-server queuing system; State and related probability; Geometric evolution; Application: Case of M/M/1 queue.

Important aspects to be investigated in each chapter

❑ Chapter 4. Approximation and fitting

- ✓ **Norm approximation**
  - Introducing the basic norm approximation, the penalty function approximation and the approximation with constraints.
- ✓ **Least-norm problems**
  - Presentation of the Least-norm problems; Examples for illustration.
- ✓ **Regularized approximation**
  - Description of the concepts of Bi-criterion formulation, Regularization, Reconstruction, smoothing, and de-noising.
- ✓ **Robust approximation**
  - Analysis of Stochastic robust approximation, and Worst-case robust approximation.
- ✓ **Function fitting and interpolation**
  - Description of selected important concepts: Function families; Constraints; Fitting and interpolation problems; Sparse descriptions and basis pursuit; Interpolation with convex functions.
- ✓ **Application exercises:** Several exercises are proposed and are methodically solved.

Important aspects to be investigated in each chapter

❑ **Chapter 5. Statistical estimation**

- ✓ **Parametric distribution estimation**
  - Presentation of the “Maximum likelihood estimation” and “Maximum a posteriori probability estimation”.
- ✓ **Nonparametric distribution estimation**
  - Presentation of the Nonparametric distribution estimation
- ✓ **Optimal detector design and hypothesis testing**
  - Presentation of deterministic and randomized detectors; Detection probability matrix; Optimal detector design; Multicriteria formulation and scalarization; Binary hypothesis testing; Robust detectors.
- ✓ **Chebyshev and Chernoff bounds**
  - Description of the concepts of Chebyshev bounds, and Chernoff bounds; presentation of some concrete application examples for illustration.
- ✓ **Experiment design**
  - Application exercises
- ✓ **Selected exercises with solutions** are proposed in this section for illustration.

Important aspects to be investigated in each chapter

❑ **Chapter 6. Geometric problems**

- ✓ **Projection on a set**
  - Analysis of the projection of a point on a convex set; separating a point and a convex set; Projection and separation via indicator and support functions.
- ✓ **Distance between sets**
  - Analysis of distance between sets: Computing the distance between convex sets; Separating convex sets; Distance & separation via indicator & support functions.
- ✓ **Euclidean distance and angle problems**
  - Study of selected concepts and problems: Gram matrix and realizability; Problems involving angles only; Euclidean distance problems.
- ✓ **Extremal volume ellipsoids**
  - Study of three important concepts: The Löwner-John ellipsoid; Maximum volume inscribed ellipsoid; Affine invariance of extremal volume ellipsoids.
- ✓ **Centering**
  - Analysis of selected concepts for centering: Chebyshev center; Maximum volume ellipsoid center; Analytic center of a set of inequalities.
- ✓ **Classification**
  - Study of linear and nonlinear classification techniques.

Important aspects to be investigated in each chapter

- ✓ **Placement and location**
  - Study of linear and nonlinear location problems without constraints. Extension to the study of location problems with path constraints.
- ✓ **Floor planning**
  - Study of selected case studies: Relative positioning constraints; Floor planning via convex optimization; Floor planning via geometric programming.

❑ **Chapter 7. Selected concrete scenarios as application examples in transportation**

- ✓ **Supply Chain Networks -** Selected concrete scenarios are:
  - Dynamic supply chains with stochastic policies
  - Dynamic supply chains with stochastic demands
  - Modelling of a supply chain network driven by stochastic fluctuations
- ✓ **Railway transportation -** Selected concrete scenarios are:
  - Stochastic analysis of dynamic interaction between train and railway turnout
  - Simulation of train track interaction with stochastic track properties
  - Stochastic modeling of track irregularities using experimental measurements
- ✓ **Road transportation -** Selected concrete scenarios are:
  - (1) Stochastic modeling and simulation to vehicle system dynamics; Stochastic modeling and simulation of traffic flow; (3) Chaotic behavior of traffic flow.

INTRAS

Co-funded by the
Erasmus+ Programme
of the European Union

A didactic lecture on the topic:

"Statistical analysis of stochastic phenomena"

(Analysis of selected examples for illustration)

Univ.-Prof. Dr.-Ing. Kyandoghere Kyamakya
Kyandoghere.Kyamakya@aau.at

Assoc. Prof. PD. Dr. Dr.-Ing. Jean Chamberlain Chedjou
jean.chedjou@aau.at

UNIVERSITÄT
KLAGENFURT

Intelligent Transport Systems: New ICT – based Master's Curricula in Uzbekistan (INTRAS)
Agreement number: 2017-3516/001-001
Project reference number: 586292-EPP-1-2017-1-PL-EPPKA2-CBHE-JP

Contents

- ❑ Objectives
- ❑ Introduction
  - ✓ What is a deterministic system?
  - ✓ What is a stochastic system?
  - ✓ “Deterministic formalism” versus “Stochastic formalism”
  - ✓ Why is the estimation process important?
  - ✓ What is the essence of an estimation process?
- ❑ Fundamental parameters of stochastic processes/scenarios
  - ✓ Mean
  - ✓ Median
  - ✓ Mode
  - ✓ Variance
  - ✓ Standard deviation
  - ✓ Covariance
  - ✓ Percentile
  - ✓ Quartile
  - ✓ Confidence interval (CI)

Contents

❑ Definition of the confidence interval (CI)

❑ Elements/Factors affecting the CI range

❑ Confidence limit for population mean

❑ Interval and level of confidence

❑ Confidence interval estimates

❑ Z-score Vs. Measurement scale

❑ Determination/Derivation of the CI

❑ Importance of the CI in Engineering

❑ Case study 1. choice of 5 examples/exercises in Railway traffic;

❑ Case study 2. Choice of 5 examples/exercises in Road transportation;

❑ Case study 3. Choice of 5 examples in Supply Chain Networks and Logistics.

❑ Recommendation. All three case studies must be solved by the Lecturer as illustration of how to analyze concrete stochastic problems in transportation.

1. Objectives

- □ Understanding what is a deterministic system
- □ Understanding what is a stochastic system
- □ Learning how to test deterministic/stochastic systems in engineering
- □ Understanding the importance of the probabilistic analysis
- □ Learning construction of the probability distribution for a random variable (using experimental data)
- □ Learning construction of the cumulative distribution for a random variable (using experimental data)
- □ Learning calculation of selected parameters of a distribution: Average/ Mean; Mode; Median; Variance; Standard deviation; Likelihood; Percentile; Quartile; Z-score value; Confidence interval (CI).
- □ Mastering plotting of the probability density functions and cumulative distribution functions describing stochastic variables/scenarios.
- □ Mastering the identification of the properties of a normal distribution.

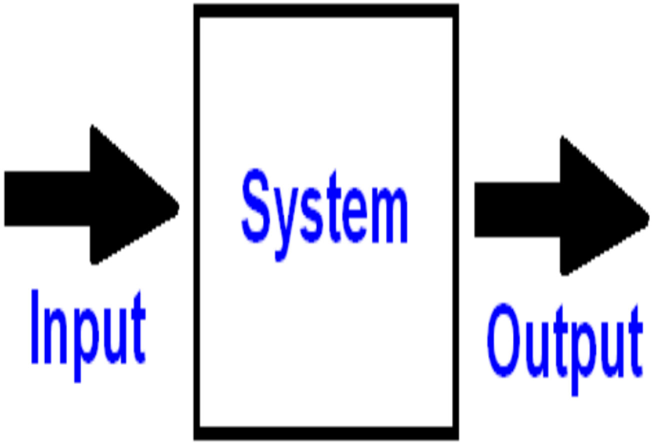
1. Objectives

- □ Learning derivation of probabilities for a normally distributed variable by transforming it into a standard normal variable.
- □ Learning calculation of specific data values for given percentages, using the standard normal distribution: The Resident- and Travel- Times Calculation.
- □ Learning application of the “Central Limit Theorem” to solve problems involving “sample means” for large samples.
- □ Understanding the analytical expressions of the “mean/average” and “standard deviation” expressed into continuous forms.
- □ Understanding what is the Likelihood and learning how to calculate the Likelihood analytically.
- □ Application: Consideration of concrete application examples in: (1) Railway traffic, (2) Road traffic, and (3) Supply chain networks to be solved as case- studies (for both illustration and validation of the theoretical concepts developed in this chapter).

2. “Deterministic formalism” versus “Stochastic formalism”

What is a deterministic system?

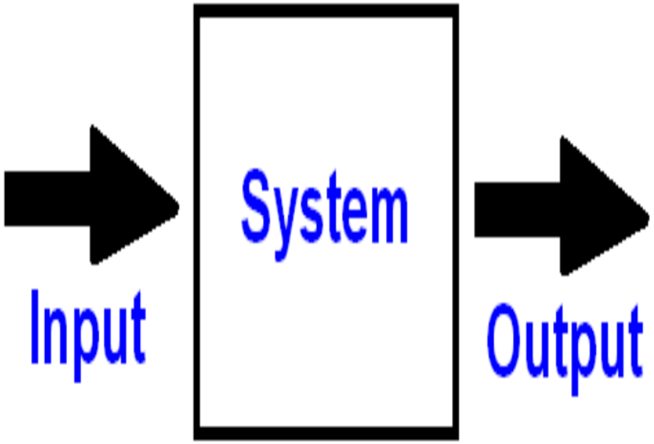
- ❑ A system that produces the same output for a given starting condition (or input).
- ❑ A system in which the next state is uniquely determined by the current state.
- ❑ A system with a fully predictable state/Behavior (i.e. no randomness is observed in the behavior of the system).
  - ✓ **Examples:**
    - ○ Classical Mechanics: Pendulum, motors, etc. Electrical engineering: Electrical- and Electronic circuits, etc.



2. “Deterministic formalism” versus “Stochastic formalism”

What is a stochastic system?

- ❑ A system that produces different output for a given starting condition (or input) (Note: this is observed through repeated measurements).
- ❑ A system in which the next state cannot be predicted/determined by the current state.
- ❑ A system in which the next state (output) is only probabilistically determined by the current state (i.e. there are several possible next states that can occur subsequently to the same activity (input), each with a given probability.

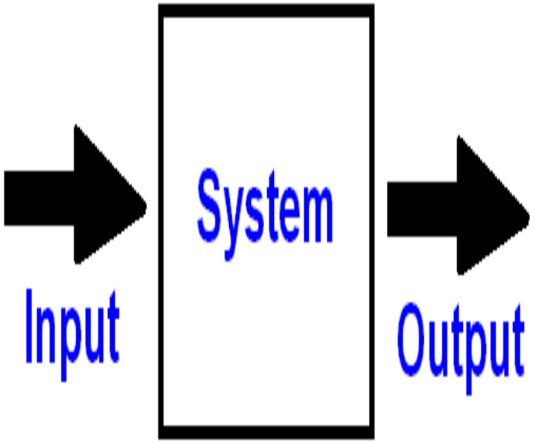


✓ **Examples of stochastic phenomena:**

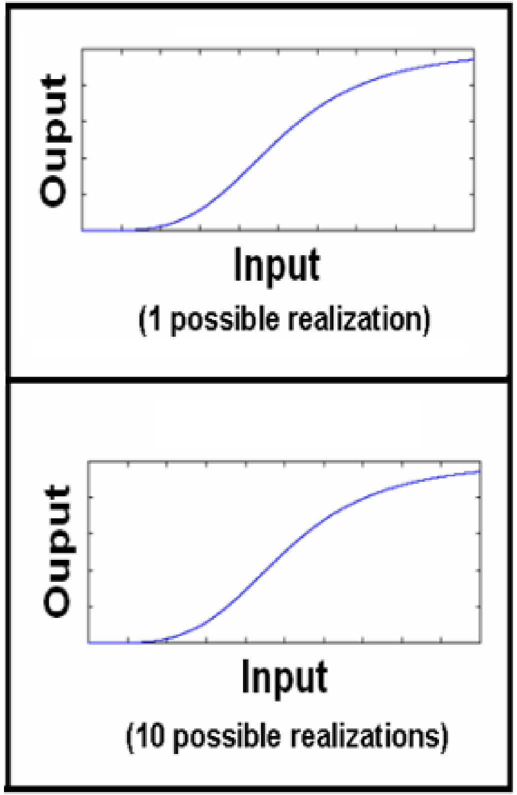
- ○ Phenomena/scenarios in transportation and traffic Engineering.
- ○ Phenomena in Quantum Mechanics, etc..

2. “Deterministic formalism” versus “Stochastic formalism”

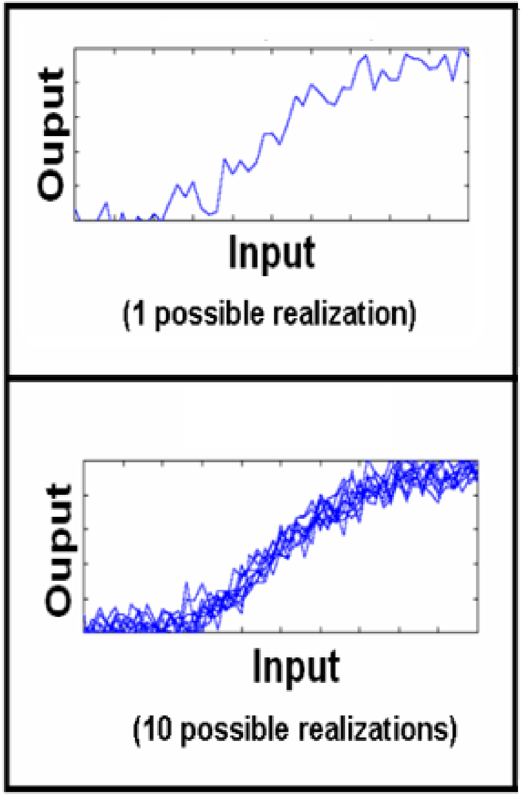
Experimental checking/test of a “Deterministic system” or a “Stochastic system”



“Deterministic system”



“Stochastic system”

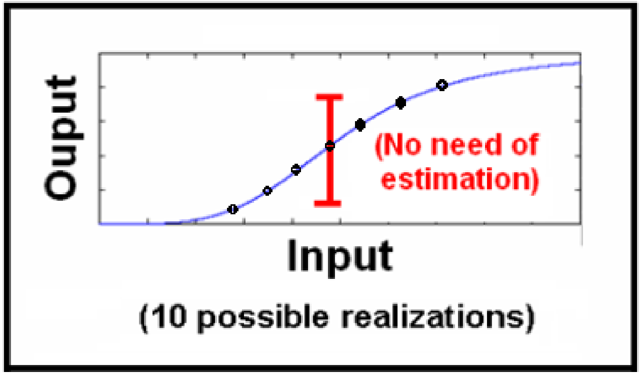


3. Importance and essence of the estimation process in data analysis

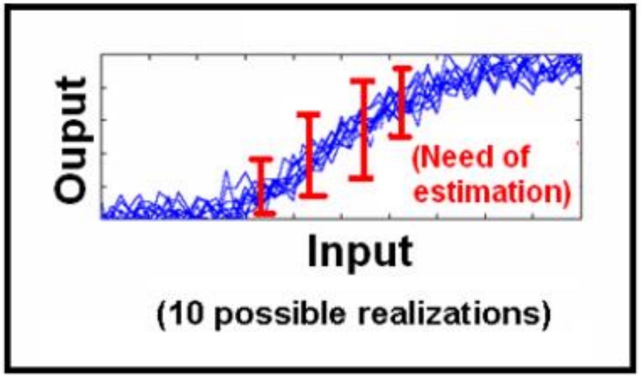
Why is the estimation process important ?

- ❑ **The estimation process is important to** analyze stochastic (or random) scenarios/events.
- ❑ **The result of the „estimation process“ provides** a range around the „Mean“.
- ❑ **This range generally reveals** the domain in which the expected values can be located. This range, which is called „Confidence Interval“ (CI) is obtained at/for a specific level of confidence.

“Deterministic system”

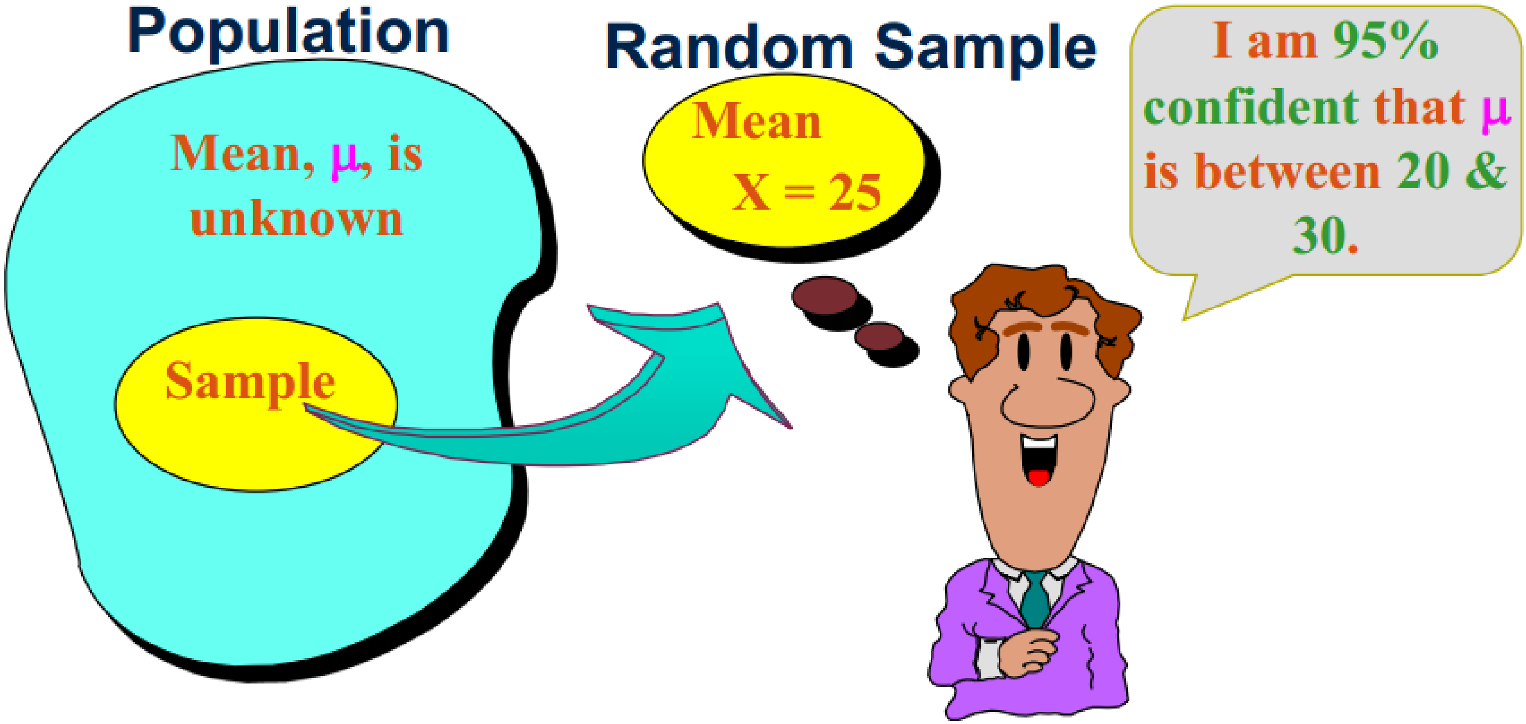


“Stochastic system”



3. Importance and essence of the estimation process in data analysis

Why is the estimation process important ?



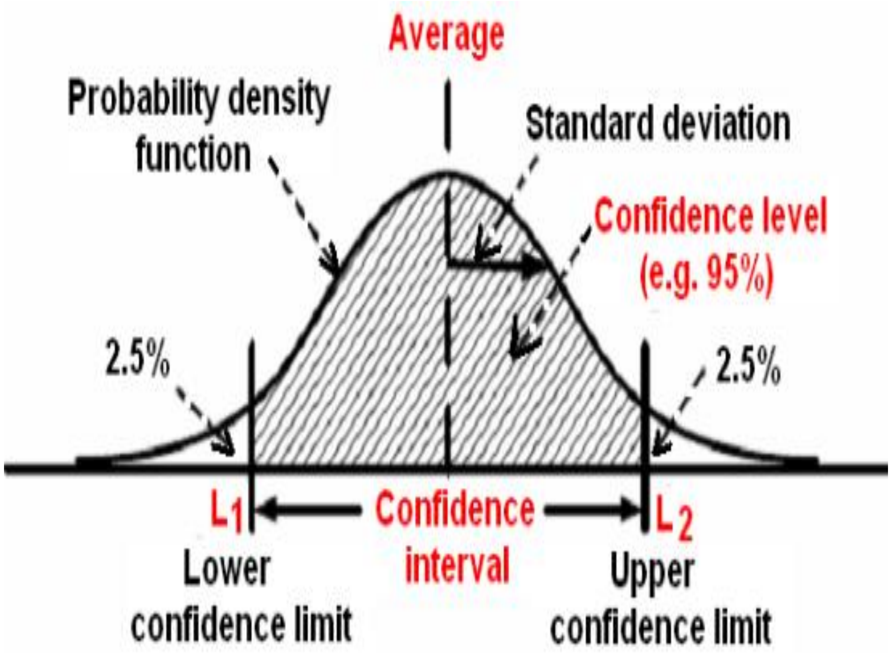
4. Fundamental parameters of a stochastic process and measurement

→ (1)

What are the fundamental/key parameters for a "Population Estimation"?

- ❑ **The fundamental parameters are** key quantities in the analysis, prediction and control of stochastic processes/scenarios

- ❑ **Cases of Univariate/Multivariate processes**
  - ✓ Mean
  - ✓ Median
  - ✓ Mode
  - ✓ Variance /Covariance
  - ✓ Standard deviation
  - ✓ Quantile
  - ✓ Confidence interval



4. Fundamental parameters of a stochastic process and measurement → (2)

Analytical expression of the fundamental/key parameters for a "Population Estimation"?

- ❑ Mean: $$\mu = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
- ❑ Mode/Top: $$x_{\mathrm{mod}} = \arg\max_{x} \hat{f}(x)$$
- ❑ Median $$x_{med} = \begin{cases} x_{((\mathrm{n}+1)/2)} & \text{if } n \text{ is odd} \\ \dfrac{x_{(n/2))+} x_{((n/2)+1))}}{2} & \text{if } n \text{ is even} \end{cases}$$
- ❑ Variance: $$Var(X) = E\left[(X-\mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

4. Fundamental parameters of a stochastic process and measurement → (3)

Analytical expression of the fundamental/key parameters for a "Population Estimation"?

❑ Standard deviation $$\sigma = \sqrt{E\left[(X-\mu)^2\right]} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}$$

❑ Confidence interval $$CI = \mu \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

❑ Covariance $$Cov(X,Y) = E(XY) - \mu_X . \mu_Y$$

$$\mu_X = E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

$$E(X) = \sum_{i=1}^{n} x_i . p_i$$

4. Fundamental parameters of a stochastic process and measurement → (4)

Analytical expression of the fundamental/key parameters for a "Population Estimation"?

❑ Moment of order 1 (Mean/Average)

$$\mu = \int_{-\infty}^{+\infty} x \cdot f(x)dx$$

❑ Moment of order "k"

$$\mu_{k} = \int_{-\infty}^{+\infty} x^{k} \cdot f(x)dx$$

❑ Probability density function (PDF)

$$f(x) = \frac{dF(x;\lambda)}{dx}$$

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$

❑ Cumulative density function (CDF)

$$F(x;\lambda) = \int_{-\infty}^{x} f(\tau)d\tau$$

$$\lim_{x \to +\infty} F(x;\lambda) = 1 \qquad \lim_{x \to -\infty} F(x;\lambda) = 0$$

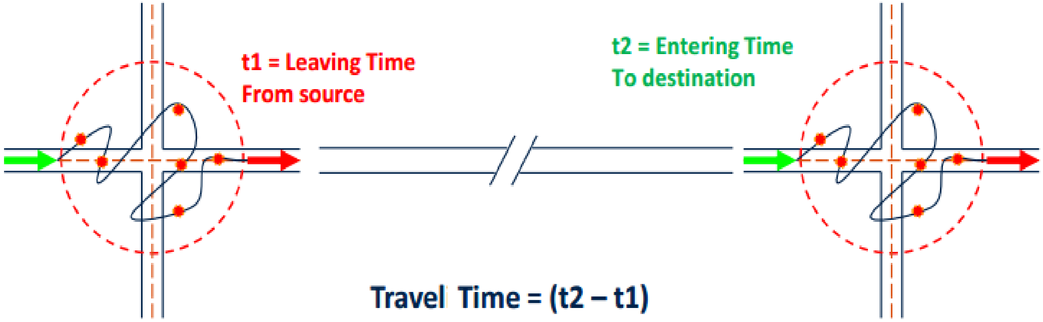
4. Fundamental parameters of a stochastic process and measurement

→ (5)

❑ Exercise

- ✓ The table below gives the travel time of vehicles between two junctions/ nodes (or the travel time of trains between two neighboring stations).
  - ○ Calculate the Mean, the variance, the standard deviation, the median, the PDF, and the CDF, the mode of the distribution.

Mathematics in traffic and transport ...



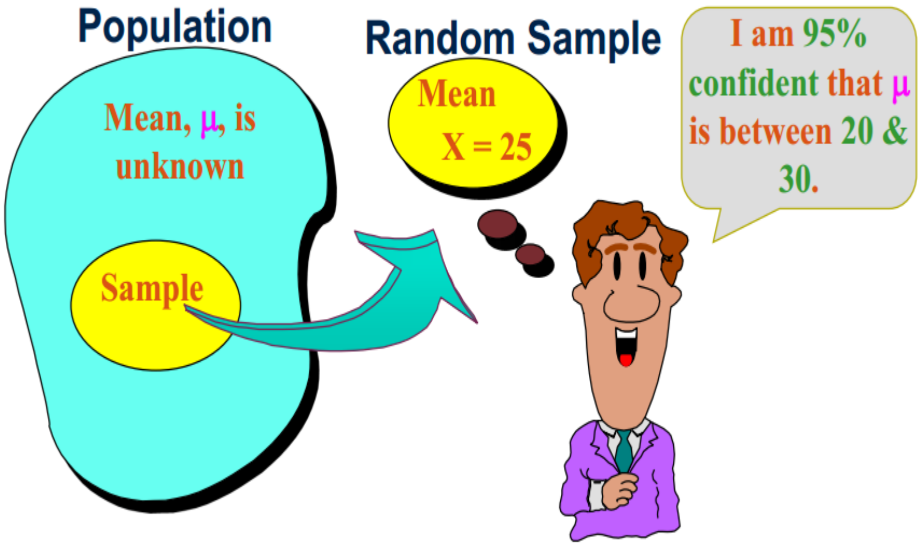
Note: This table contains fictitious data

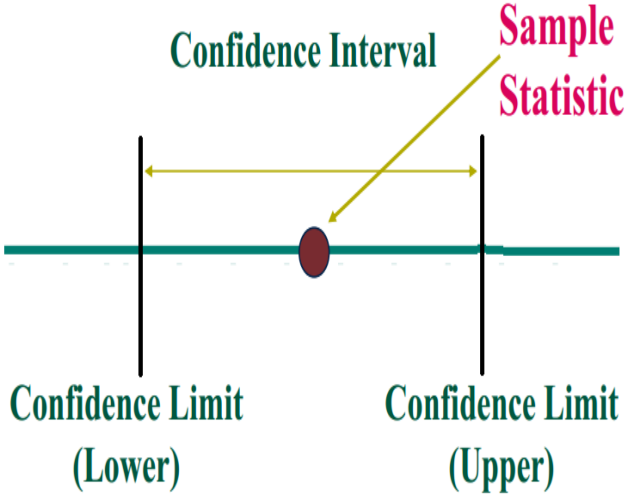
Number of cars/trains	1	5	6	8	6	3	1
Travel time (s)	10	20	25	30	40	50	80

Mean=32.3333s
std= 13.3735s

5. Estimation of the confidence interval (CI)

- **The estimation of the confidence interval -**
  - ✓ Provides range of values based on observations from 1 sample . This range is stated in terms of probability (because never 100% sure)
  - ✓ Gives information about closeness to unknown population parameter
  - ✓ A probability that the population parameter falls somewhere within the interval.





CI = [Lower Band , Upper Band]

6. Confidence limits for population mean and importance of the CI

- ❑ **Confidence limits:**

$$\text{Parameter} = \text{Statistic} \pm \text{Its } \mathit{Error}$$

$$\mu = \overline{X} \pm \mathit{Error}$$

$$\overline{X} - \mu = \mathit{Error} = \mu - \overline{X}$$

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}} = \frac{\mathit{Error}}{\sigma_{\overline{X}}}$$

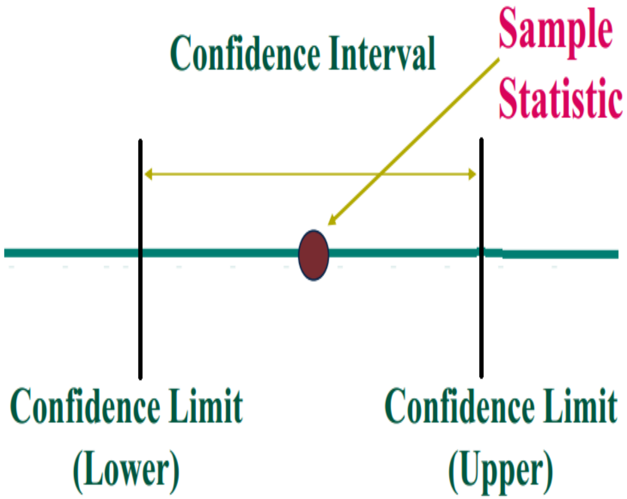
$$\mathit{Error} = Z\sigma_{\overline{x}}$$

$$\mu = \overline{X} \pm Z\sigma_{\overline{X}}$$

- ❑ **Importance of the confidence interval (CI)**
  - ✓ The confidence interval provides the range/window in which the expectation is located. This range contains the acceptable values of the random variable.

7. Elements of the “confidence interval estimation” and importance of the Z-score value

- ❑ **The parameters needed to calculate the CI are:**
  - ✓ The mean
  - ✓ The standard deviation
  - ✓ The sample size
  - ✓ The confidence level (Quantile)
  - ✓ The Z-score values
  - ✓ The sampling error
- ❑ **Importance of Z-score**
  - ✓ The value of Z-score significantly affects the length of the confidence interval
  - ✓ The Z-score value is obtained for a fixed/known value of the quantile.
  - ✓ Various tables of Z-score are available in the literature and each value of a quantile is used to obtain the corresponding Z-score value.



CI = [LB , UB]

8. Z-score tables: Description, reading techniques, and importance in data analysis

- ❑ Learning how to adapt the value of a quantile to a specific table of Z-score.
- ❑ Learning how to read the various Z-score tables available in the literature
- ❑ Each area of the figures below corresponds to a specific table of Z-score
- ❑ For a given/known/fixed value of a quantile the corresponding value of Z-score is same/unique when reading through all tables defined by the figures below. **It should be clearly explained how to read the Z-score value through different tables below.**

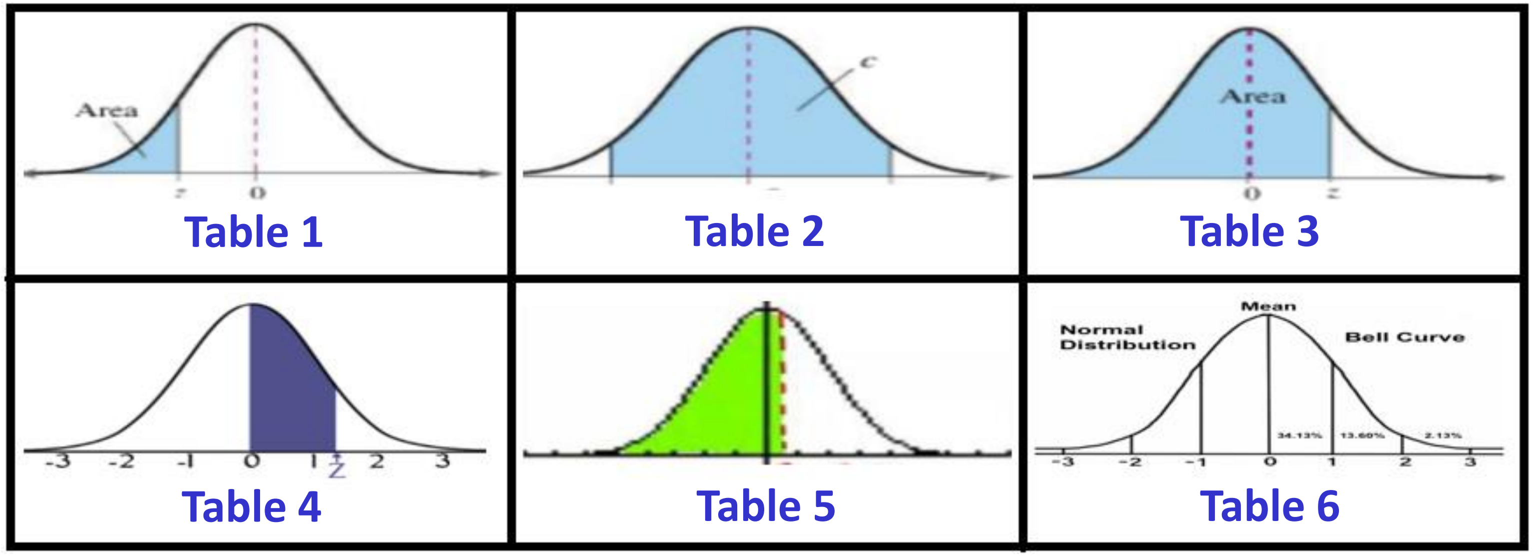


Table 1

Table 2

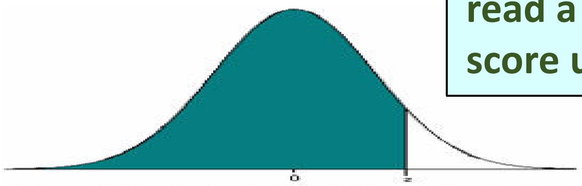
Table 3

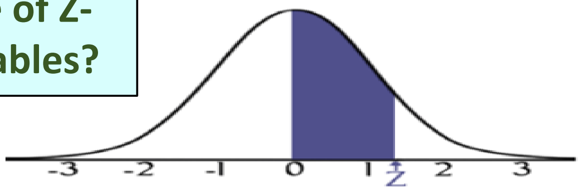
Table 4

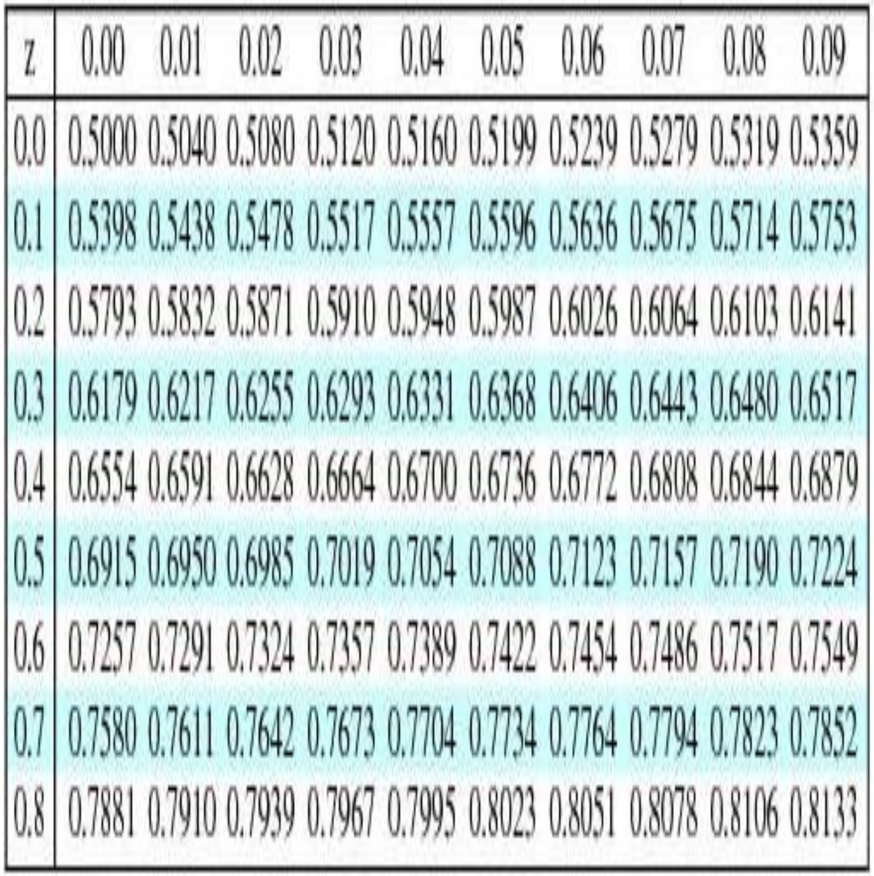
Table 5

Table 6

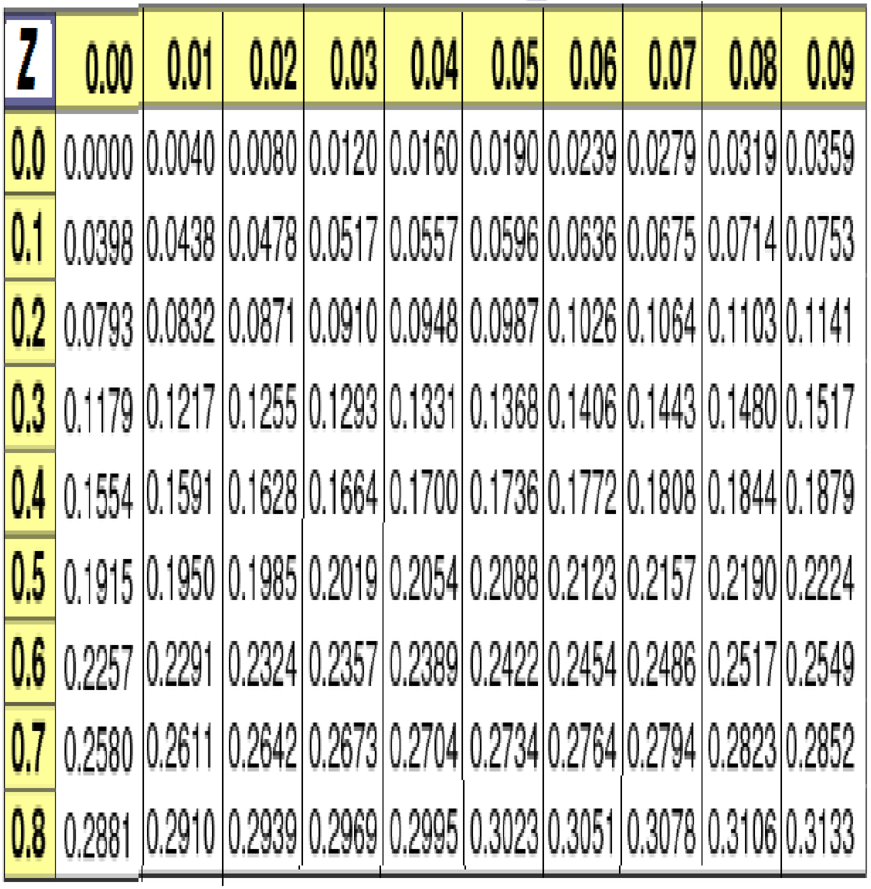
Application example. How to read a unique (same) value of Z-score using two different tables?





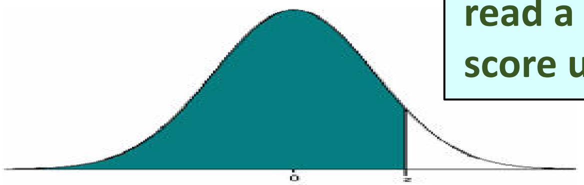


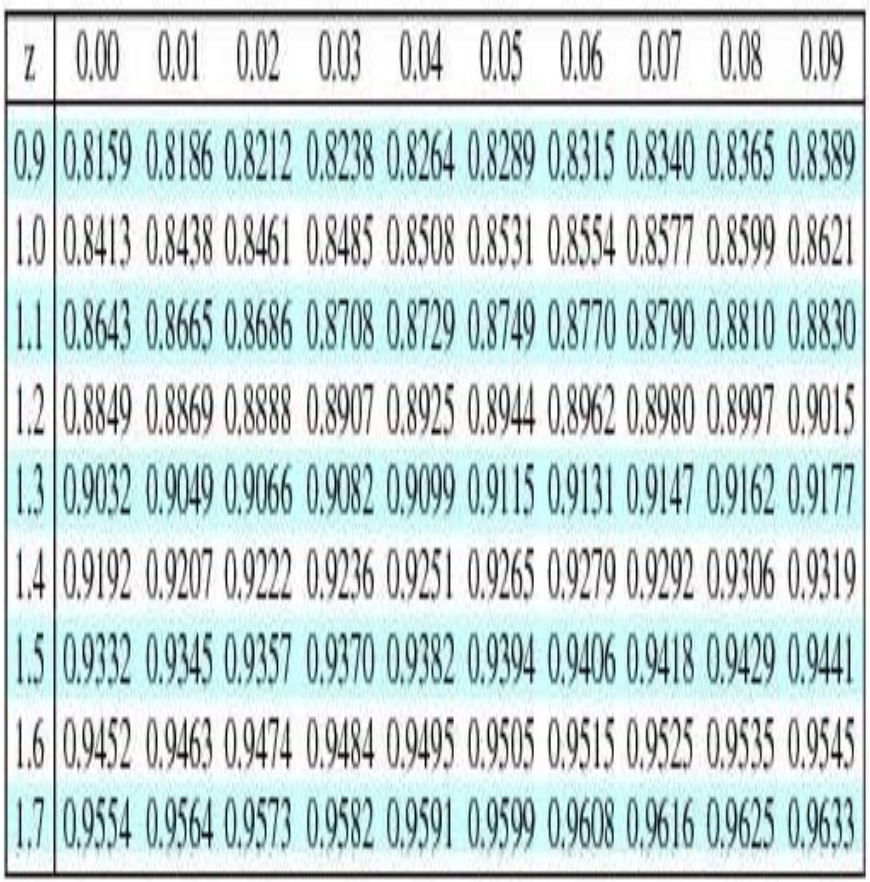
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133



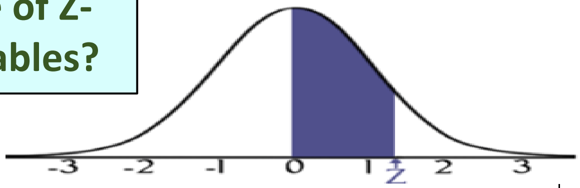
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133

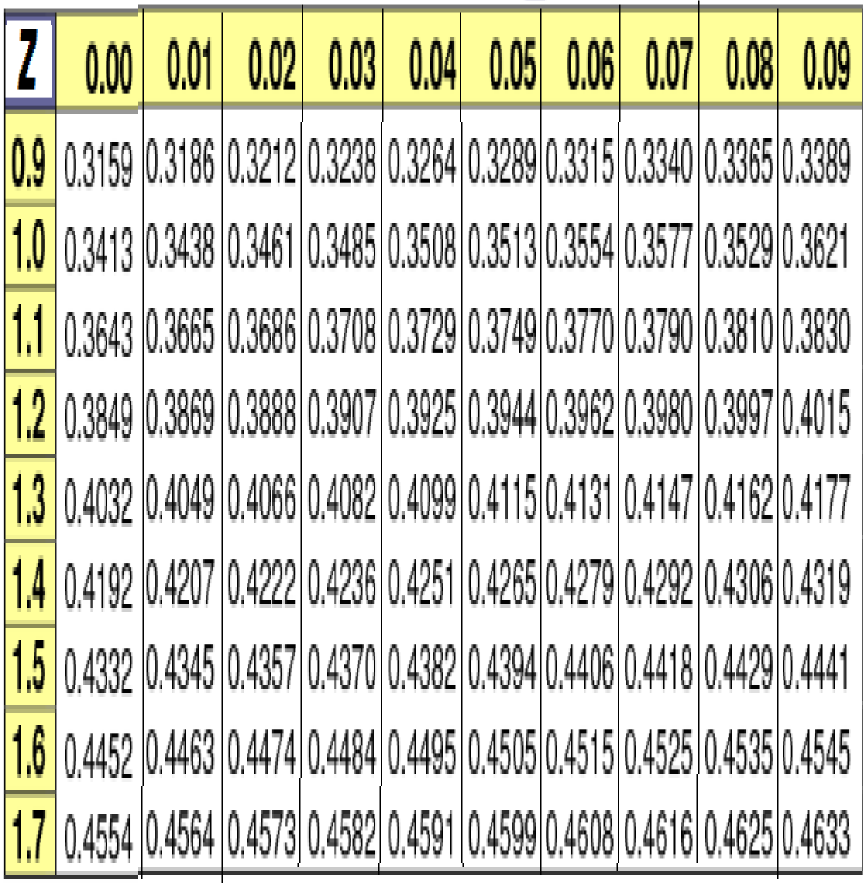
Application example. How to read a unique (same) value of Z-score using two different tables?





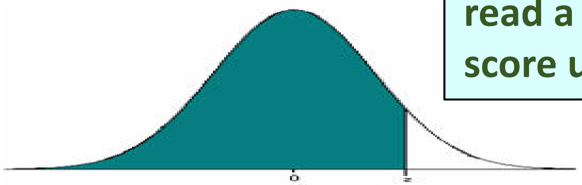
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

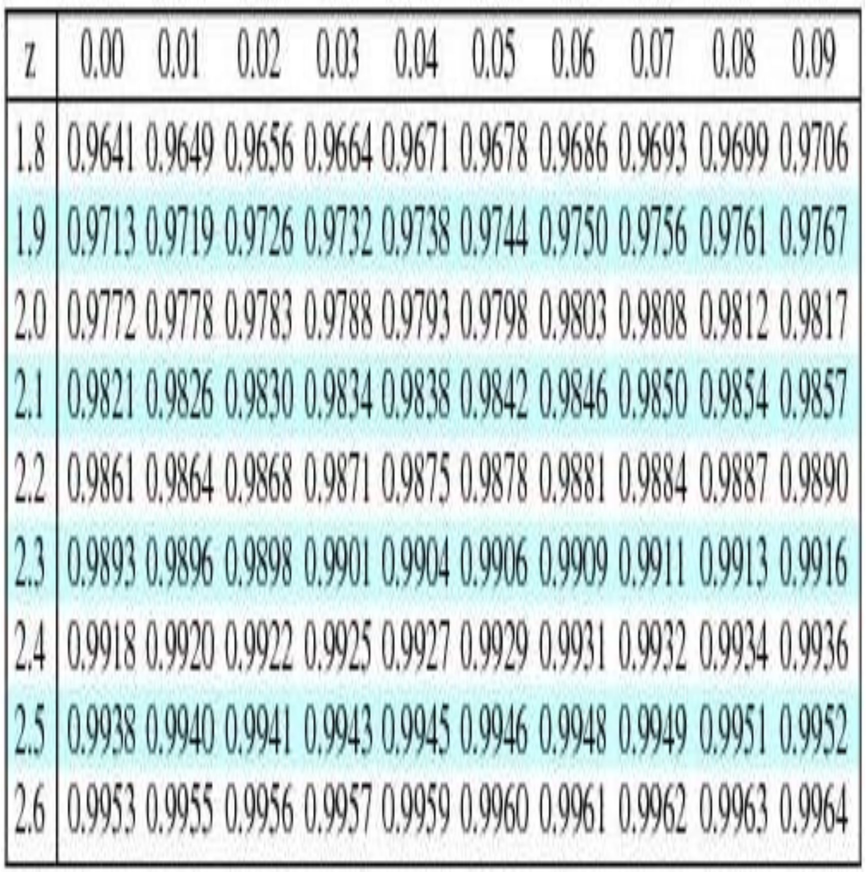




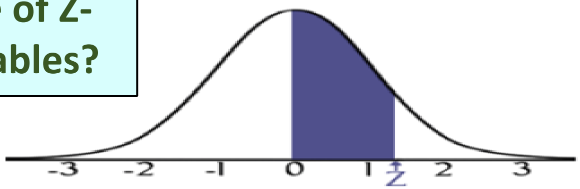
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633

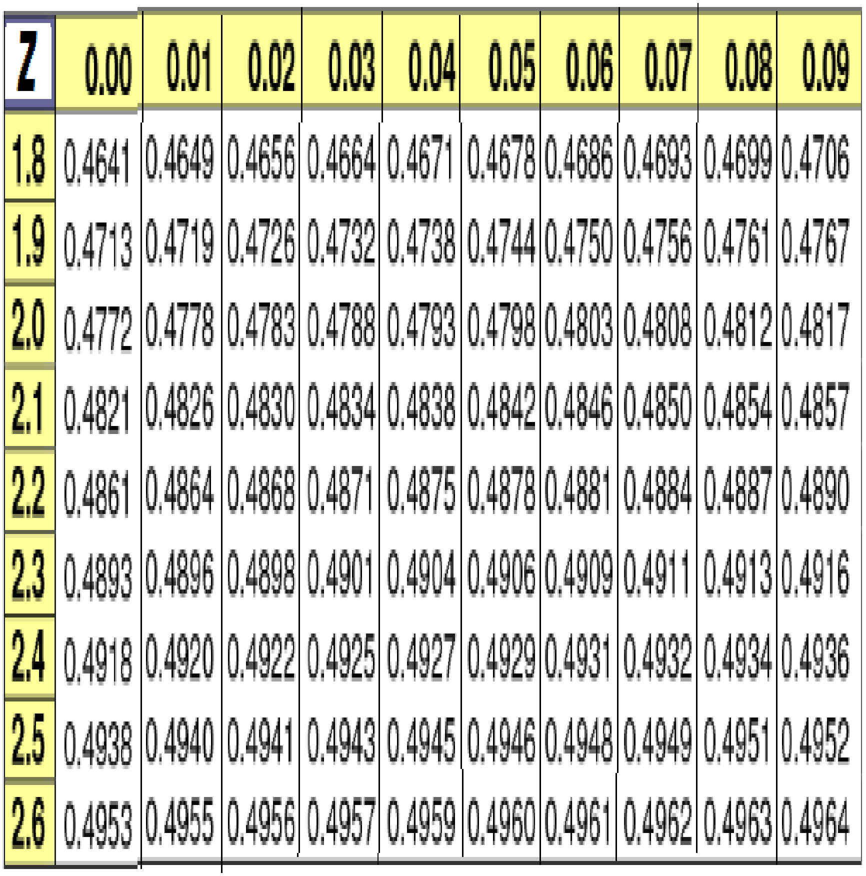
Application example. How to read a unique (same) value of Z-score using two different tables?





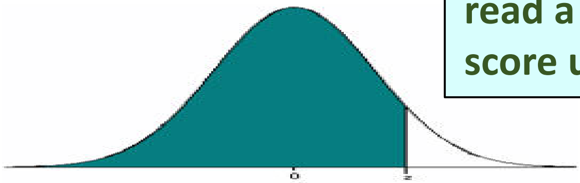
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964

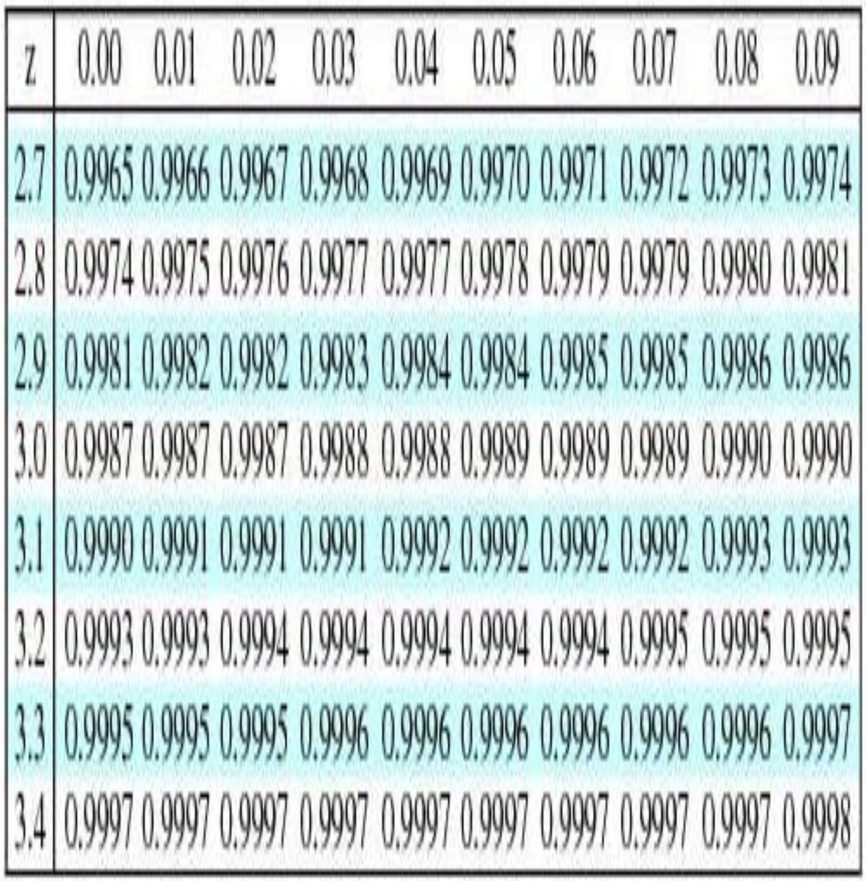




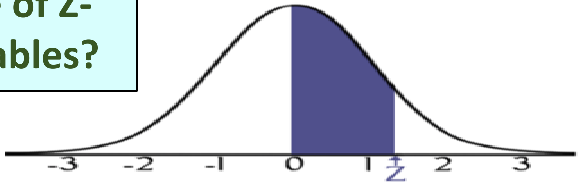
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964

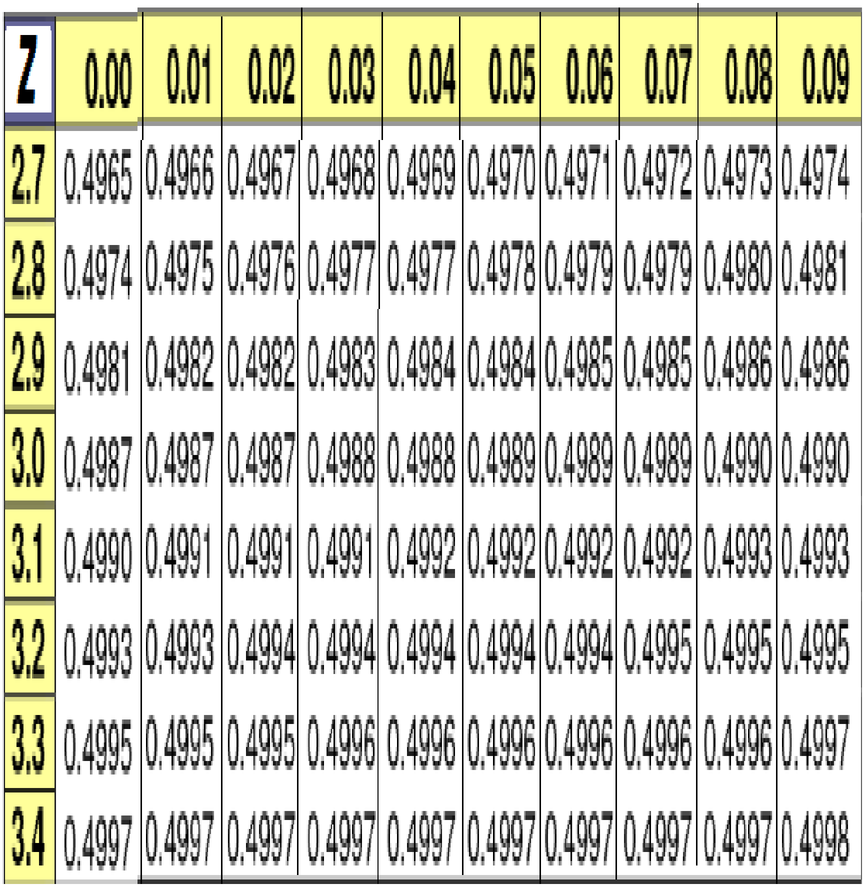
Application example. How to read a unique (same) value of Z-score using two different tables?





z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998





Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998

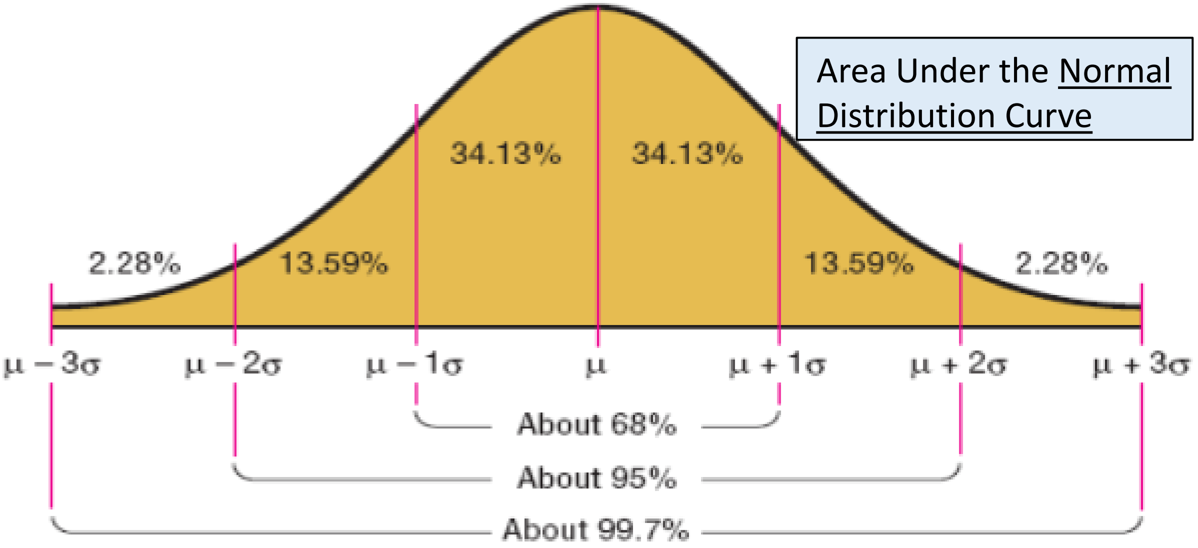
9. The central limit theorem (CLT)

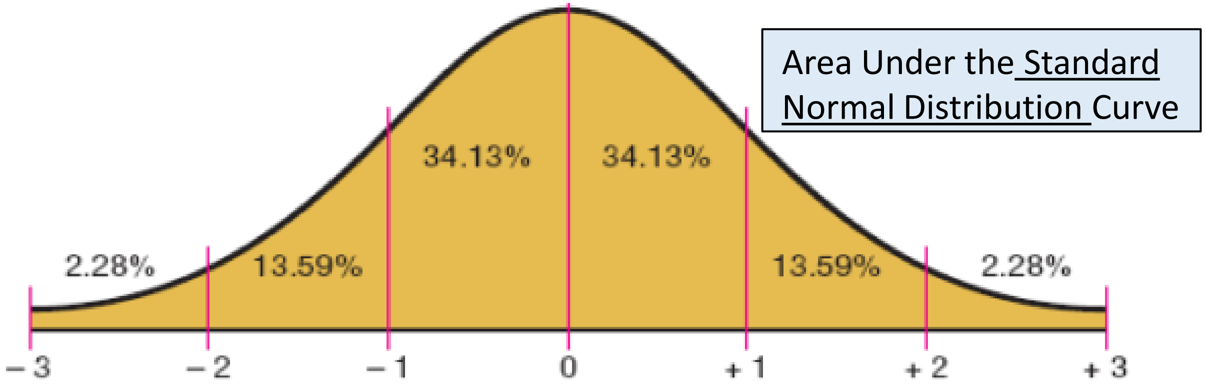
- □ Let $(X_1,............,X_n)$ be a simple random sample from a population with mean μ and variance σ^2.
- □ Let $\overline{X}=(X_1,............,X_n)/\,n$ be the sample mean.
- □ Let $S_n=(X_1,............,X_n)$ be the sum of the sample observations.
- □ Then if „n" is sufficiently large,

  $$\overline{X} \cong N\left(\mu,\ \frac{\sigma^2}{n}\right) \text{ and } S_n \cong N\left(n\mu,\sigma^2\right) \text{ approximately}$$

- □ For most populations, if the sample size is greater than "30", the central limit theorem approximation is good.
- □ Thus, the appropriate formula for the Z value is $Z=\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$

10. Normal distribution versus standard normal distribution (Z-score)

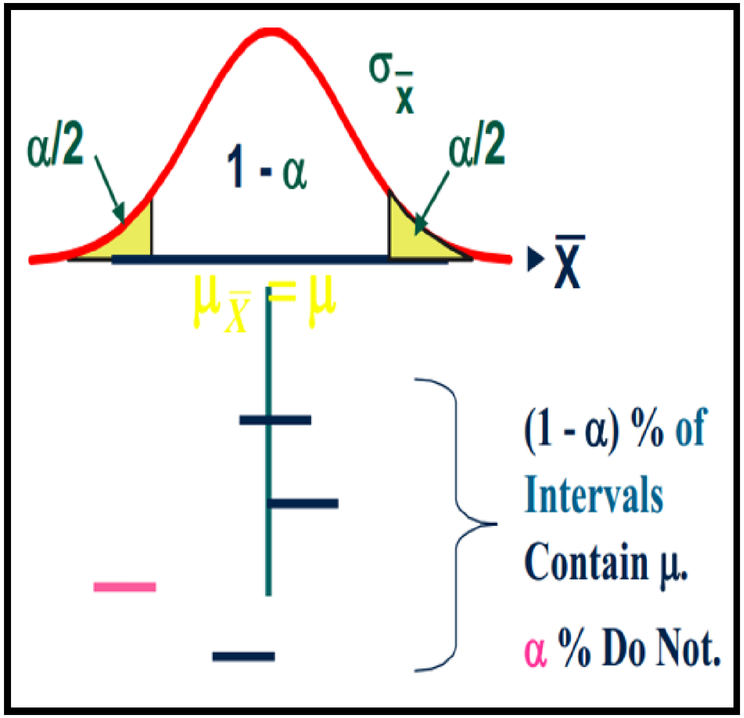

Area Under the Normal Distribution Curve


Area Under the Standard Normal Distribution Curve

11. Factors affecting the confidence interval (CI) range → (1)

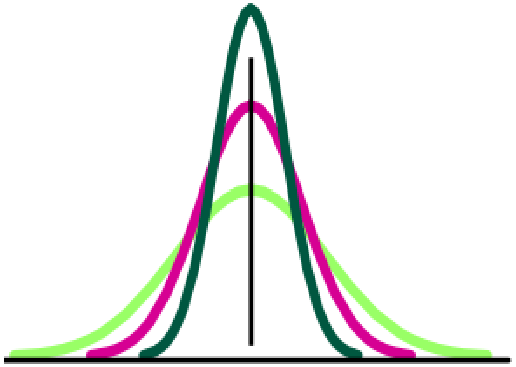
❑ **Level of confidence**

- ✓ Probability that the unknown population parameter is in the confidence interval in 100 trials.
- ✓ Denoted (1-α)% = level of confidence (e.g., 90%, 95%, 99%, etc.). The quantity **α** is the probability that the unknown population parameter is not within the interval **in any other** 100 of trials performed.

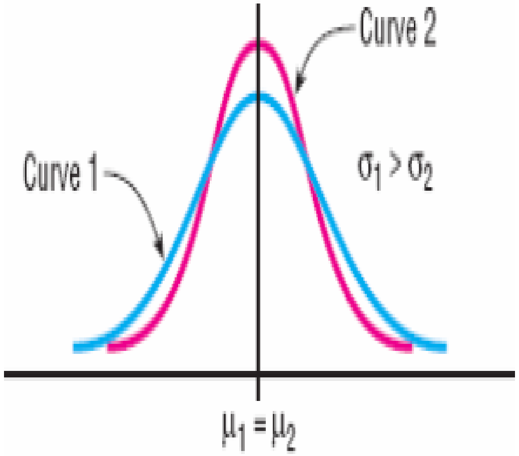


$$CI = \left[\bar{X} - Z \frac{\sigma}{\sqrt{n}} \qquad \bar{X} + Z \frac{\sigma}{\sqrt{n}} \right]$$

11. Factors affecting the confidence interval (CI) range → (2)



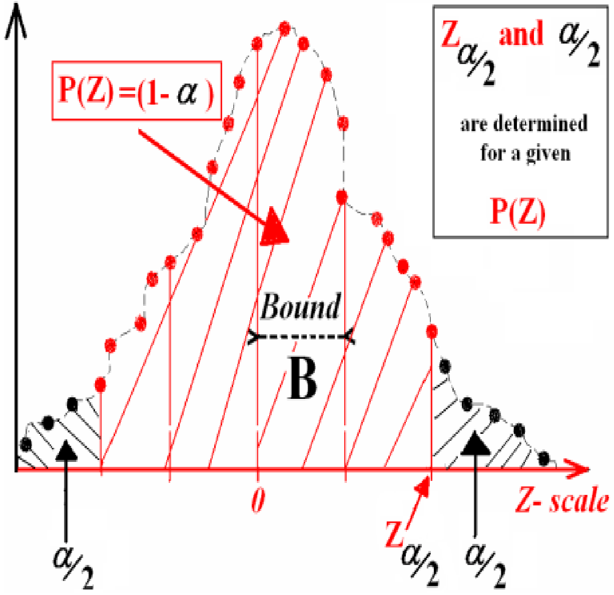
- ❑ **Data variation** σ_X
- ❑ **Sample size** $\sigma_{\overline{X}} = \sigma_X / \sqrt{n}$
- ❑ **Level of confidence** $(1-\alpha)$
- ❑ **Confidence interval** $\text{CI} = \left[\overline{X} - Z\frac{\sigma}{\sqrt{n}} \qquad \overline{X} + Z\frac{\sigma}{\sqrt{n}}\right]$
- ❑ **Sampling error** $Z\frac{\sigma}{\sqrt{n}}$



These Figures are with the same mean but different standard deviations

12. Confidence interval “estimates” under some assumptions

- ❑ Assume the population standard deviation is known
- ❑ Assume the population is normally distributed. If not, we must use a large sample data.
- ❑ Under these assumptions, the CI is estimated as follows:

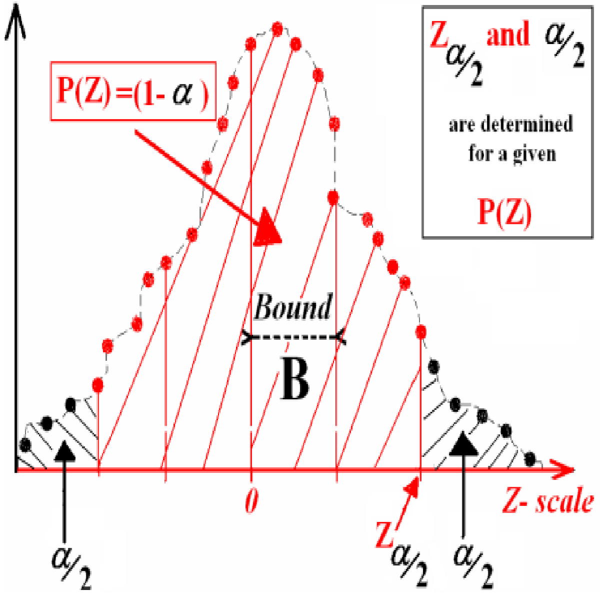


$$\text{CI} = \mu-\left(\frac{\sigma}{\sqrt{n}}\right)*Z_{\left(\frac{\alpha}{2}\right)} \leq \mu_{\text{expected}} \leq \mu+\left(\frac{\sigma}{\sqrt{n}}\right)*Z_{\frac{\alpha}{2}}$$

13. Z-score versus measurement scale

- ❑ **What is a Quantile ?**
  - ✓ A percentage expected for a given distribution.
    - o Example of quantiles:
      - 100*P(Z)=10%;
      - 100*P(Z)=50% (Median);
      - 100*P(Z)=68%;
      - 100*P(Z)=95%;

- ❑ **How to determine $Z_{\alpha/2}$?**
  - ✓ **Step 1:** Definition of the precision we want to achieve when analyzing data which average is denoted µ.
  - ✓ **Step 2:** Represent P(Z) in the plan Z-score.
  - ✓ **Step 3:** Use the value of P(Z) in the well-known table of Z-score to determine $Z_{\alpha/2}$



14. How to exploit confidence level- values in the table of z-score ?

Very important: Students must learn how to obtain all the values of Z-score mentioned in this table. **Please note that** three different Tables of Z-score must be used by students and it should be verified that the same/identical results are obtained regardless the table of Z-score used.

Quantiles: 100*P(Z)	$Z_{(\langle/ 2)}$
10%	0.13
30%	0.39
50%	0.675
68%	0.99
90%	1.645
95%	1.96
99%	2.575

$$\text{CI} = \mu - \left(\frac{\sigma}{\sqrt{n}}\right) * Z_{\left(\frac{\alpha}{2}\right)} \leq \mu_{\text{expected}} \leq \mu + \left(\frac{\sigma}{\sqrt{n}}\right) * Z_{\frac{\alpha}{2}}$$

Co-funded by the
Erasmus+ Programme
of the European Union

15. Application example

- ❑ Consider the exercise above (see section 4) and determine the confidence interval on the travel time between the two junctions/nodes (or the travel time of trains between two neighboring stations) for each of the following levels of confidence:
  - ○ 10%;
  - ○ 40%
  - ○ 50% (Median);
  - ○ 60%
  - ○ 68%;
  - ○ 80%
  - ○ 95%;
- ❑ Please comment the results obtained.

Mathematics in traffic and transport ...

16. The normal distribution

- ❑ What is the normal distribution?
  - ✓ The Gaussian distribution [Carl Friedrich Gauss (1777–1855)] defined as follows (where X is a continuous variable):

$$\boxed{X \cong N\left(\mu,\ \sigma^2\right)}$$

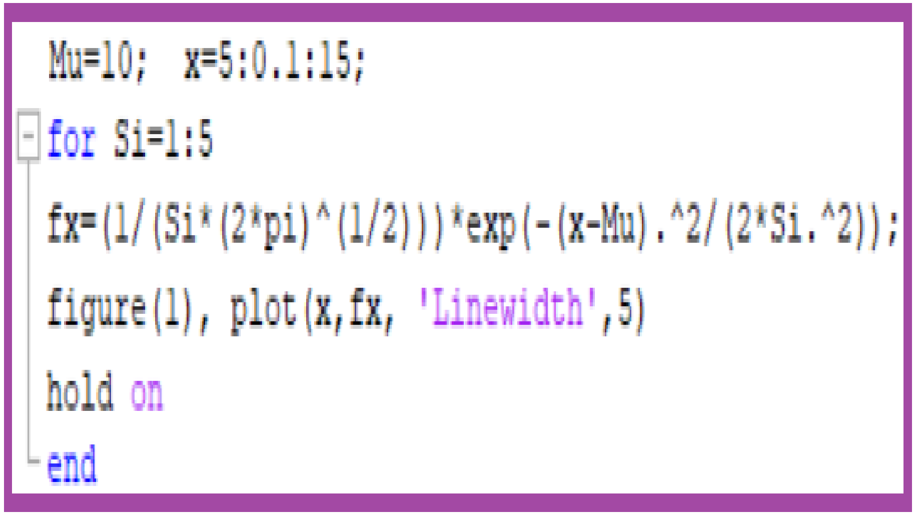
$$\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < +\infty}$$

17. MATLAB Script for the plot of the normal distribution

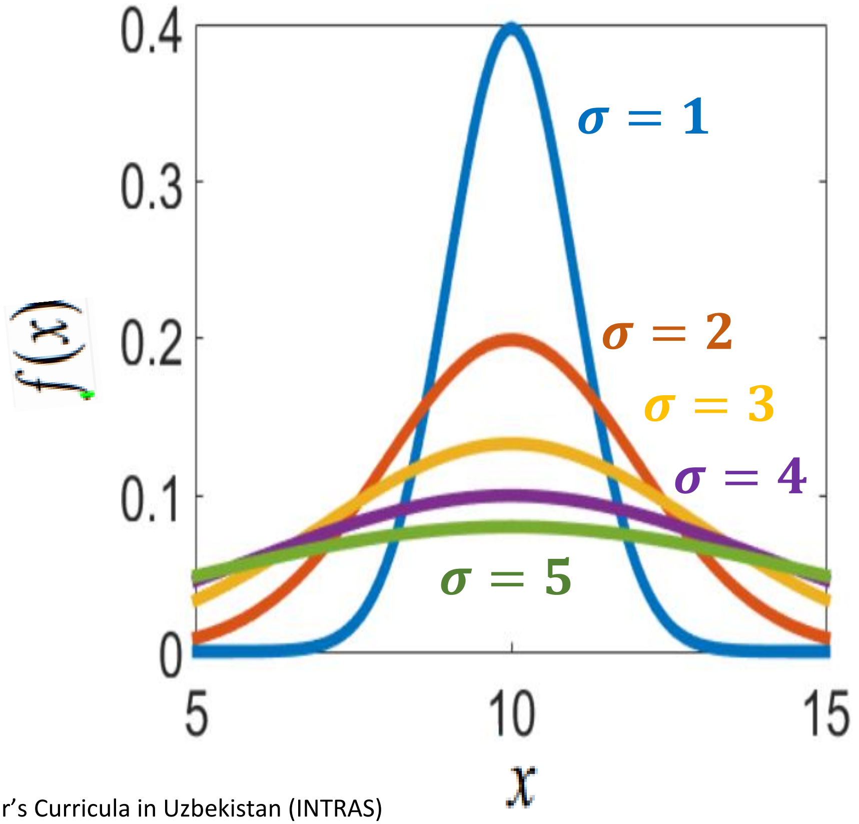
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty<x<+\infty$$

$$X \cong N(\mu,\ \sigma^2)$$

MATLAB Script for plotting *f(x)*



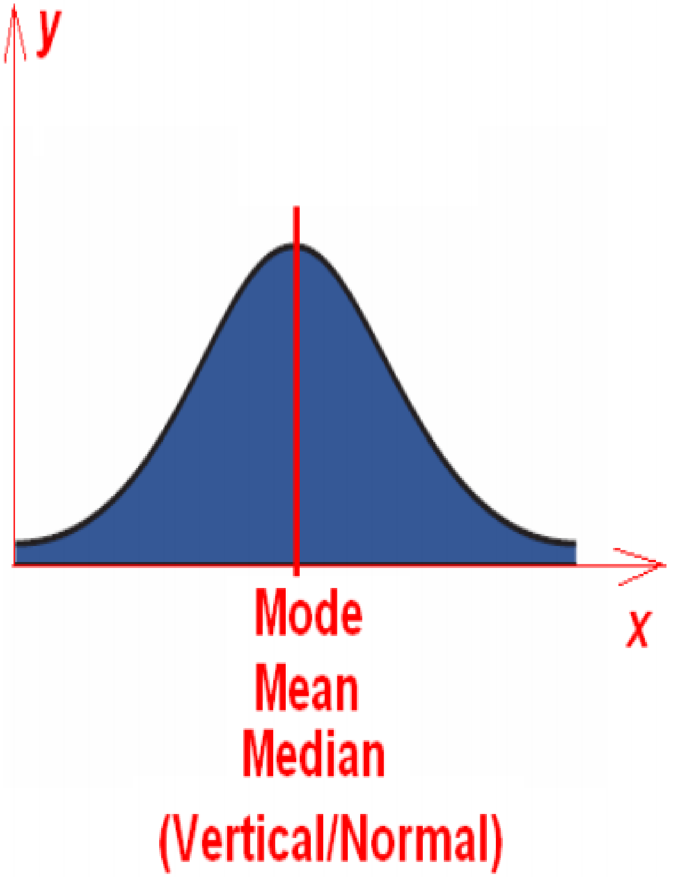
```
Mu=10;  x=5:0.1:15;
for Si=1:5
fx=(1/(Si*(2*pi)^(1/2)))*exp(-(x-Mu).^2/(2*Si.^2));
figure(1), plot(x,fx, 'Linewidth',5)
hold on
end
```



Co-funded by the
Erasmus+ Programme
of the European Union

18. The normal distribution: Localization of Mode, Mean, Median

- □ **Normal Probability of the** Gaussian distribution
  - ✓ The mean, median, and mode are equal and located at the center of the distribution.
  - ✓ The curve is unimodal (i.e., it has only one mode).
  - ✓ The curve is symmetric about the mean, (its shape is the same on both sides of a vertical line passing through the center.
  - ✓ The curve is continuous, (there are no gaps or holes). For each value of *X*, there is a corresponding value of Y.
  - ✓ The total area under the normal distribution curve is equal to 1.00, or 100%.



19. Likelihood: Difference between Likelihood and Probability.

❑ **What is Likelihood? and what is Probability?**

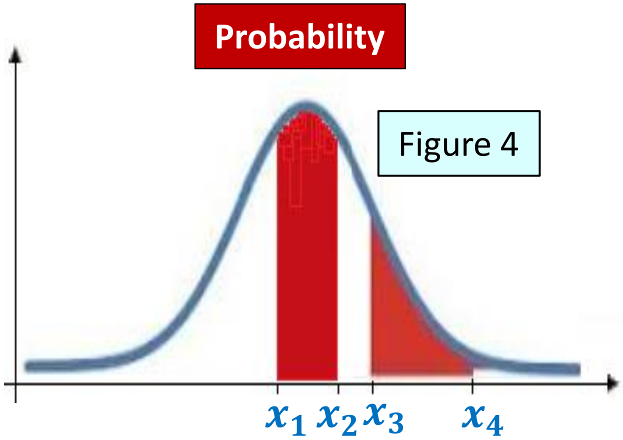
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty$$

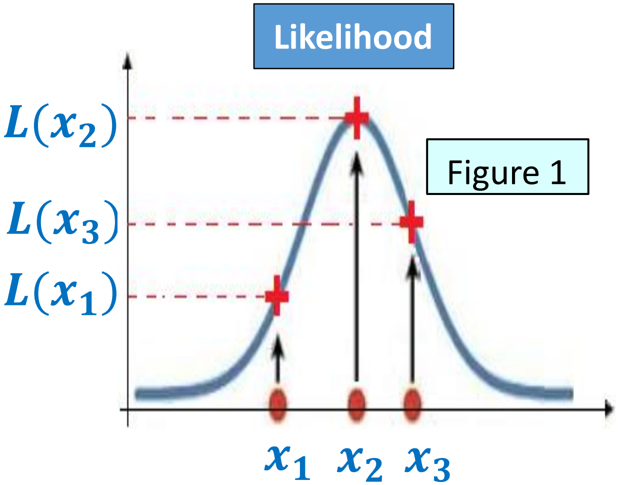
$$X \cong N(\mu, \sigma^2)$$

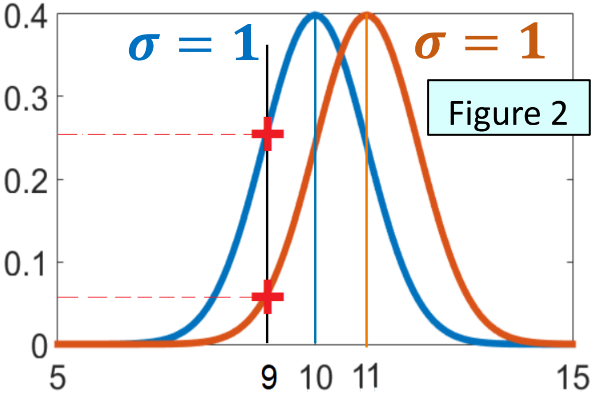
❑ **Likelihood depends on the following parameters?**

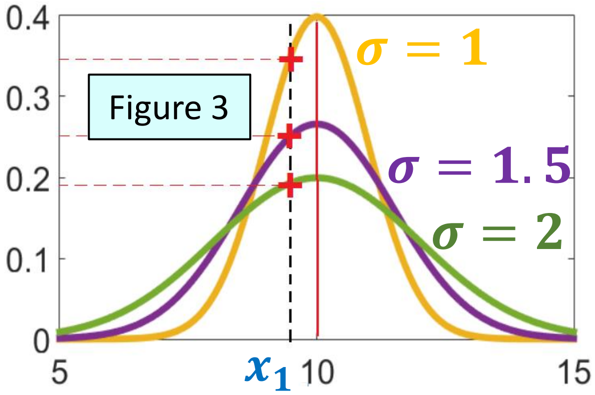
✓ The Mean

✓ The standard deviation









20. Likelihood of a single state: Extension to multiple states

- ❑ **Likelihood for a single sample dataset:**

$$L(\mu,\sigma,x_i)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

- ❑ **Likelihood for two samples dataset:**

$$L(\mu,\sigma,x_1\,,x_2)=L(x_1).L(x_2)$$

- ❑ **Likelihood for multiple sample dataset:**

$$L(\mu,\sigma,x_1\,,x_2\,,\dots,x_n\,)=\prod_{i=1}^{n}L(x_i)$$

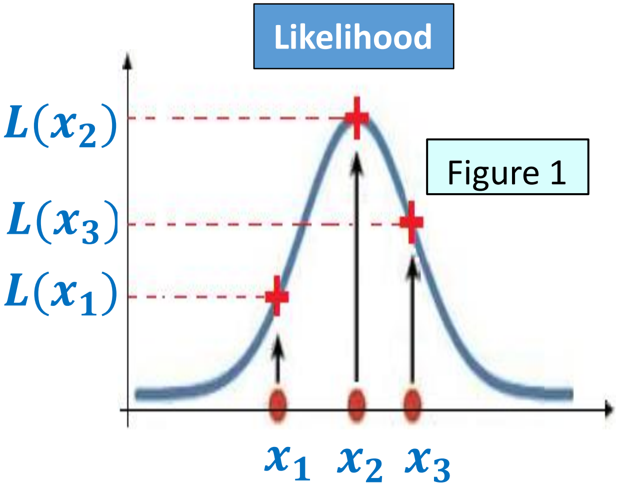


Figure 1

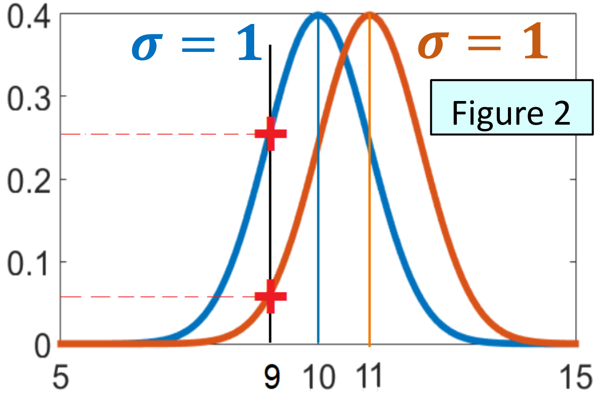


Figure 2

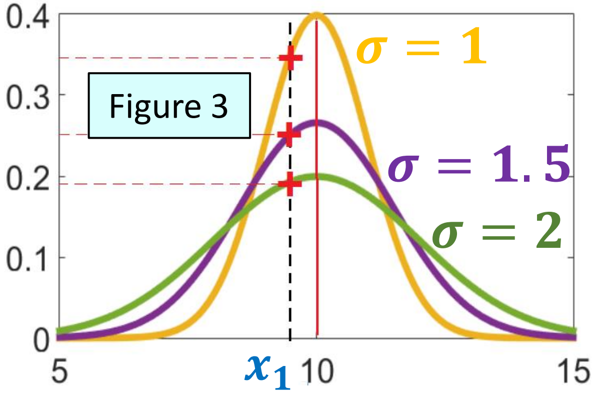


Figure 3

21. Maximum Likelihood of a normal distribution → (1)

$$L(set) = \prod_{i=1}^{n} L(x_i) = \underbrace{\frac{e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}}_{\boxed{L(x_1)}} * \underbrace{\frac{e^{\frac{-(x_2-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}}_{\boxed{L(x_1)}} *, \dots * \underbrace{\frac{e^{\frac{-(x_n-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}}_{\boxed{L(x_1)}} = \frac{e^{\frac{-(x_1-\mu)^2-(x_2-\mu)^2,\dots,-(x_n-\mu)^2}{2\sigma^2}}}{\left(\sigma\sqrt{2\pi}\right)^n}$$

- □ **Maximum Likelihood estimate for μ:** The partial derivative of $L(set)$ in the μ dimension

$$\frac{\partial L(set)}{\partial \mu} = \left(\frac{x_1 + x_2 + \dots + x_n - n\mu}{\sigma^2}\right) * L(set) \qquad (1)$$

- □ **Maximum Likelihood estimate for μ:** Using (1) the maximum Likelihood corresponds to:

$$\mu_{opt} = \left(\frac{x_1 + x_2 + \dots + x_n}{n}\right) = Mean \qquad (2)$$

- □ **Maximum Likelihood estimate for μ:** Eq. (2) shows that for a normal distribution (Gaussian distribution), the maximum Likelihood estimate for μ is the Mean of the measurements.

21. Maximum Likelihood of a normal distribution → (2)

$$L(set) = \prod_{i=1}^{n} L(x_i) = \frac{e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} * \frac{e^{\frac{-(x_2-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} *, \dots * \frac{e^{\frac{-(x_n-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} = \frac{e^{\frac{-(x_1-\mu)^2-(x_2-\mu)^2,\dots,-(x_n-\mu)^2}{2\sigma^2}}}{\left(\sigma\sqrt{2\pi}\right)^n}$$

$\boxed{L(x_1)}$ $\boxed{L(x_1)}$ $\boxed{L(x_1)}$

□ **Maximum Likelihood estimate for σ:** The partial derivative of $L(set)$ in the σ dimension

$$\frac{\partial L(set)}{\partial \sigma} = \left(\frac{(x_1-\mu)^2 + (x_2-\mu)^2, \dots, +(x_n-\mu)^2}{\sigma^3} - \frac{n}{\sigma}\right) * L(set) \qquad (3)$$

□ **Maximum Likelihood estimate for σ:** Using (3) the maximum Likelihood corresponds to:

$$\sigma_{opt} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}} = Std = Standard\ deviation \qquad (4)$$

□ **Maximum Likelihood estimate for σ:** Eq. (4) shows that for a normal distribution (Gaussian distribution), the maximum Likelihood estimate for σ is the ***Std.*** of the measurements.

Selected application examples to illustrate the theoretical results obtained in this chapter

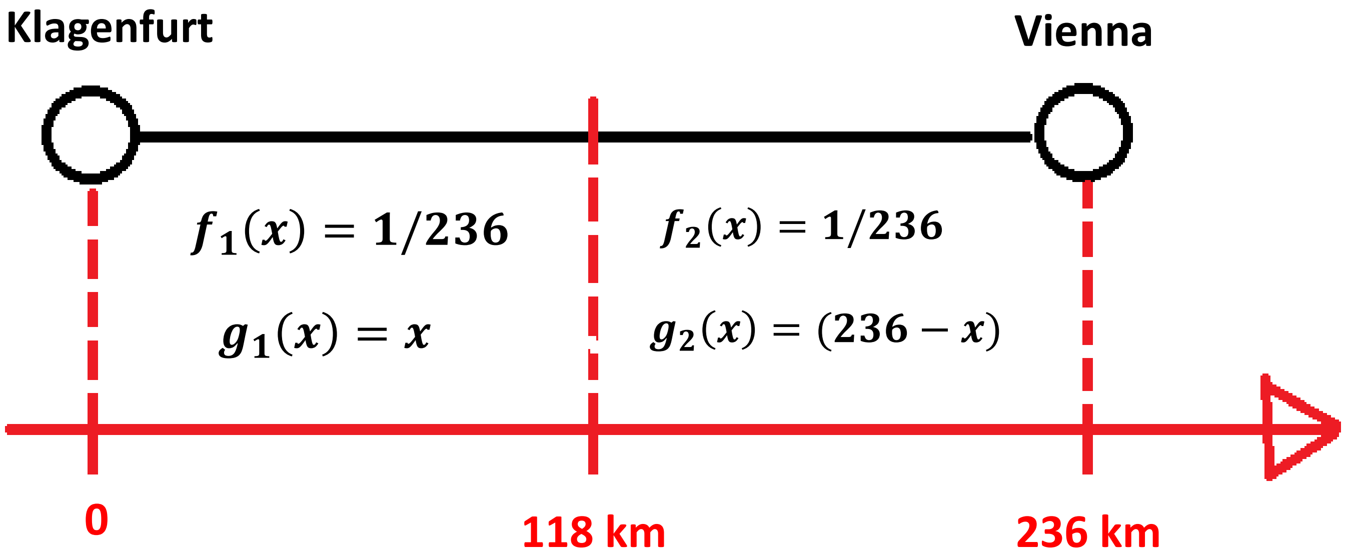
22. Examples in Railway Traffic → (1)

- ❑ **Example 1 (a concrete example):** The train distance between Klagenfurt (Austria) and Vienna (Austria) is about 236km. Motor cars often break down on this route. It is assumed that this failure can occur uniformly over the entire route. When a train breaks down, a spare locomotive must be brought in. It is assumed that ÖBB has two spare locomotives for this line. Of course, the nearest locomotive always goes to the point of breakdown.
  1. It is assumed that the locomotives are arranged respectively in the departure and arrival stations. What is the average distance traveled by the spare locomotive?
  2. An ÖBB engineer thinks that: it would be smarter to place the locomotives respectively 1/3 and 2/3 of the way. What do you think?
  3. Determine the optimal position of the two emergency locomotives to minimize the distance traveled by the spare locomotive.

22. Examples in Railway Traffic → (2)

❑ **Solution to Example 1:**

1. Calculation of the average distance traveled by the spare locomotive.

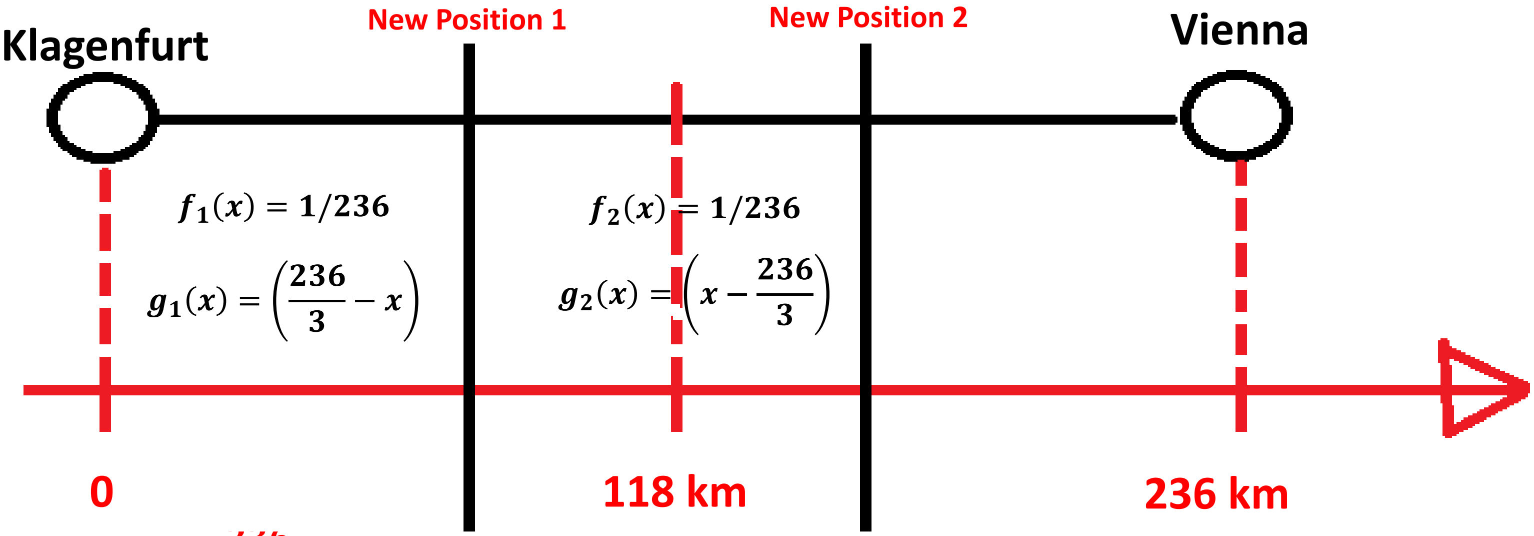


$$\mathrm{E}[g_1(x)] = \int_{0}^{118} f_1(x) * g_1(x)dx + \int_{118}^{236} f_2(x) * g_2(x)dx = 59\ km$$

22. Examples in Railway Traffic → (3)

❑ **Solution to Example 1:**

2. Practical thinking about where to place the Locomotives (for a good Engineer).



$$\mathrm{E}[g_1(x)] = \int_0^{\frac{236}{3}} f_1(x) * g_1(x)dx + \int_{\frac{236}{3}}^{118} f_2(x) * g_2(x)dx = 16.39\ km$$

Average distance traveled by the spare locomotive $= 2 * \mathrm{E}[g_1(x)] = 32.78km$

22. Examples in Railway Traffic → (4)

- **Example 2 (a concrete example):** A railway manager records the minutes late (or early) of intercity trains arriving at his station. He takes a random sample of 20 arrivals and finds that sample Mean and variance are 6.8 and 120 respectively.
  1. Construct a 99 per cent confidence interval estimator for mean minutes late of all intercity trains arriving at this station.
  2. What is the probable sampling error?
  3. What size of sample would be required to reduce the sampling error to five minutes?
- **Example 2 (Solution):**

1. $\boldsymbol{n = 20}$; $\boldsymbol{\mu = 6.8}$; $\boldsymbol{\sigma = \sqrt{120} = 10.9545}$;

   $\boldsymbol{Z - score(99\%) = 2.575}$.

   The confidence interval is: $\boldsymbol{CI = [0.5mn\ ,\ 13.1mns]}$

2. The sampling error is: $Zscore * \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}} = \boldsymbol{6.3mns}$

3. New size of sample corresponding to the Sampling Error of 5mns: Solving the equation << $Zscore * \frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}} = \boldsymbol{5mns}$ >> leads to << $\boldsymbol{n = 32\ arrivals}$ >>.

CONCLUDING REMARKS

- □ We have presented and explained the full content of the Lecture entitled “Advanced Statistics and Data Analysis”

- □ The importance of this Lecture for the „New ICT-based Master‘s Curriculum in Uzbekistan“ has been clearly demonstrated.

- □ This Lecture is also part of the Key Module of the Double-degree Master- program between the „University of Klagenfurt“ and „Some partners Universities in Uzbekistan“

- □ The Lecture entitled “Advanced Statistics and Data Analysis” is of huge potential applications in Road Transportation, Railway Transportation as well as in Supply Chain Networks and Logistics.

END

Q!