BTU ბიზნესისა და ტექნოლოგიების უნივერსიტეტი

python

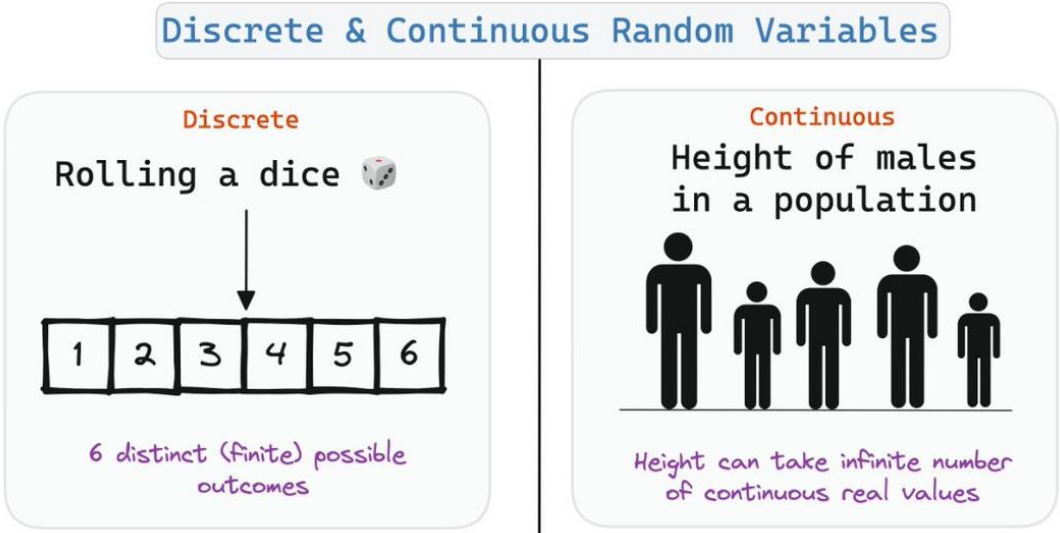
მონაცემთა ანალიტიკა Python

ლექცია 4:შესავალი ალბათობაში. შემთხვევითი სიდიდე, ალბათური განაწილება დისკრეტული და უწყვეტი ცვლადებისათვის. დიდ რიცხვთა კანონი და ცენტრალური ზღვარითი თეორემა.

ლიკა სვანაძე
lika.svanadze@btu.edu.ge

შემთხვევითი სიდიდე

- **შემთხვევითი სიდიდე** არის მათემატიკური ფუნქცია, რომელიც ანიჭებს რიცხვით მნიშვნელობას შემთხვევით ხდომილებას (თითოეულ შესაძლო შედეგს).
- შემთხვევით სიდიდეს ეწოდება **დისკრეტული** ტიპის თუ ის ღებულობს ცალკეულ, იზოლირებულ შესაძლო მნიშვნელობებს.
- შემთხვევით სიდიდეს ეწოდება **უწყვეტი ტიპის** თუ მისი შესაძლო მნიშვნელობების სიმრავლე მთლიანად ავსებს რაიმე რიცხვით შუალედს.



ალბათური განაწილება

ცხრილს, რომელშიც ჩამოთვლილია დისკრეტული ტიპის შემთხვევითი სიდიდის შესაძლო მნიშვნელობები და მათი შესაბამისი ალბათობები, **ალბათური განაწილების კანონი** ეწოდება

Colour	Red	Blue	Green	Yellow
Probability	0.3	0.35	0.15	0.2

უნდა აკმაყოფილებდეს შემდეგ ორ პირობას:

1. თითოეული ალბათობა უნდა იყოს 0-სა და 1-ს შორის

$$0 \leq P(x) \leq 1.$$

2. ჯამი ყველა შესალო შემთხვეივითი სიდიდის უნდა იყოს 1-ის ტოლი

$$\sum P(x) = 1.$$

Probability

Probability: An Alternative Method of Expressing Uncertainty

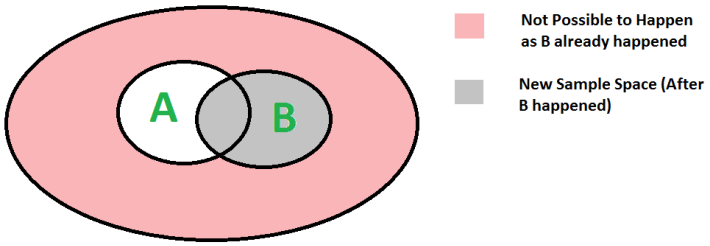
- **Probability** is represented numerically by a number between 0 and 1.
- Statements with a probability of 0 are **false**. Statements with a probability of 1 are **true**. An event that is **certain** to occur has a probability of 1; An event that is **certain not** to occur has a probability of 0.
- Probability of 0.5 or 50% are just as likely to be true as false.
- The probability of event A is written p[A].
- The sum of the probabilities of all possible, collectively exhaustive outcomes of a chance event must be equal to 1. e.g., p[heads]+ p[tails] = 1.0.
- Events A and B are considered **independent** if the occurrence of one does not influence the probability of the occurrence of the other. The probability of two independent events A and B both occurring is given by the product of the individual probabilities:

$$p[A,B] = p[A] \times p[B]$$

Head Tail

Conditional Probability

- The probability that event A will occur given that event B is known to occur is called the **conditional probability**
- conditional probability of event A given event B is denoted by p[A|B] and read as “the probability of A given B.”
- Thus a post-test probability is a conditional probability predicated on the test or finding. For example, if 30 % of patients who have a swollen leg have a blood clot, we say the probability of a blood clot given a swollen leg is 0.3, denoted: p[blood clot | swollen leg] = 0.3.
  - $p[A \cap B] = p[A] * p[B|A]$ (Bayes’ theorem)



Probability

Additive Rule

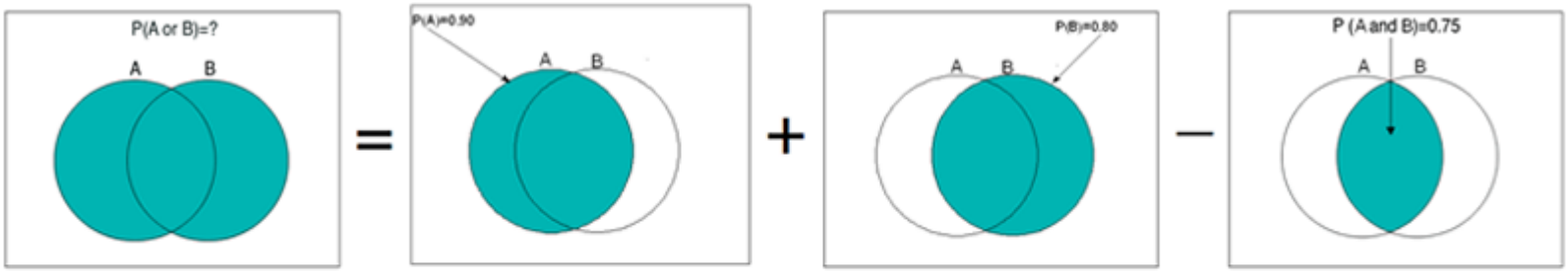
1. Mutually Exclusive
$P(A \cup B) = P(A)+P(B)$

2. Not Mutually Exclusive
$P(A \cup B) = P(A)+P(B)-P(A \cap B)$

Multiplication Rule

1. Independent Events
$P(A \cap B) = P(A)*P(B)$

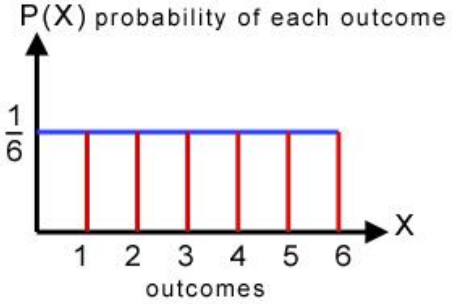
2. Dependent Events
$P(A \cap B) = P(A)*P(B|A)$



We therefore get:

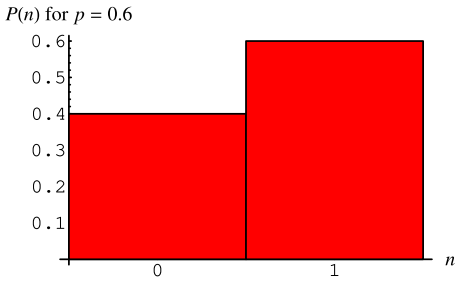
$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

ალბათური განაწილება დისკრეტული ცვლადებისთვის



1. **Discrete Uniform Distribution** - All possible outcomes are equally likely. Example: Rolling a fair six-sided die.

2. **Bernoulli Distribution** - Describes a single binary outcome (success/failure) with a probability of success "p". Example: Coin toss (Heads/Tails).



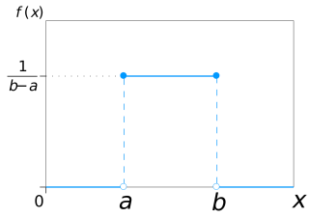
3. **Binomial Distribution** - Models the number of successes in a fixed number of independent Bernoulli trials. Parameters: Number of trials "n" and probability of success "p". Example: Number of heads in 10 coin tosses.

4. **Poisson Distribution** (პუასონის): Describes the number of events occurring in a fixed interval of time or space, given a known average rate of occurrence. Parameter: Average rate "λ". Example: Poisson Distribution for Phone Calls with an Average of 5 Calls per Hour

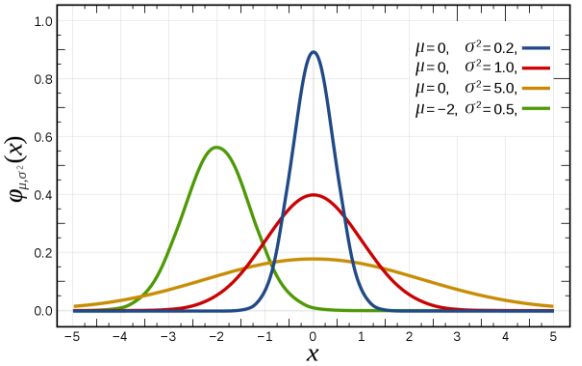
$$P(X = x) = \frac{e^{-\lambda} \lambda^{x}}{x!}$$

ალბათური განაწილება უწყვეტი ცვლადებისთვის

1. Continuous Uniform Distribution - All values within a specified interval are equally likely. Example: Selecting a random real number between 0 and 1.

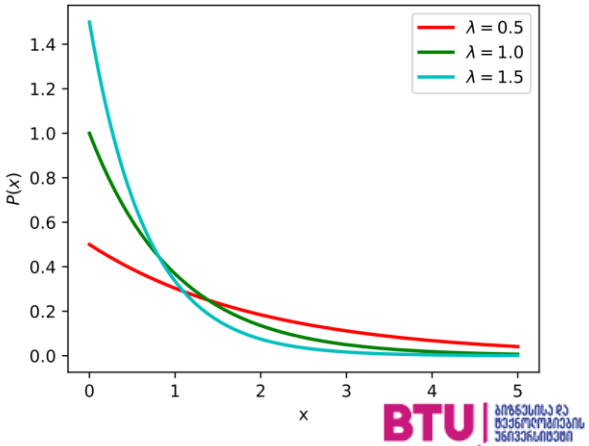


2. Normal (Gaussian) Distribution - Describes a symmetric, bell-shaped curve characterized by its mean and standard deviation. Many natural phenomena tend to follow this distribution due to the **Central Limit Theorem**. Example: Height, weight, IQ scores in a large population.



3. Exponential Distribution: - Models the time until an event occurs in a Poisson process (events occurring continuously and independently at a constant average rate). Parameter: Rate parameter "λ" (average number of events per unit time). Example: Time until a radioactive atom decays.

$$f(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases}$$



ჭეშმარიტი საშუალო / true mean

The **true mean** refers to the actual, exact average or expected value of a population. This parameter represents the central tendency of the entire population under consideration.

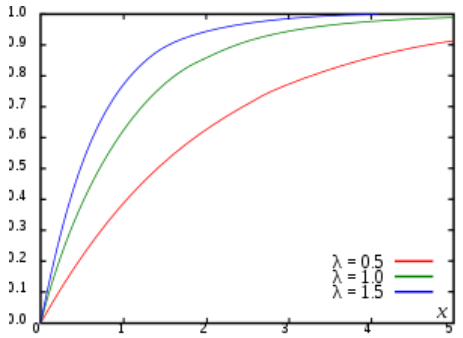
For example, if you're studying the heights of all adult males in a particular country, the true mean height would be the average height of every single adult male in that country.

In practice, it's often impossible or impractical to measure every individual in a population, so we rely on sample data to estimate the true mean. The **sample mean** is the average of a subset of individuals from the population.

Cumulative distribution function (cdf)

CDF gives the probability that the random variable X is less than or equal to x and is usually denoted F(x).

$$F_X(x) = P(X \leq x)$$



In Python's **scipy.stats.norm** module, the function **cdf(x, loc=0, scale=1)** computes the CDF of a normal (Gaussian) distribution.

x: This is the value at which you want to evaluate the cumulative distribution function. It represents a point on the x-axis of the normal distribution.

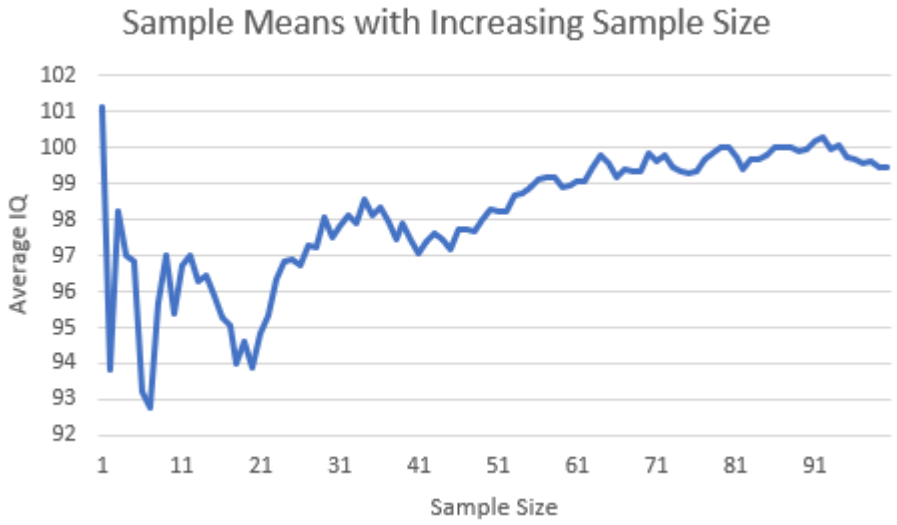
loc: This parameter specifies the mean of the normal distribution. It indicates the center of the distribution. By default, it is set to 0.

scale: This parameter represents the standard deviation of the normal distribution. It determines the spread or dispersion of the distribution. By default, it is set to 1.

The function **norm.cdf(x, loc, scale)** calculates the probability that a random variable from a normal distribution with mean loc and standard deviation scale is less than or equal to x.

დიდ რიცხვთა კანონი

The Law of Large Numbers and the Central Limit Theorem play important roles in statistics because they allow us to make inferences about the true mean based on sample data. They give us confidence that, as sample sizes increase, the sample mean will converge towards the true mean.



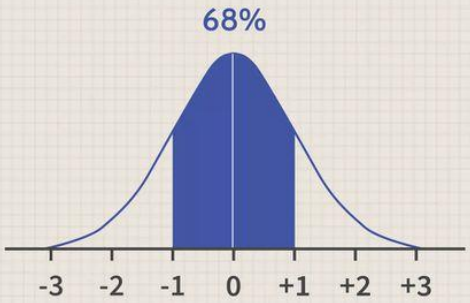
ცენტრალური ზღვარითი თეორემა

დიდ რიცხვთა კანონი არ იკვლევს შემთხვევით სიდიდეთა ჯამის განაწილების კანონის სახეს. ეს საკითხი შეისწავლება თეორემების ჯგუფში, რომლებსაც ცენტრალური ზღვარითი თეორემა ეწოდება.

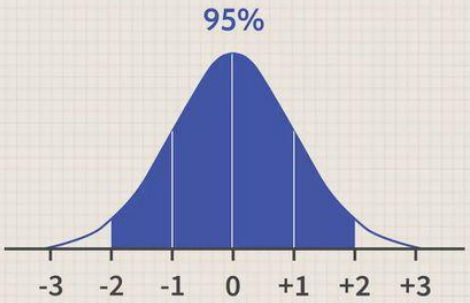
ეს თეორემა ამტკიცებს, რომ შემთხვევით სიდიდეთა ჯამის განაწილების კანონი, რომელთაგან ცალკეულ შესაკრებს შეიძლება ჰქონდეს განსხვავებული განაწილება, უახლოვდება **ნორმალურ განაწილებას** შესაკრებთა საკმაოდ დიდი რიცხვის შემთხვევაში.

ამით აიხსნება ნორმალური განაწილების კანონის უაღრესად დიდი მნიშვნელობა პრაქტიკულ გამოყენებებში.

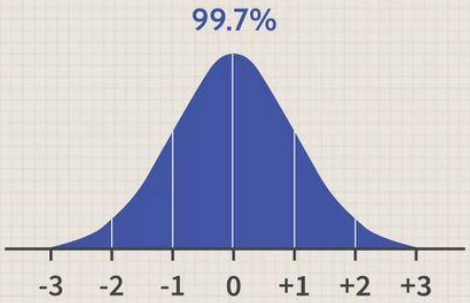
The Central Limit Theorem states that the distribution of the sum (or average) of a large number of independent, identically distributed random variables will be approximately normally distributed, regardless of the shape of the original distribution.



68% of all values are within 1 standard deviation of mean value



95% of all values are within 2 standard deviations of mean value



99% of all values are within 3 standard deviations of mean value