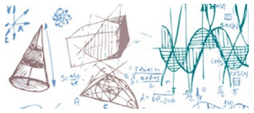
Introduction to probability, statistics and data handling

Agnieszka Obłąkowska-Mucha
Tomasz Szumlak

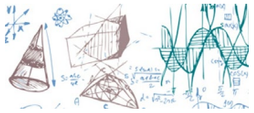
Faculty of Physics and Applied Computer Science
AGH University of Krakow



Data and models

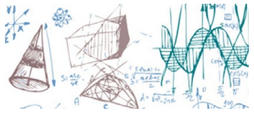
- We have already accumulated a lot of knowledge at this point. Let's stop for a bit and think, how can we enhance our data analysis techniques.
- **First**, we always need data. We make an observation (also known as an experiment) and with given data set trying to come up with **a model** describing it. Here a model means some mathematical formulae.
- Usually, we will not be able to reproduce all features of our data set – need to use **simplifications**. At the same time, we need to assert that our model is **good enough** (sufficient accuracy).
- This reasonably good approach is vital and will drive the **quality of our conclusions** (predictions). Also, need take into account the purpose of our studies (i.e., sometimes it just does not make sense to fight for too much precision!)

Models



- Central concepts of collecting data are **sample** and **population**.
- Having data, we usually go with an attempt to use probability theory to **model the real-world phenomena**
- Models may be very simple (like yes-no situation) or increasingly complicated. Usually, we **start** with **the simplest possible** and build more and more intricate one.
- Say, we analyse insurance data. At first, we may be just interested in whether a claim was made or not. Then, how many claims are typically made by a single person, then the size of the claim… You see, where we are going…
- The **type of data** also has a significant impact: restricted class of outcomes (die roll), discrete data, continuous data.
- We start discussing popular probabilistic models with two special cases: Bernoulli distribution and binomial distribution

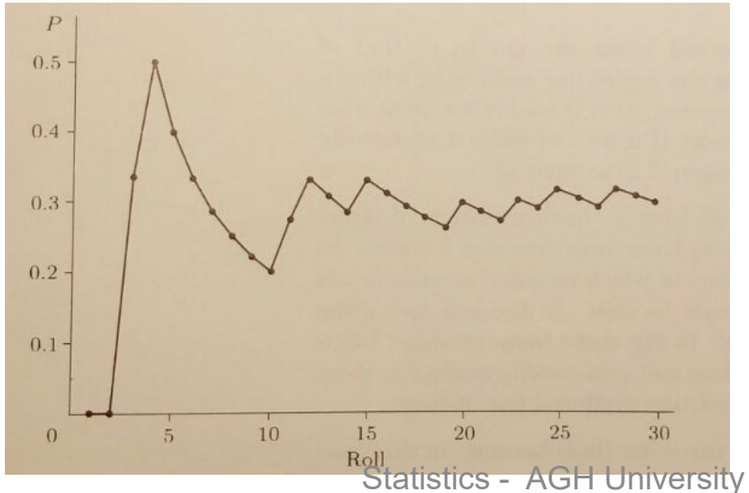
Real world

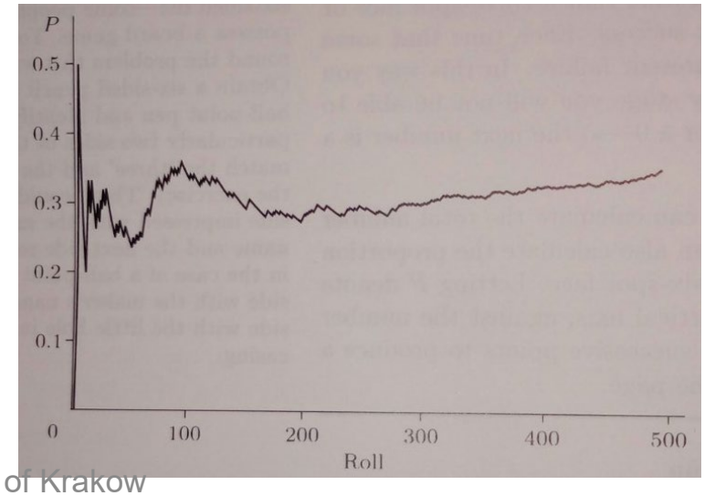


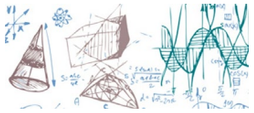
- Imagine the following: *a company acquired a lot of electronics components that are used for the assembly lines. At some point, the quality control started to suspect that this batch is bad*.
- How would you **confirm/reject** such a hypothesis?
- One way would be to start testing the components **one by one**. Say, somebody has been tasked with this job and started to check them.
- The results are: *amongst the first 10 he finds 2 bad, for the first 50 he finds 11 faulty and then 16 for the first 100*.
- The point of the data analysis process is to provide an argument that we can analyse a sample of the whole population and then make a statement about the latter.
- Inspect the data – as we increase the sample, the proportion of faulty components **gets stable**. We need to come up with a rule that makes it formal and depend on the sample size!

Real world

- Another example: *we have a fair die and we want to perform the following experiment – we roll it 30 times and we want to count events with either a two or a four*. How can me make it similar to the previous example?
- Say, we obtained results as below. **Any conclusions**…?
- What about longer trending, say 500 tosses. That would get boring.
- We can use computer and simulate it. We would need a model to perform such simulation. This is a typical situation in science!



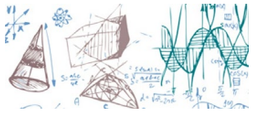




Bernoulli trials

- Formally, we say that a statistical experiment with a **binary** outcome is called **Bernoulli trial**. Repeating such an experiment leads to a sequence of Bernoulli trials.
- By using such a procedure, we can **test a particular model** describing important aspects of given random phenomena.
- The key point that should be stressed is that we assume that we are not able in any way **to predict** the outcome of any of the trials.
- In „stat pro" language the value we observed in described situations are called **sample frequencies** and **sample relative frequencies**.
- For instance, if we perform 1000 tosses of a single coin and obtain **491 heads**, we call it sample frequency. And **by dividing it by the number of tosses**, we get the relative frequency.
- Remember this – *it will be vital for estimation with confidence and hypothesis testing*!

Asking questions



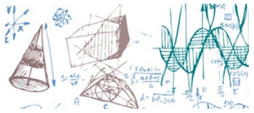
- Remember this – formulating questions is an art. Think about what you would like to know!
- Roulette wheel. *In Monte-Carlo the wheels have 37 compartments numbered from 0 to 36*. Check this out. Our R.V. *X* takes the value 1 if ball stops at ‚19' and 0 otherwise. Assume that the wheel is fair. The R.V. *X* follows a Bernoulli distribution with the p.d.f. given by:

$$p(x) = \left(\frac{1}{37}\right)^{x} \left(\frac{36}{37}\right)^{1-x}, x = 1{,}0$$

- Different question: *at what compartment will the ball rest*? Still assuming the wheel is fair:

$$p(x) = \frac{1}{37}, x = 0{,}1, \dots, 35{,}36$$

Creating a model



- *Fine, we studied a number of experiments, are we ready to build our first model*?
- Consider a fair die. Using symmetry rule and denoting the number of dots rolled by X we can propose the following model:

$$p(x) = \frac{1}{6}, x = 1, 2, \dots, 6$$

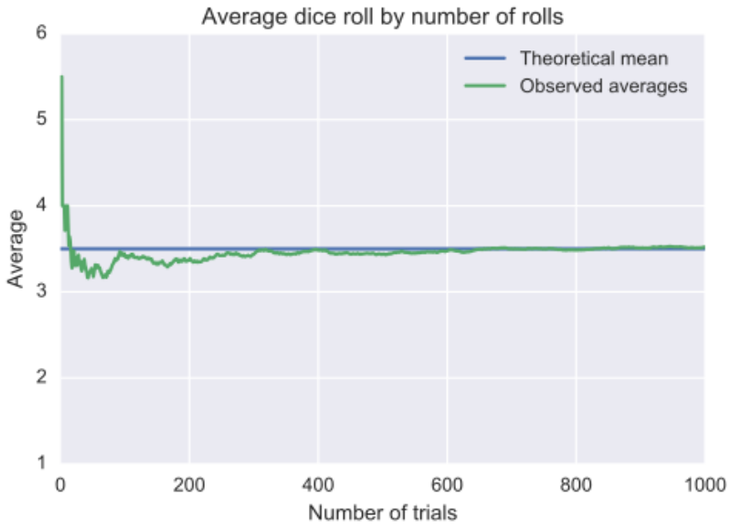
- What can be done with it? **Plenty**!
  - ✓ Estimate probabilities: $P(X = 2 \; or \; X = 6) = p(2) + p(6) = \frac{1}{3}$
  - ✓ Create p.d.f and c.d.f
  - ✓ Compare with experiments to check if a given die is fair
  - ✓ Make simulation experiments
  - ✓ Quite a lot!

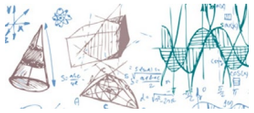
Law of large numbers

- Knowing a few important distributions, we can formulate a very advanced theorem, that is considered fundamental for statistics.
- **Theorem 1**. Let $X_1, X_2, \cdots, X_n$ be mutually independent RV (discrete or continuous), each having finite mean μ and variance σ^2. Then if we take into consideration a new

  RV: $\mathbf{S_n = X_1 + X_2 + \cdots + X_n}$, then:

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





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 $k = 2$:

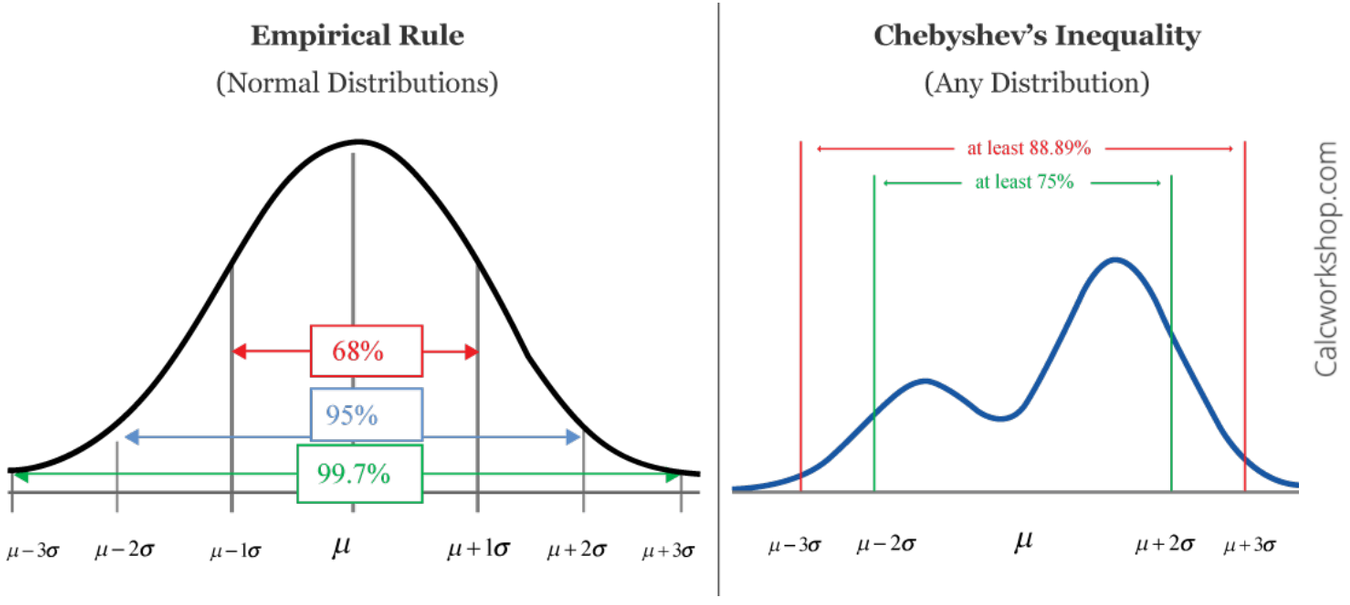
$$p(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \rightarrow p(|X-\mu| \geq 2\sigma) \leq \frac{1}{4}$$

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

  - 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's Inequality

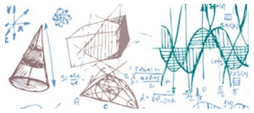
- A minimum of just 75% of values must lie within two standard deviations of the mean and 88.89% within three standard deviations



- 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.

The Law of Large Numbers



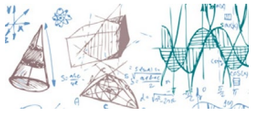
- It is essential to understand, the last point – remember frequency will always depend on a particular sample! Different sample will yield a different frequency.
- Having said that, we can however write (remember that X_j is a binomial R.V.!):

$$E[f_j] = E\left[\frac{X_j}{n}\right] = \boldsymbol{p_j}, E[X_j] = np_j$$

$$\sigma^2(f_j) = \sigma^2\left(\frac{X_j}{n}\right) = \frac{1}{n^2}\sigma^2(X_j) = \frac{1}{n}p_j q_j = \boldsymbol{\frac{1}{n}p_j(1-p_j)}$$

- In words: the expectation value of the frequency (event A_j) is equal to the probability of success. The variance of the frequency about its mean value can, in turn, be made arbitrarily small – just need to collect enough data! (large n).
- **This, actually, is the law of large numbers**!

The Law of Large Numbers



- Imagine that our sample space can be divided into k events (or outcomes that we wish to study): $\{A_j\}, j = 1, \dots, k$. What are the respective probabilities of such events p_j?
- Well, in principle we should conduct an experiment, collect a data sample and then calculate the frequency f_j (we assume that n below means the number of events of type j observed):

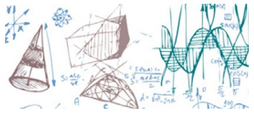
$$f_j = \frac{1}{n}\sum_{i/1}^{i/n} X_{ij} = \frac{1}{n} X_j$$

- So, note that X_j is a binomial R.V. that takes the following values:

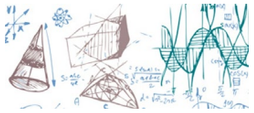
$$X_j = \begin{cases} 1 \; if \, A_j \; occured \\ 0 \quad otherwise \end{cases}$$

- Now, how is f_j related to the probability p_j? Remember, the probability is just a number, whilst the frequency is a R.V.

The Law of Large Numbers



- Let's just think about the variance for a second. It is a product of these two elements: $1/n$ and $p_j(1-p_j)$. The latter is always less than unity (the max value is: $max\{p_j(1-p_j) = 1/4\}$), so the „smallness" of departure **will be governed by the number of observed events**.
- Using this argument and the result from previous slide we can justify that the approach, where **respective probabilities** of events that are estimated by **frequencies measured directly** in experiments, is **the right one**!
- The square of the error we make doing so is inversely proportional to the number of measurements in an experiment – this kind of error is called a **statistical** one
- This is essence of, so called, **counting experiments** such as: number of decaying particles, number of animals with a given traits, number of defective items, …



The Law of Large Numbers

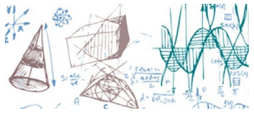
And finally to sum up the **LoLN**

- Let $X_1, X_2, \dots, X_n$ 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.: $S_n = X_1 + X_2 + \cdots + X_n, n = 1, 2, \dots$
- The probability that the arithmetic mean of $X_1, X_2, \dots, X_n$ differs from its expected value more than ϵ approaches zero as $n \to \infty$:

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

as a probabilistic process is repeated a *large number of times*, the relative frequencies of its possible outcomes will get closer and closer to their respective probabilities.

Central Limit Theorem



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

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

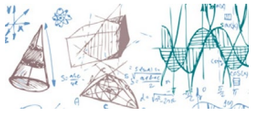
- Now, let's define the NEW RV:

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

- 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:

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

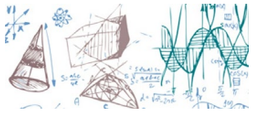
- Now, let's define the NEW RV:

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

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

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

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:

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

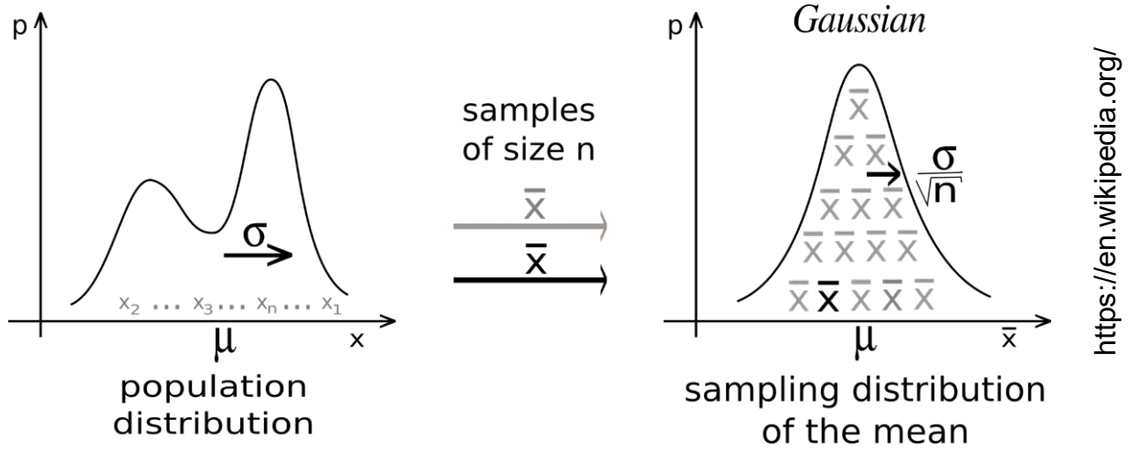
$$VAR(S_n) = n\sigma^2$$

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

Central Limit Theorem

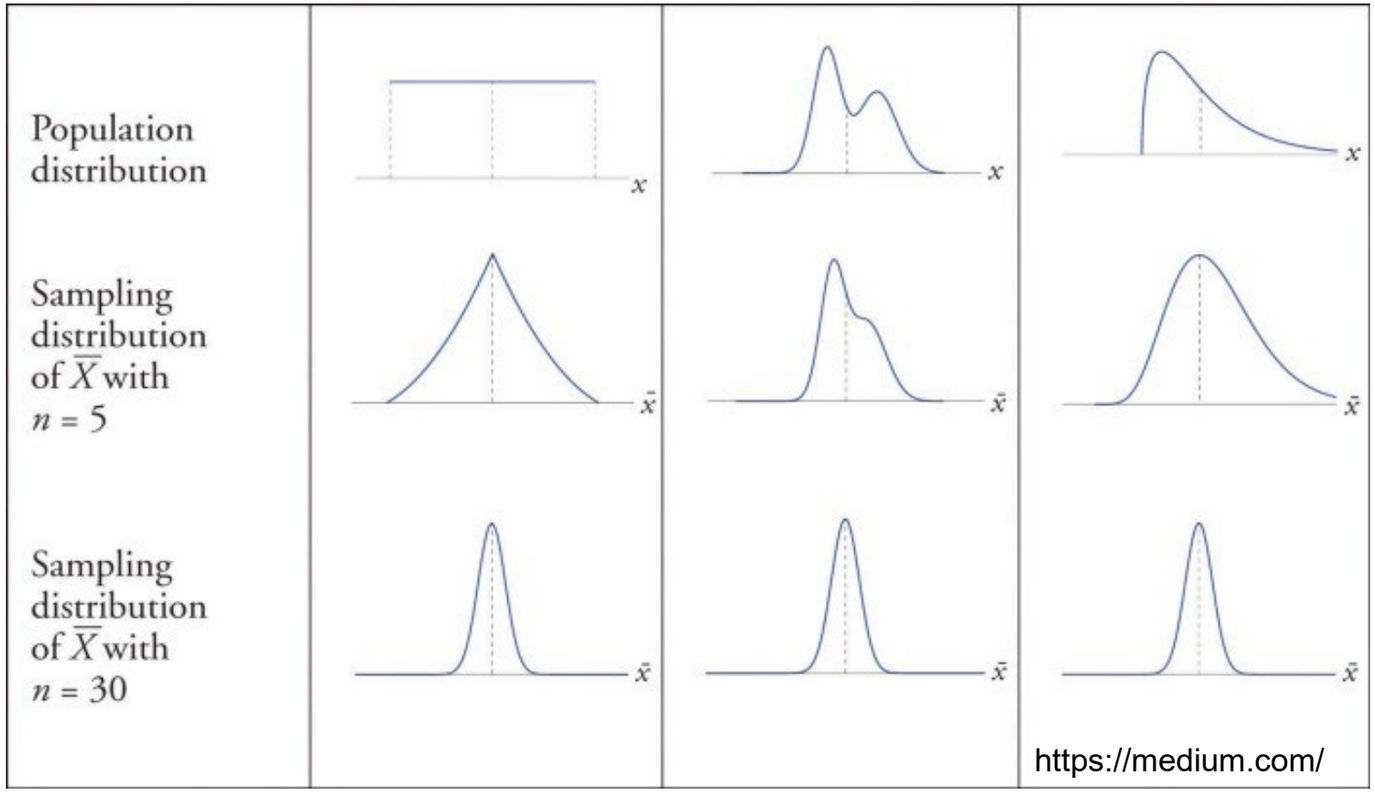
$$Y_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} \to \mathcal{N}(0,1)$$

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.**

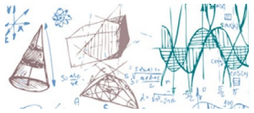


Central Limit Theorem

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



The Law of Large Numbers



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

- Let $X_1, X_2, \dots, X_n$ 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.: $S_n = X_1 + X_2 + \cdots + X_n, n = 1, 2, \dots$ The probability that the arithmetic mean of $X_1, X_2, \dots, X_n$ differs from its expected value more than ϵ approaches zero as $n \to \infty$:

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

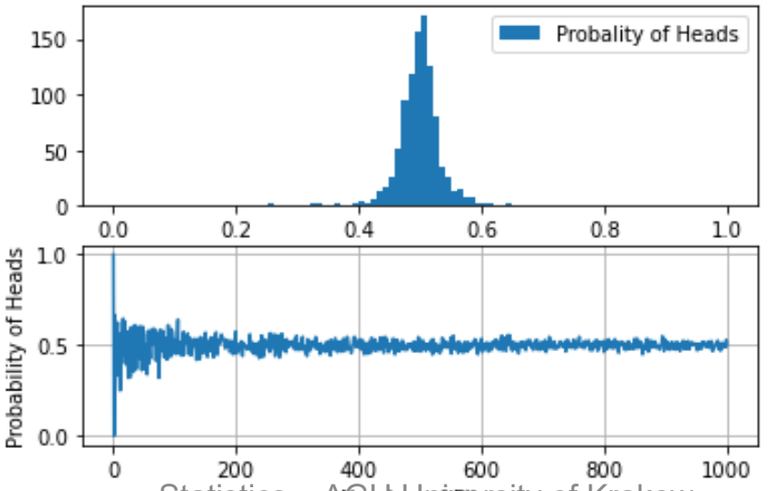
- The most important property for us is that by repeating the observation we are getting closer to the true answer…
- Well, there is also the systematic uncertainty… unfortunately

The Law of Large Numbers

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

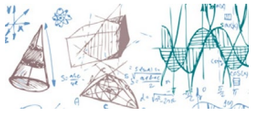
- Let $X_1, X_2, \dots, X_n$ be mutually independent random variables (no particular P.D.F. is assumed here), each of which have finite mean, μ, and variance, σ^2. Now let's us define new R.V.: $S_n = X_1 + X_2 + \cdots + X_n, n = 1, 2, \dots$ The probability that the arithmetic mean of $X_1, X_2, \dots,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

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



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:

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

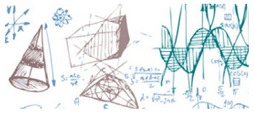
- So, Y_i is the indicator R.V. of the event $X_i \in A$, and its expectation value:

$$\boldsymbol{E[Y_i] = 1 \cdot P(X \in A) + 0 \cdot P(X \notin A) = P(X \in A) = p}$$

- 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

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

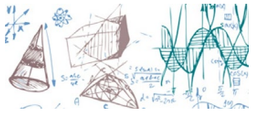
- Here, $\phi(c)$ is CDF of $\mathcal{N}(0,1)$ distribution, so, we say that the CDF of RV Z_n **is almost identical** to the CDF of a **standardised normal distribution**!!! And this is true for any PDF!
- Note that Z_n is in fact the average of standardised sample mean

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.

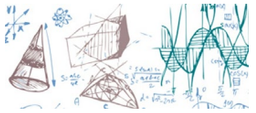
Central Limit Theorem

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 $\left(\sqrt{n}\right)(\sigma_x)$ 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

openstax

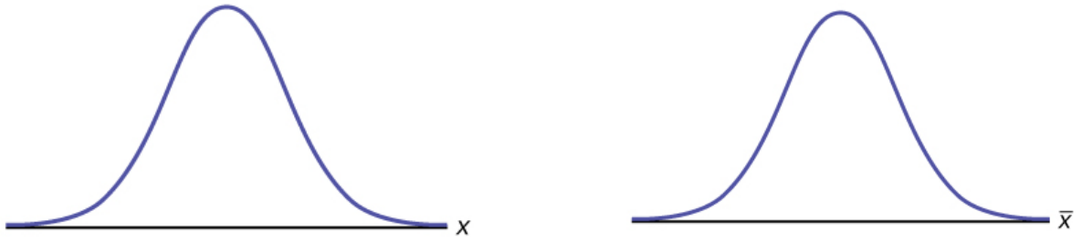
Introductory Statistics

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

openstax™

Introductory Statistics

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

openstax™

< Introductory Statistics

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

openstax™

< Introductory Statistics

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.