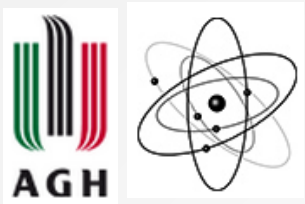
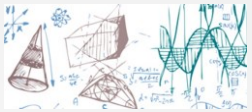


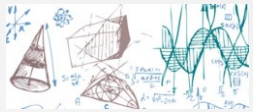


Introduction to probability, statistics and data handling

Tomasz Szumlak, Agnieszka Obłąkowska-Mucha
Faculty of Physics and Applied Computer Science

AGH UST Krakow

2021



2 Chebyshev's Inequality

- There is an extraordinary theorem related to the fundamental properties of RV (both discrete and continuous). We just need both the **expectation value and variance to be finite**.
- **Theorem 5**. Suppose that X is a random variable. Let the mean and variance of this RV be μ and σ^2 respectively. If we assume that they are both finite, then if ϵ is any positive number:

$$p(|X-\mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

$$\epsilon = k\sigma \rightarrow p(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

- For instance, let $k = 2$:

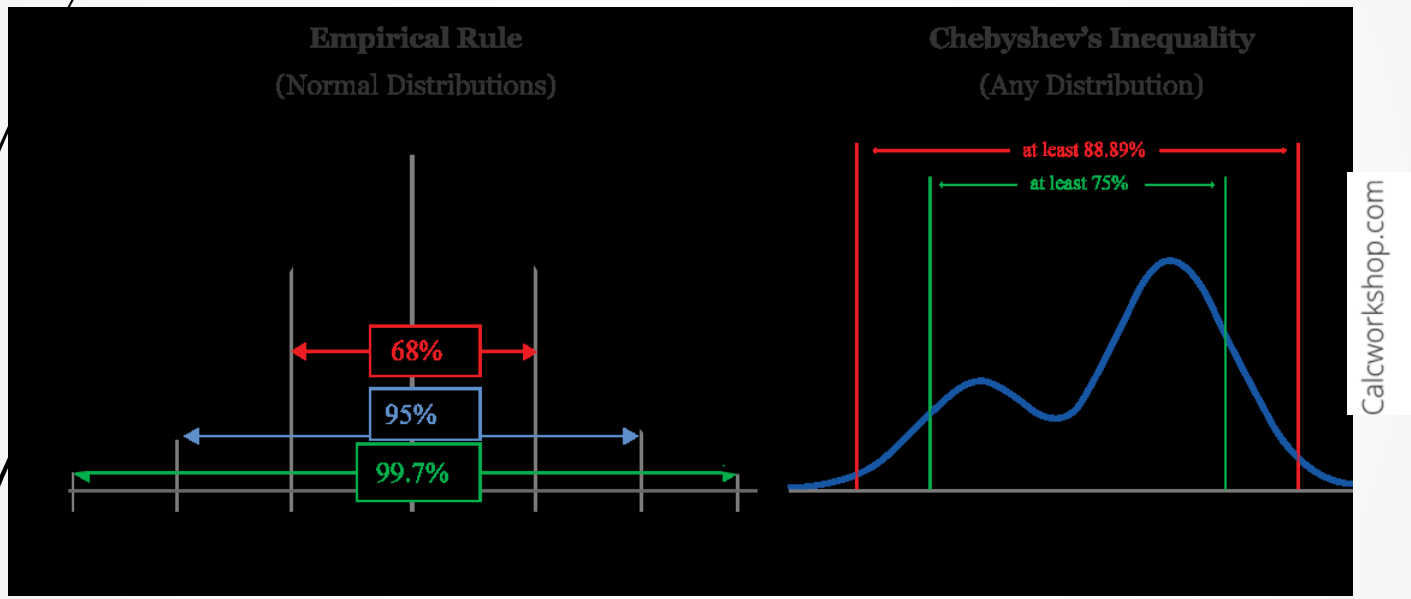
$$p(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \rightarrow p(|X-\mu| \geq 2\sigma) \leq \frac{1}{4}$$

$$p(|X-\mu| < 2\sigma) \geq \frac{3}{4}$$

- No more than a small fraction of X $\left(\frac{1}{k^2}\right)$ can be more than a small distance (k standard deviations) from the mean.

Chebyshev inequality

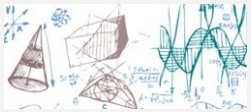
- A minimum of just 75% of values must lie within two standard deviations of the mean and 88.89% within three standard deviations



- It can be applied to any probability distribution in which the mean and variance are defined

- Chebyshev's inequality is more general than 68–95–99.7 rule, which applies only to normal distributions.

Central Limit Theorem



- We have n independent Random Variables X_i.
- X_is follow unknown (but the same type) distribution with parameters:

$$E(X_i) = \mu_i$$
$$VAR(X_i) = \sigma_i^2$$

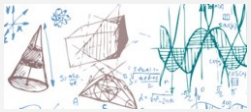
- Now, let's define the NEW RV:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

- What is:

$$E(S_n) = ?$$
$$VAR(S_n) = ?$$

Central Limit Theorem



- We have n independent Random Variables X_i.
- X_is follow unknown (but the same type) distribution with parameters:

$$E(X_i) = \mu_i$$
$$VAR(X_i) = \sigma_i^2$$

- Now, let's define the NEW RV:

$$S_n = X_1 + X_2 + \cdots + X_n.$$

$$E(S_n) = \sum \mu_i$$

$$VAR(S_n) = \sum \sigma_i^2$$

Central Limit Theorem

$$S_n = X_1 + X_2 + \cdots + X_n.$$

- If $n \to \infty$, we have…..

$$Y_n = \frac{S_n - \sum \mu_i}{\sqrt{\sum \sigma_i^2}} \to \mathcal{N}(0,1)$$

- If RV X_i are „the same" (?):

$$\mu_i \equiv \mu$$

$$\sigma_i \equiv \sigma$$

then S_n has:

$$E(S_n) = n\mu$$

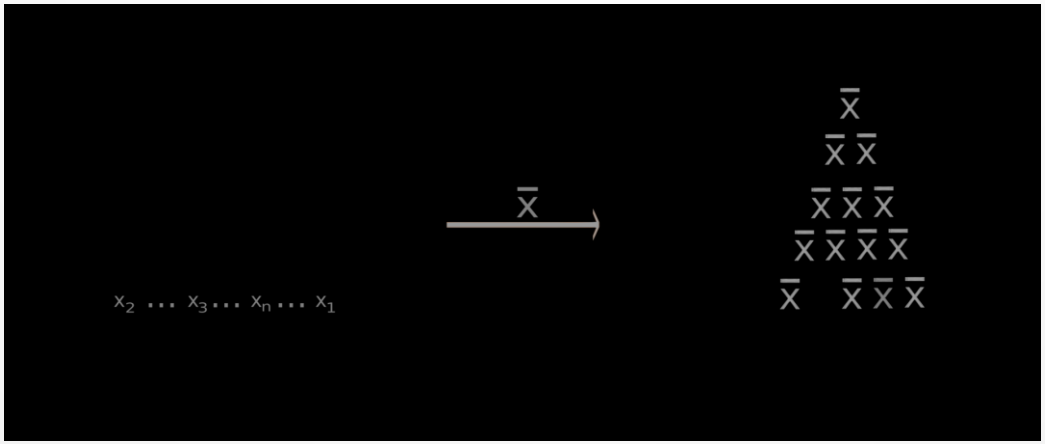
$$VAR(S_n) = n\sigma^2$$

$$Y_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} \to \mathcal{N}(0,1)$$

Central Limit Theorem

$$Y_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \mathcal{N}(0,1)$$

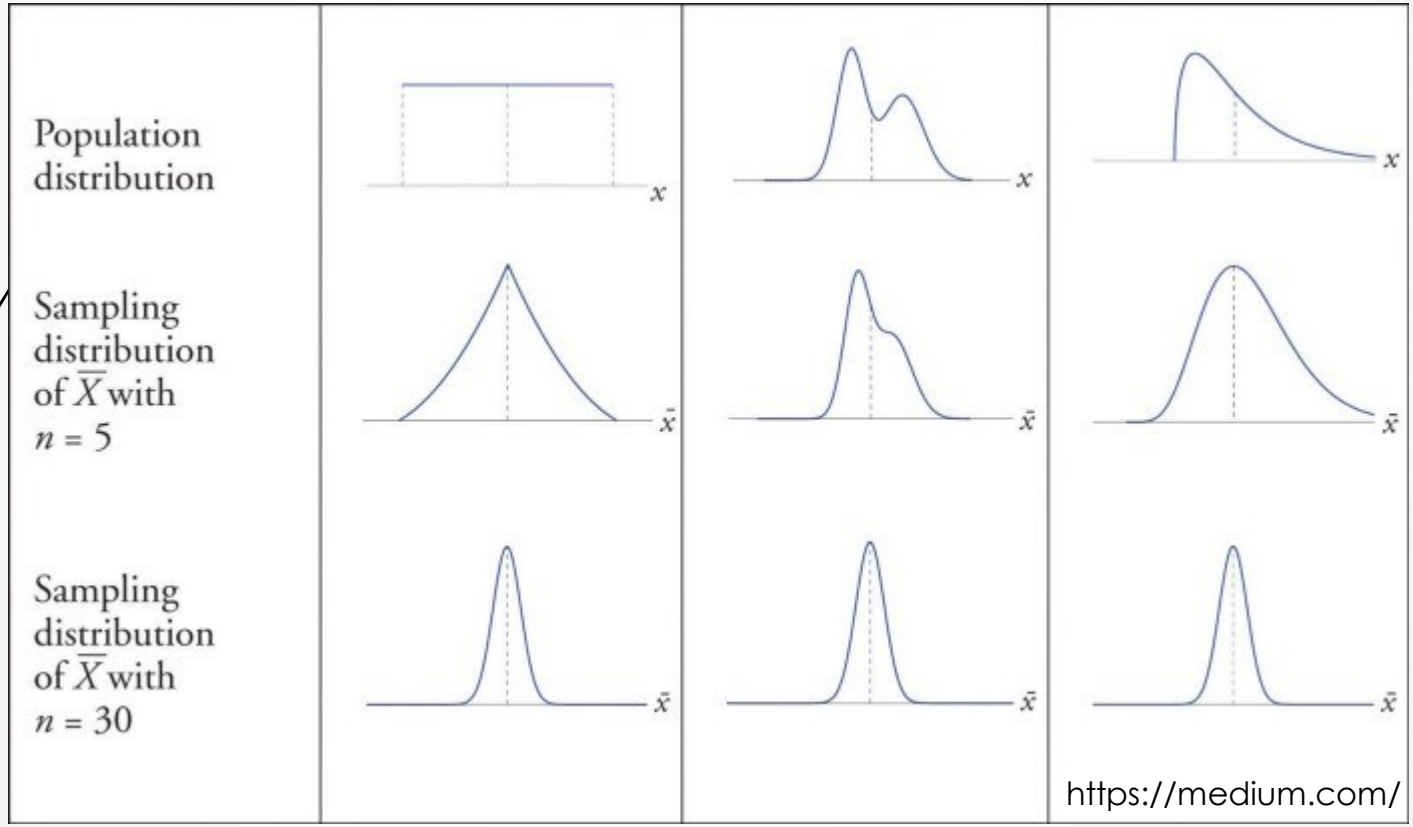
If we are sampling from a population with unknown distribution (finite of Infinite), the distribution of the means $\bar{X}$ is approximately **normal** with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma^2/n}$ provided that the **sample size is large.**

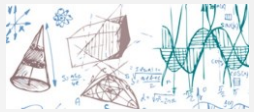


https://en.wikipedia.org/

Central Limit Theorem

$$Y_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} \to \mathcal{N}(0,1)$$





The Law of Large Numbers

- A reminder of the formal definition of **LoLN**

Let $X_1, X_2, \ldots, X_n$ be mutually independent random variables (no particular P.D.F. is assumed here), each of which have finite mean, μ, and variance, σ^2. Now let's us define new R.V.: $S_n = X_1 + X_2 + \cdots + X_n, n = 1, 2, \ldots$ The probability that the arithmetic mean of $X_1, X_2, \ldots, X_n$ differs from its expected value more than ϵ approaches zero as $n \to \infty$:

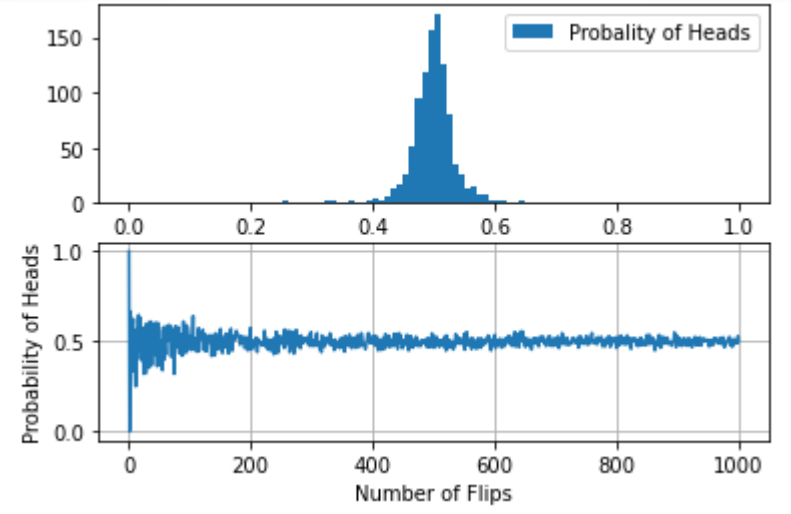
$$\lim_{n\to\infty} p\left(\left|\frac{S_n}{n} - E\left[\frac{S_n}{n}\right]\right| \geq \epsilon\right) = \lim_{n\to\infty} p\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0$$

- The most important property for us is that by repeating the observation we are getting closer to the true answer…
- Well, there is also the systematic uncertainty… unfortunately

The Law of Large Numbers

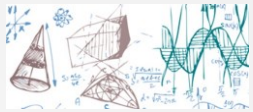
- A reminder of the formal definition of **LoLN**

Let $X_1, X_2, \ldots, X_n$ be mutually independent random variables (no particular P.D.F. is assumed here), each of which have finite mean, μ, and variance, σ^2. Now let's us define new R.V.: $S_n = X_1 + X_2 + \cdots + X_n, n = 1, 2, \ldots$ The probability that the arithmetic mean of $X_1, X_2, \ldots, X_n$ differs from its expected value more than ϵ approaches zero as $n \to \infty$:



https://towardsdatascience.com/law-of-large-numbers-in-finance-using-python-86945eaee444

Estimating probabilities-example

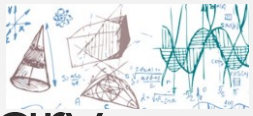


- Imagine that we are interested in estimating the probability of some event: $p = P(X \in A), A = (c, d)$
- What would be the procedure to estimate this? Experiment!!
- Collect a sample $\{X_1, X_2, \dots, X_n\}$ and estimate how often we see $\{X_i \in A\}$, then we calculate the relative frequency by dividing it by sample size n
- How to put it in the context of the LoLN?
- Introduce an indicator R.V. Y_i:

$$Y_i = \begin{cases} 1 \rightarrow X_i \in A \\ 0 \rightarrow X_i \notin A \end{cases}$$

- So, Y_i is the indicator R.V. of the event $X_i \in A$, and its expectation value: $\boldsymbol{E[Y_i] = 1 \cdot P(X \in A) + 0 \cdot P(X \notin A) = P(X \in A) = p}$
- We can write for the relative frequency of the indicator R.V.:

$$\lim_{n \to \infty} P(|\bar{Y}_n - p| > \epsilon) = 0, \bar{Y}_n = (X_1 + X_2 + \cdots + X_n)/n$$



The Central Limit Theorem-summary

- Let assume that we have n I.R.V. identically distributed $X_1, X_2, \dots, X_n$ with defined mean and variance (this means it is finite and positive in case of variance). Now, we define a RV Z_n (we are going to call it **the score** soon…):

$$Z_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$$

- Then, for any number c we have

$$\lim_{n\to\infty} F_{Z_n}(c) = \phi(c)$$

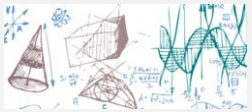
- Here, $\phi(c)$ is CDF of $\mathcal{N}(0,1)$ distribution, so, we say that the CDF of RV Z_n **is almost identical** to the CDF of a **standardised normal distribution**!!! And this is true for any PDF!
- Note that Z_n is in fact the average of standardised sample mean

Central Limit Theorem – Key Terms

Let's discuss the CLT with our textbook:

https://openstax.ohttps://en.wikipedia.org/rg/books/introductory-statistics/pages/7-key-terms

Central Limit Theorem – Key Terms



Average

a number that describes the central tendency of the data; there are a number of specialized averages, including the arithmetic mean, weighted mean, median, mode, and geometric mean.

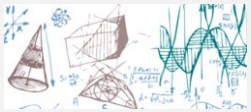
Central Limit Theorem

Given a random variable (RV) with known mean μ and known standard deviation, σ, we are sampling with size n, and we are interested in two new RVs: the sample mean, $\overline{X}$, and the sample sum, ΣX. If the size (n) of the sample is sufficiently large, then $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ and $\Sigma X \sim N(n\mu, (\sqrt{n})(\sigma))$. If the size ($n$) of the sample is sufficiently large, then the distribution of the sample means and the distribution of the sample sums will approximate a normal distributions regardless of the shape of the population. The mean of the sample means will equal the population mean, and the mean of the sample sums will equal n times the population mean. The standard deviation of the distribution of the sample means, $\frac{\sigma}{\sqrt{n}}$, is called the standard error of the mean.

Mean

a number that measures the central tendency; a common name for mean is "average." The term "mean" is a shortened form of "arithmetic mean." By definition, the mean for a sample (denoted by $\bar{x}$) is $\bar{x} = \frac{\text{Sum of all values in the sample}}{\text{Number of values in the sample}}$, and the mean for a population (denoted by μ) is $\mu = \frac{\text{Sum of all values in the population}}{\text{Number of values in the population}}$.

Central Limit Theorem - Review



7.1 The Central Limit Theorem for Sample Means (Averages)

In a population whose distribution may be known or unknown, if the size (n) of samples is sufficiently large, the distribution of the sample means will be approximately normal. The mean of the sample means will equal the population mean. The standard deviation of the distribution of the sample means, called the standard error of the mean, is equal to the population standard deviation divided by the square root of the sample size (n).

7.2 The Central Limit Theorem for Sums

The central limit theorem tells us that for a population with any distribution, the distribution of the sums for the sample means approaches a normal distribution as the sample size increases. In other words, if the sample size is large enough, the distribution of the sums can be approximated by a normal distribution even if the original population is not normally distributed. Additionally, if the original population has a mean of μ_X and a standard deviation of σ_x, the mean of the sums is $n\mu_x$ and the standard deviation is $(\sqrt{n})(\sigma_x)$ where n is the sample size.

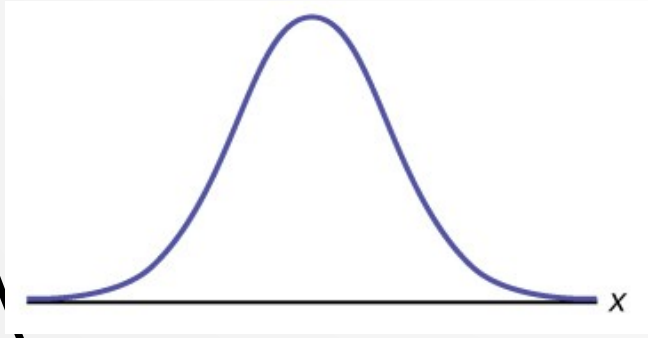
7.3 Using the Central Limit Theorem

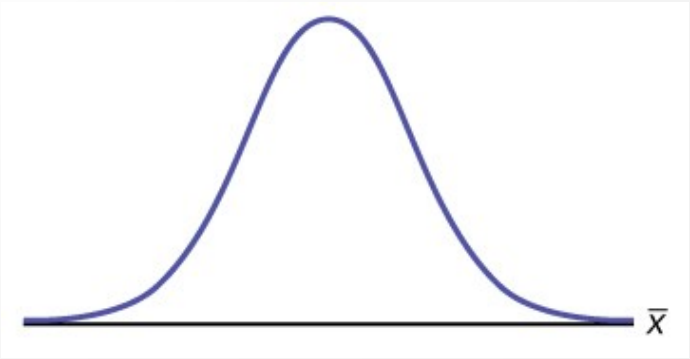
The central limit theorem can be used to illustrate the law of large numbers. The law of large numbers states that the larger the sample size you take from a population, the closer the sample mean $\bar{x}$ gets to μ.

Central Limit Theorem - Practice

Example: **The Central Limit Theorem for Sample Means (Averages)**
Yoonie is a personnel manager in a large corporation. Each month she must review 16 of the employees. From past experience, she has found that the reviews take her approximately four hours each to do with a population standard deviation of 1.2 hours. Let *X* be the random variable representing the time it takes her to complete one review. Assume *X* is normally distributed.

1. What is the mean, standard deviation, and sample size?
2. What is the distribution of: X, $\bar{X}$?
3. Find the probability that **one** review will take Yoonie from 3.5 to 4.25 hours.
4. Find the probability that the **mean** of a month's reviews will take Yoonie from 3.5 to 4.25 hrs.
5. Find the 95th percentile for the mean time to complete one month's reviews.





Central Limit Theorem - Practice

Example: **The Central Limit Theorem for Sums**
An unknown distribution has a mean of 45 and a standard deviation of eight. A sample size of 50 is drawn randomly from the population. Find the probability that the sum of the 50 values is more than 2,400.

1. What is the distribution of X, $\bar{X}$?
2. What is the distribution of ΣX?
3. What is a z-score associated with $\mathrm{Y_n} = \Sigma X$

Central Limit Theorem - Practice

Example: **The Central Limit Theorem for Means and Sums**

1. The measurement of the number of hours in front of the mobile follow **a uniform distribution** with the lowest stress score equal to one and the highest equal to 15. Using a sample of 75 students, find:
   a) The probability that the **mean the number of hours** for the 75 students is less than two.
   b) The probability that the **total of the 75 numbers of hours** is less than 300.

Central Limit Theorem - Practice

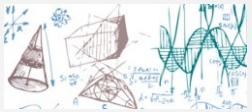
Example: **The Central Limit Theorem for Means and Sums**

2. The time for the next customer to come follows an **exponential distribution** with a mean of 22 minutes.

a) What is the probability that the owner has to wait more than 20 minutes for the customer?

b) Consider 80 shops and find the probability that the mean waiting time is longer than 20 minutes.

Statistical Sample and Population



POPULATION

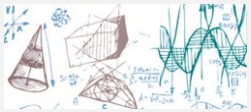
is the set of all the objects (or the totality of observation results with which we are concerned) that posses the property X (our RV), whether their number be finite or infinite.
The statistical inference consists in arriving at (quantitative) conclusions concerning a population where it is impossible or impractical to examine the entire set of observations that make up the population. Instead, we depend on a **subset** of observations — a **sample**.

Property (RV) X — has the pdf: $f(X)$
We may form n samples, each of size m:

$$\begin{array}{rccccc} \text{RV}X: & X_1 & X_2 & \dots & X_m \\ \text{Sample No: } 1 & x_{11} & x_{12} & \dots & x_{1m} \\ 2 & x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots \\ j & x_{j1} & x_{j2} & \dots & x_{jm} \\ \dots & \dots & \dots & \dots & \dots \\ n & x_{n1} & x_{n2} & \dots & x_{nm} \end{array}$$

every $j-$th sample has a pdf: $g_j = g_j(x_{j1}, x_{j2}, \dots, x_{jm})$

Statistical Sample and Population



Random Sample

- Sample which:

  all sample constitutions X_{jl} are independent of each other:

$$g_j(x_{j1}, x_{j2}, \ldots, x_{jm}) = g_{j1}(x_{j1}) g_{j2}(x_{j2}) \ldots g_{jm}(x_{jm})$$

  have the same pdf as X:

$$g_{j1}(x_{j1}) = g_{j2}(x_{j2}) = \ldots = g_{jm}(x_{jm}) = f(x)$$

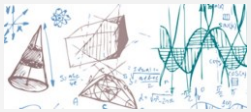
- What is the pdf of X?

  we should write it as a function of set of parameters $\boldsymbol{\lambda}$:

$$f = f(x, \boldsymbol{\lambda})$$

$$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_2) \text{ e.g: } \mu, \sigma, etc$$

Statistical Sample and Population



$\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_2)$ e.g: μ, σ, etc

How do we go about finding $\boldsymbol{\lambda}$?

Any function of the random variables constituting a RV that is used for estimation of unknown parameters is called a

STATISTIC:

$$S = S(X_1, X_2, \ldots, X_n)$$

do we already know any statistic?