Key statistical concepts

Chuvakin Sergey

«School of Advanced Studies»

October 26, 2020

Outline

- Random variable
- Distribution
- Distribution types
- Central measures
- Dispersion
- Standart deviation
- Types of variables
- Population
- Sample
- law of large numbers (maybe)
- central limit theorem (maybe)

Random Variable

Definition

Variable - ***varying*** values.
RV - aka random quantity, aleatory variable, or stochastic variable - is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes.

Examples: - tips for waiter, number of people in a line, number of insects under a bed and a lot of other examples. The idea that it's unlimited number. Everything potetntilaly could be a random variable.

Distribution

Definition

In statistics, a probability distribution is a mathematical description of a random variable in terms of the probabilities of it's particular possible values.

Put it simpler - distribution - possible values of random variable.

NB: It's a function, therefore there are input and output.

Distribution

Types

Tap here! - interactive types of various distributions!

Let us see it together!
What worth noticing:

1. it can be descrite and continious
2. each has extra parameters
3. shape varies
4. some combinations of parameters of different distributions looks alike!

Distribution

Types

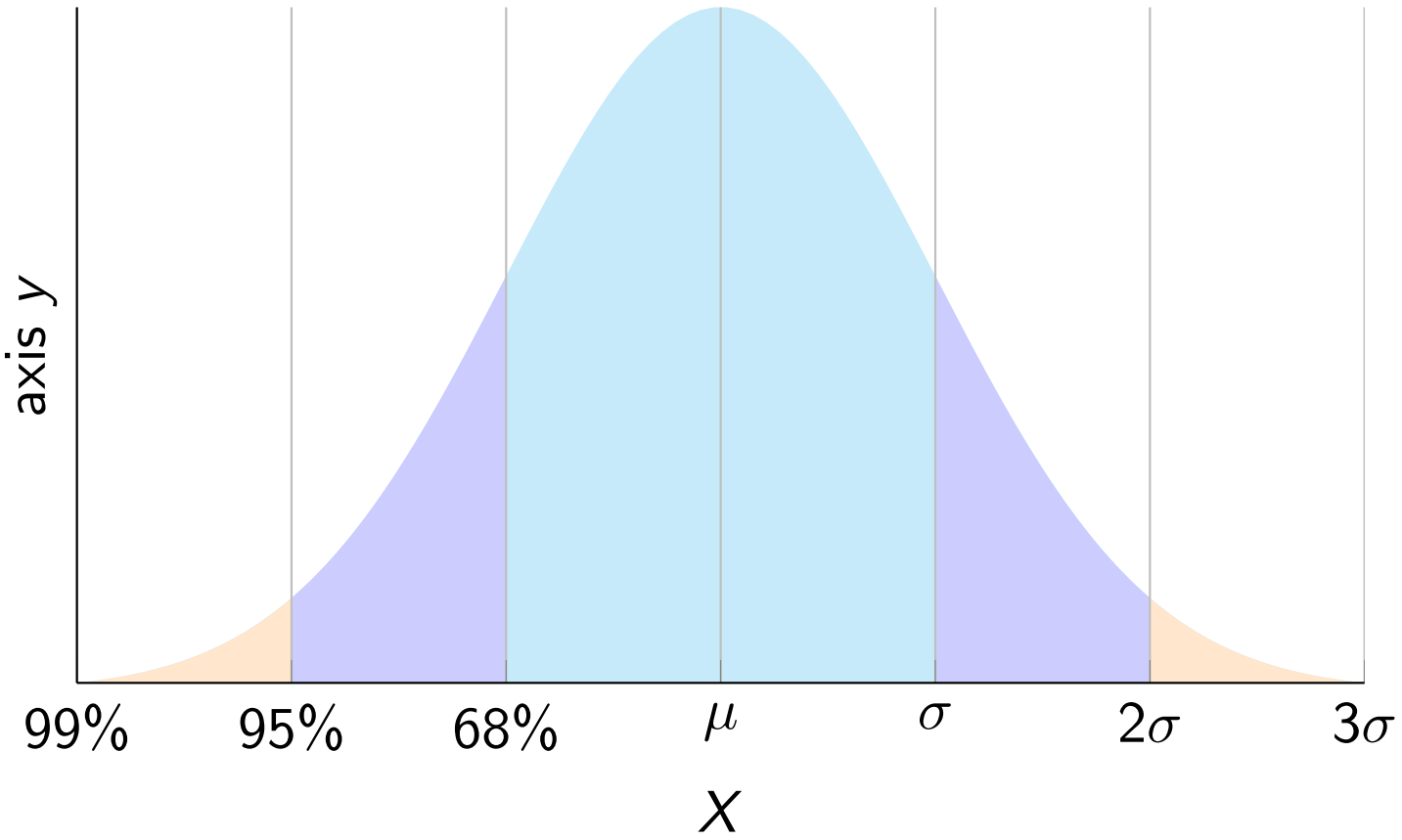
Exercise - try to guess the following!

1. meaning of (at least some of) parameters
2. difference between descrete and continious
3. what is probability mass (density) function (PDF) $f(x)$
4. what if cumulative distribution function (CDF) $F(x)$

Do not be upset if not all of above is clear! The more important is to grasp intuition...

Normal distribution

Plot



Normal distribution

Fomula

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- $P(x)$ aka y
- σ - standart deviation
- μ - mean aka Expected value $E(x)$
- e - e number (2.7~)
- π Pi number (3.14~)
- x - value from random variable

Intuition behind - probability of x assumed it distributed normally.

Back to variables

Examples in sociology and anthropology

RV - frequently used as quality (or feature) of person or some phenomena.
Examples of pseudo normal distribution:

- Age
- Heigh
- Salary
- Number of robberies in a country
- Number of votes during elections
- Number of cigarettes smoked

Formal Statistics

central tendency

- Mean
- Mode
- Median

⚠ In trully normal distrubution Mean $\simeq$ Mode $\simeq$ Median

Formal Statistics

central tendency

The mean of a distribution is the arithmetic mean, or the «average»

$$A = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

$$\mu(X) = \frac{\sum_{i=1}^{n} x^i}{n}$$

Formal Statistics

central tendency

The median is the value separating the higher half of sample,a population, or a probability distribution, from the lower half.

1. The median is the "middle" value of a [ordered] data set.
2. Let there be a variable v1: 20,7,23,17,21,5,19,3,11
3. To compute the median of v1, sort the variable into ascending order: 3,5,7,11,17,19,20,21,23
4. Pick one in center (17)

Formal Statistics

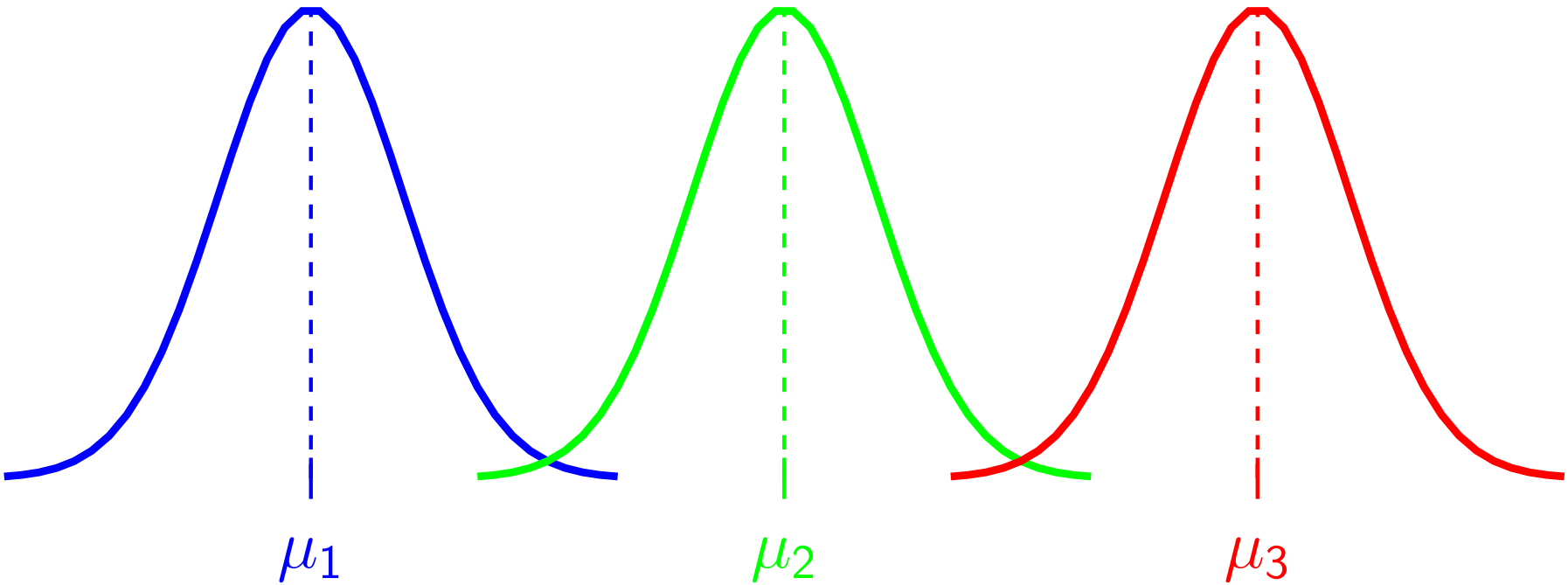
central tendency

The mode is the value that appears most often in a set of data

1. Consider data 2,5,7,6,7,9,2,0,5,3,3,7,7,8.
2. The mode is 7
3. What about 3,7,4,2,3,1,0,7,9,6,3,7,4,11?
4. The mode of a continuous probability distribution is the value x at which its probability density function has its maximum value
5. The mode is at the peak of the distribution

Formal Statistics

central tendency



Formal Statistics

central tendency

⚠ 1 Guess why median is more representative one?

⚠ 2 What is the synonym for representative?

Dispersion and std

Formula

Range is the difference between the smallest and the largest observation in the sample.

- Variance is the average of all squared deviations from the mean:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu(X))^2}{n}$$

- The larger this value, the greater the dispersion of the observations around the mean value, the more heterogeneous sample (the less informative mean).
- The standard deviation (denoted as σ or s) is the square root of the variance

Variable types

list

1. continious $(-\infty : +\infty)$
2. nominal (colors)
3. ordered (ranks in the army)

Variable types

Examples

- Continious:
  1. Age of person
  2. Salaries in a company
- Nominal:
  1. Religion of person
  2. Prefered candidate at elections
  3. Gender
- Ordered:
  1. Number of smoked sigaretes
  2. Number of children in a family

Variable types

Notions

- Continious - takes any real number - it can be normally distributed or somehow else.

- Nominal - always descrete. The only thing can be done upon this - count unique values. The values *can not be* compared!

- Ordered - unlike nominal scal it can be compared, but we do not know the granularity of difference. Descrete as well!

Population and Sample

Formula

Popuation - theoretical measure of your object of research.

Sample - Real number of observations

Representetivness - How sample reflects quality of population

Important Laws

Law of large numbers

The law of large numbers, in probability and statistics, states that as a sample size grows, its mean gets closer to the average of the whole population

Important Laws

Central Limit Theorem

The central limit theorem states that if you have a population with mean μ and standard deviation $\sigma$and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distribute