Statistics: The Central Limit Theorem

S. Maset

Dipartimento di Matematica e Geoscienze, Università di Trieste

PEM 2016-2017

Outline

The Central Limit Theorem

- Given a finite or infinite sequence $X_1, X_2, X_3, \ldots$ of discrete or continuous or mixed random variables (here and in the following for the same experiment), we say that the random variables of the sequence are **Independent and Identically Distributed (IID)** if they are independent and have the same distribution.

- One of the most important results in Probability Theory, known as the **Central Limit Theorem**, states that the sum of a large number of IID random variables is approximately normally distributed.

  In order to introduce it, we need the following notion. Let X be a discrete or continuous random variable. The random variable

$$X^* := \frac{X - \mathbb{E}(X)}{\mathrm{SD}(X)}$$

  is called the **standardized form** of X.

  We have already considered the standardized form $Z = \frac{X-\mu}{\sigma}$ of a normal random variable X with distribution $\mathrm{N}(\mu, \sigma)$. Z is a standard normal random variable: it has distribution $\mathrm{N}(0, 1)$.

Exercise. Find mean and variance of the standardized form X^* of a general random variable X.

Here is the precise statement of the Central Limit Theorem.

Theorem

*(**Central Limit Theorem**) Let $X_1, X_2, X_3, \ldots$ be an infinite sequence of IID discrete or continuous or mixed random variables. For any $n \in \{1, 2, 3, \ldots\}$, let*

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

be the sum of the first n random variables. Then

$$\lim_{n \to \infty} F_{S_n^*}(x) = \Phi(x) \quad \text{for any } x \in \mathbb{R},$$

where $F_{S_n^*}$ is the distribution function of the standardized form S_n^* of S_n and Φ is the distribution function of a standard normal random variable.

The proof of this theorem is outside the scope of the course.

The normal approximation for S_n

- Let μ and σ be the common mean and standard deviation of the IID random variables $X_1, X_2, X_3, \ldots$ For $n \in \{1, 2, 3, \ldots\}$, we have

$$\begin{aligned} \mathbb{E}(S_n) &= \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n) = n\mu \\ \mathrm{Var}(S_n) &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) = n\sigma^2. \end{aligned}$$

The Central Limit Theorem says that, for n large,

$$S_n^* = \frac{S_n - \mathbb{E}(S_n)}{\mathrm{SD}(S_n)} = \frac{S_n - n\mu}{\sqrt{n}\sigma}.$$

is approximately distributed as a standard normal variable Z, whose distribution is $\mathrm{N}(0, 1)$.

Then, for n large,

$$S_n = n\mu + \sqrt{n}\sigma S_n^*$$

is approximately distributed as the random variable

$$Y_n = n\mu + \sqrt{n}\sigma Z,$$

which has the normal distribution

$$\mathrm{N}\left(n\mu + \sqrt{n}\sigma \cdot 0, (\sqrt{n}\sigma \cdot 1)^2\right) = \mathrm{N}(n\mu, (\sqrt{n}\sigma)^2).$$

This means that, for any $a, b \in \mathbb{R}$ with $a < b$, we have, for n large,

$$\mathbb{P}(S_n \leq a) \approx \mathbb{P}(Y_n \leq a) = \Phi\left(\frac{a - n\mu}{\sqrt{n}\sigma}\right)$$

$$\mathbb{P}(S_n > b) \approx \mathbb{P}(Y_n > b) = 1 - \Phi\left(\frac{b - n\mu}{\sqrt{n}\sigma}\right)$$

$$\mathbb{P}(a < S_n \leq b) \approx \mathbb{P}(a < Y_n \leq b) = \Phi\left(\frac{b - n\mu}{\sqrt{n}\sigma}\right) - \Phi\left(\frac{a - n\mu}{\sqrt{n}\sigma}\right).$$

- Observe that if $X_1, X_2, X_3, \ldots$ have distribution $\mathrm{N}(\mu, \sigma^2)$, then , for any n, S_n is exactly (and not only approximately) distributed as

$$\mathrm{N}(n\mu, n\sigma^2) = \mathrm{N}(n\mu, (\sqrt{n}\sigma)^2)$$

and so

$$S_n^* = \frac{S_n - n\mu}{\sqrt{n}\sigma} = Z$$

is a standard normal random variable and

$$Y_n = n\mu + \sqrt{n}\sigma Z = S_n.$$

- Example. A US insurance company has around 10^4 car policyholders. Assume that the distribution of the yearly claim (a mixed random variable) is the same for all policyholders with mean 260$ and standard deviation 800$.

  What is the probability that the total yearly claim exceeds 2.8 million $?

  Let $n \approx 10^4$ be the number of car policyholders. For $k \in \{1, 2, \ldots, n\}$, let X_k be the yearly claim of the $k-$th policyholder. The total yearly claim is

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

  The random variables $X_1, X_2, \ldots, X_n$ have a common distribution of mean $\mu = 260\$$ and standard deviation $\sigma = 800\$$.

  It is also reasonable assume that $X_1, X_2, \ldots, X_n$ are independent.

Therefore, since n is large, the Central Limit Theorem says that

$$\begin{aligned}
\mathbb{P}\left(S_n > b = 2.8 \cdot 10^6\right) &\approx \mathbb{P}\left(Y_n > b\right) \\
&= 1 - \Phi\left(\frac{b - n\mu}{\sqrt{n}\sigma}\right) \\
&= 1 - \Phi\left(\frac{2.8 \cdot 10^6 - 10^4 \cdot 260}{10^2 \cdot 800}\right) \\
&= 1 - \Phi\left(\frac{20 \cdot 10^4}{8 \cdot 10^4}\right) = 1 - \Phi\left(\frac{20}{8}\right) \\
&= 1 - \Phi\left(2.5\right) = 1 - 0.9938 \\
&= 0.62\%.
\end{aligned}$$

So, although we do not know the distribution of the yearly claim, we can conclude that it is almost sure that the insurance company has not to pay the amount of 2.8 million \$ for claims.

When the normal approximation for S_n is valid?

- The Central Limit Theorem leaves open the following question: how large has to be n in order to have a valid normal approximation for S_n?

  The answer, of course, depends on the common distribution of $X_1, X_2, X_3, \ldots$ For example, if this distribution is normal, then, for any n, S_n has exactly a normal distribution.

  A general rule of thumb is the following: for any common distribution of $X_1, X_2, X_3, \ldots$, one can be confident that the normal approximation is valid whenever n is at least 30.

  In many cases the normal approximation is valid for a much smaller n.

- Example. Consider the **exponential distribution**

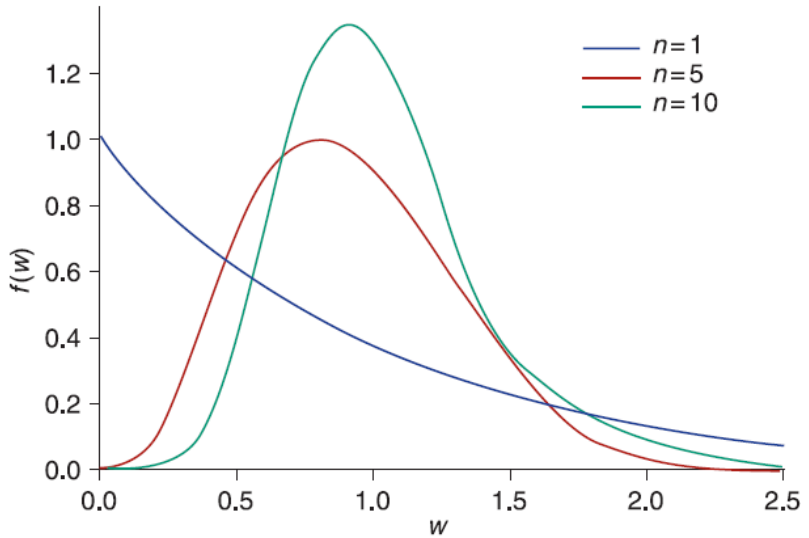
$$f_X(x) = \begin{cases} 0 \text{ if } x < 0 \\ e^{-x} \text{ if } x \geq 0 \end{cases}, \ x \in \mathbb{R}.$$

Exercise. Check that this a valid pdf, i.e.

$$\int_{x \in \mathbb{R}} f_X(x) dx = 1.$$

Find for a random variable with this distribution, the moment generating function and then the mean μ and the variance σ^2.

In figure, we see the distribution of $\frac{S_n}{n}$ in case of the exponential distribution for $n = 1, 5, 10$.



Since, for n large, S_n has distribution close to

$$\mathrm{N}(n\mu, (\sqrt{n}\sigma)^2),$$

the distribution of $\frac{S_n}{n}$ is close, for n large, to the distribution

$$\mathrm{N}\left(\frac{n\mu}{n}, \left(\frac{\sqrt{n}\sigma}{n}\right)^2\right) = \mathrm{N}\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right).$$

The normal approximation of the binomial distribution

- Consider a Bernoulli process of length n, where, at any trial, there are two outcomes α and β with probability p and q, respectively.

  Let, for $i \in \{1, 2, \ldots n\}$, X_i be the discrete random variable given by

$$X_i = \begin{cases} 1 \text{ if the outcome of the } i-\text{th trial is } \alpha \\ 0 \text{ if the outcome of the } i-\text{th trial is } \beta. \end{cases}$$

  Since $X_1, X_2, \ldots, X_n$ are IID with common mean

$$\mu = 1 \cdot p + 0 \cdot q = p$$

  and common variance

$$\sigma^2 = \left(1^2 \cdot p + 0^2 \cdot q\right) - p^2 = p - p^2 = pq,$$

  the Central Limit Theorem says that, for large n,

$$S_n = X_1 + X_2 + \cdots + X_n = \text{number of occurrences of } \alpha \text{ in the } n \text{ trials}$$

  has approximately the normal distribution $\mathrm{N}\left(np, \left(\sqrt{n} \cdot \sqrt{pq}\right)^2\right)$.

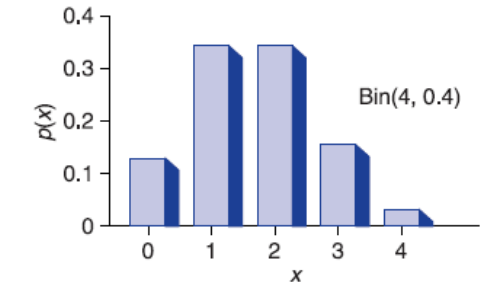
On the other hand, we know that S_n has distribution $\text{binomial}\,(n, p)$.

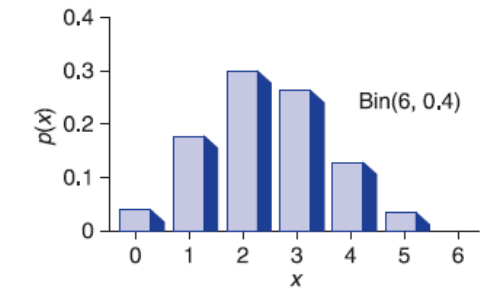
Thus, we can conclude that: for n large, the distribution $\text{binomial}\,(n, p)$ is approximately the normal distribution $\text{N}\left(np, (\sqrt{n} \cdot \sqrt{pq})^2\right)$.

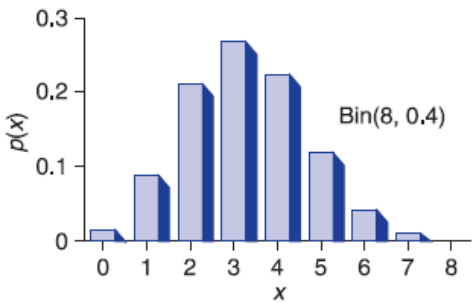
Therefore, we can approximate the probabilities related to a binomial random variable with probabilities related to a normal random variable.

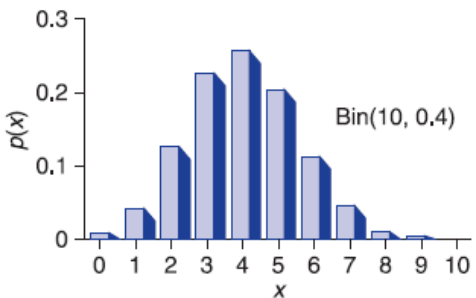
Rule of thumb: the normal approximation to the binomial distribution is quite good when np and nq are greater than 5.

Pmfs for $\text{binomial}(n, 0.4)$ in case of $n = 4, 6, 8, 10$. As n increases, the pmfs approaches to a normal pdf.









- Since S_n is a discrete random variable, the probabilities $\mathbb{P}(S_n = k)$, $k \in \{0, 1, \ldots, n\}$, are not zero, as in case of a continuous random variable.

  Therefore, for n large, we approximate such probabilities with

$$\begin{aligned} \mathbb{P}(S_n = k) &= \mathbb{P}(k - 0.5 < S_n \leq k + 0.5) \\ &\approx \mathbb{P}(k - 0.5 < Y_n \leq k + 0.5), \end{aligned}$$

  instead of $\mathbb{P}(Y_n = k) = 0$, where Y_n is the normal random variable whose distribution $\mathrm{N}\left(np, (\sqrt{n} \cdot \sqrt{pq})^2\right)$ is close to the distribution of S_n. This approximation is called **integer correction**.

  Exercise. Compute the exact pmf of the distribution $\mathrm{binomial}\,(8, 0.4)$ and the approximate pmf computed by the integer correction.

- We have previously seen figures where there is evidence that, as n increases, the pmfs of a distribution $\text{binomial}(n, p)$ approaches to a normal pdf.

  This can explained by the integer correction: for $k \in \{0, 1, \ldots, n\}$, we have, for n large,

$$\begin{aligned} \mathbb{P}(S_n = k) &\approx \mathbb{P}(k - 0.5 < Y_n \leq k + 0.5) \\ &= \int_{k-0.5}^{k+0.5} f_{Y_n}(x)dx \\ &= f_{Y_n}(\xi) \text{ by the Mean Value Theorem} \end{aligned}$$

  where $\xi \in [k - 0.5, k + 0.5]$ and f_{Y_n} is the normal pdf $\text{N}(np, (\sqrt{n} \cdot \sqrt{pq})^2)$.

- For n large and for any $a, b \in \{0, 1, \ldots, n\}$ with $a \leq b$, we have by the integer correction

$$\mathbb{P}(a \leq S_n \leq b) \approx \mathbb{P}(a - 0.5 < Y_n \leq b + 0.5)$$

In fact

$$\begin{aligned} \mathbb{P}(a \leq S_n \leq b) &= \sum_{k=a}^{b} \mathbb{P}(S_n = k) \approx \sum_{k=a}^{b} \mathbb{P}(k - 0.5 < Y_n \leq k + 0.5) \\ &= \mathbb{P}(a - 0.5 < Y_n \leq b + 0.5)\,. \end{aligned}$$

However, observe that

$$\begin{aligned} &\mathbb{P}(a - 0.5 < Y_n \leq b + 0.5) \\ &= \mathbb{P}(a - 0.5 < Y_n \leq a) + \mathbb{P}(a < Y_n \leq b) + \mathbb{P}(b < Y_n \leq b + 0.5) \\ &\approx \mathbb{P}(a < Y_n \leq b) \end{aligned}$$

if

$$\mathbb{P}(a - 0.5 < Y_n \leq a)\,, \mathbb{P}(b < Y_n \leq b + 0.5) \;<<\; \mathbb{P}(a < Y_n \leq b)$$

and this happens when $0.5 \;<<\; b - a$.

- Observe that the probability $\mathbb{P}(a \leq S_n \leq b)$ can be exactly computed by

$$\mathbb{P}(a \leq S_n \leq b) = \sum_{k=a}^{b} \mathbb{P}(S_n = k) = \sum_{k=a}^{b} \binom{n}{k} p^k q^{n-k}. \tag{1}$$

However, when n is large, the sum in (1) could have many terms and the products $p^k q^{n-k}$ could be very small and give underflow.

This is one of the reasons for using the normal approximation of the binomial distribution.

- Example. Consider a ballot with two candidates A and B. Assume that exactly $p = 46\%$ of the voters support the candidate A.

  If a representative sample of voters of size n is randomly chosen from the population of the voters, what is the probability that at least $\frac{n}{2}$ voters support A?

  Observe that this is the probability that a pool based on this sample will predict a win of A instead of B.

  We consider the Bernoulli process of lenght n, where at the $i-$th trial, $i \in \{1, \ldots, n\}$, the outcome is the candidate supported by the the $i-$th voter of the sample: the outcome α is the candidate A (probability $p = 46\%$) and the outcome β is the candidate B (probability $q = 54\%$).

  Therefore, the numbers of voters in the sample supporting the candidate A is given by the random variable S_n, whose distribution is $\mathrm{binomial}(n, p)$.

We have

$$\begin{aligned}
\mathbb{P}\left(S_n \geq \frac{n}{2}\right) &= \mathbb{P}\left(\frac{n}{2} \leq S_n \leq n\right) \\
&\approx \mathbb{P}\left(\frac{n}{2} - 0.5 < Y_n \leq n + 0.5\right) \\
&= \Phi\left(\frac{n + 0.5 - np}{\sqrt{n} \cdot \sqrt{pq}}\right) - \Phi\left(\frac{\frac{n}{2} - 0.5 - np}{\sqrt{n} \cdot \sqrt{pq}}\right) \\
&= \Phi\left(\sqrt{n} \cdot \frac{1 + \frac{0.5}{n} - p}{\sqrt{pq}}\right) - \Phi\left(\sqrt{n} \cdot \frac{\frac{1}{2} - \frac{0.5}{n} - p}{\sqrt{pq}}\right).
\end{aligned}$$

So, for $n = 100$, we have

$$\mathbb{P}\left(S_n \geq \frac{n}{2}\right) \approx \Phi(10.9) - \Phi(0.70) = 1 - 0.7580 = 24.2\%$$

and, for $n = 1000$, we have

$$\mathbb{P}\left(S_n \geq \frac{n}{2}\right) \approx \Phi(34.3) - \Phi(2.51) = 1 - 0.9940 = 0.6\%.$$

Exercise. In the previous example, use $\mathbb{P}\left(\frac{n}{2} < Y_n \leq n\right)$, instead of $\mathbb{P}\left(\frac{n}{2} - 0.5 < Y_n \leq n + 0.5\right)$, as an approximation of $\mathbb{P}\left(S_n \geq \frac{n}{2}\right)$.

- Exercise. A machine producing pieces produces a defective piece with probability 1%. What is the probability that in a set of 1000 pieces produced by the machine more than 1% of them are defective? Use the normal approximation.

Why the normal distribution appears so frequently?

- The preceding version of the Central Limit Theorem assumes that $X_1, X_2, X_3, \ldots$ are Independent and Identically Distributed random variables.

  There is a more general version of the Central Limit Theorem, where the assumption that the random variables are Identically Distributed is not longer necessary.

  In this version, one proves that the sum

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

  is, for n large, approximately normal under the assumption that $X_1, X_2, X_3, \ldots$ are independent and another assumption, not explicitly given here, that all the random variables tend to be of roughly the same magnitude.

The fact that all the random variables tend to be of roughly the same magnitude guarantees that none of them tends to dominate the value of the sum.

For example, suppose that X_1 has a value much larger than $X_2, \ldots, X_n$, so that $S_n \approx X_1$. Then, S_n has approximately the distribution of X_1, not a normal distribution.

- This last version of the Central Limit Theorem explains our previous observation that:
  - in a process that produces a final result, which is programmed in some measure but it is also influenced by many random factors, numerical quantities related to this final result are normally distributed.

The explanation is that these numerical quantities can be seen as the sum S_n of a large number n of independent random variables $X_1, X_2, \ldots, X_n$ related to random factors.

The fact that the result is in some measure programmed means that the mean of these random variables $X_1, X_2, \ldots, X_n$ can be nonzero.

- Example. Consider the formation process of an individual from the birth to the adulthood.

  If this formation process is split in n small successive independent steps, corresponding to short periods of time in the interval from the birth to the adulthood, we can think that, at the i-th step, $i \in \{1, \ldots, n\}$, there is a small random variation X_i of the quantity of interest (for example, the height of the individual).

  The mean of the random variations $X_1, X_2, \ldots, X_n$ is not zero, since the growth of the individual is programmed by the genes.

  The quantity of interest in adulthood is then the sum S_n of $X_1, X_2, \ldots, X_n$. Since the periods of time are short, n is large and then S_n turns out to be normally distributed.

- Exercise. By using the Central Limit Theorem explain why the time to travel by car from the city A to the city B is normally distributed.

- The Central Limit Theorem was stated and proved by the French mathematician Pierre-Simon Laplace (1749-1827)

Pierre Simon,
Marquis de Laplace

to provide a theoretical justification of the empirical fact that a

$$\text{measurement error} = \text{measured value} - \text{true value}$$

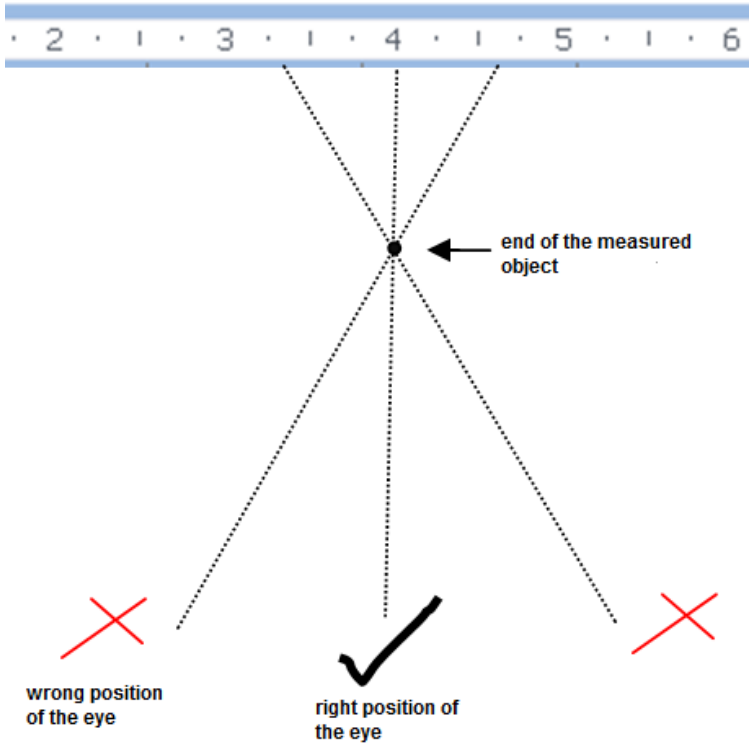
tends to be normally distributed. This fact was called **Law of Frequency of the Error**.

The Law of Frequency of the Error follows by the Central Limit Theorem by considering the measurement error as the sum S_n of a large number n of small independent errors $X_1, X_2, \ldots, X_n$.

For example, the error in measuring a distance by means of a rope can be regarded as being equal to the sum of small errors caused by random factors as:

1. Wrong positions of the initial zero of the rope;
2. Bending of the rope due to its weight;
3. Elastic effects (the rope can be more o less pulled);
4. Vibrations of the rope due to the wind;
5. Temperature modifying the length of the rope (in case of a metal rope);

6. Error in reading the length (for example parallax errors);



7. Random changes of the length of the object to be measured.

We remark that the adjective "gaussian" of the normal distribution comes from the use of the Law of the Frequency of the Error that Carl Friedrich Gauss (1777-1855)

Karl F. Gauss

did in the astronomical measurements.

Exercise. If the measurement error has the normal distribution $\mathrm{N}(\mu, \sigma^2)$ (μ is called the systematic error), what is the distribution of the measured value?

The Sample Mean

- Consider $X_1, X_2, \ldots, X_n$ IID random variables with mean μ and standard deviation σ.

  The random variable

  $$\overline{X} := \frac{S_n}{n}, \; S_n = X_1 + X_2 + \cdots + X_n,$$

  is called the **Sample Mean** of $X_1, X_2, \ldots, X_n$.

  The reason for "Sample" is that in many contexts $X_1, X_2, \ldots, X_n$ represent numerical quantities obtained by a random sample taken from a population.

  The Central Limit Theorem says that, for n large, S_n is approximately distributed as the normal random variable $Y_n = n\mu + \sqrt{n}\sigma Z$ with distribution $\mathrm{N}\left(n\mu, \left(\sqrt{n}\sigma\right)^2\right)$.

Then, for n large, the Sample Mean

$$\overline{X} = \frac{S_n}{n}$$

is approximately distributed as the random variable

$$\frac{Y_n}{n} = \frac{n\mu + \sqrt{n}\sigma Z}{n} = \mu + \frac{\sigma}{\sqrt{n}} Z$$

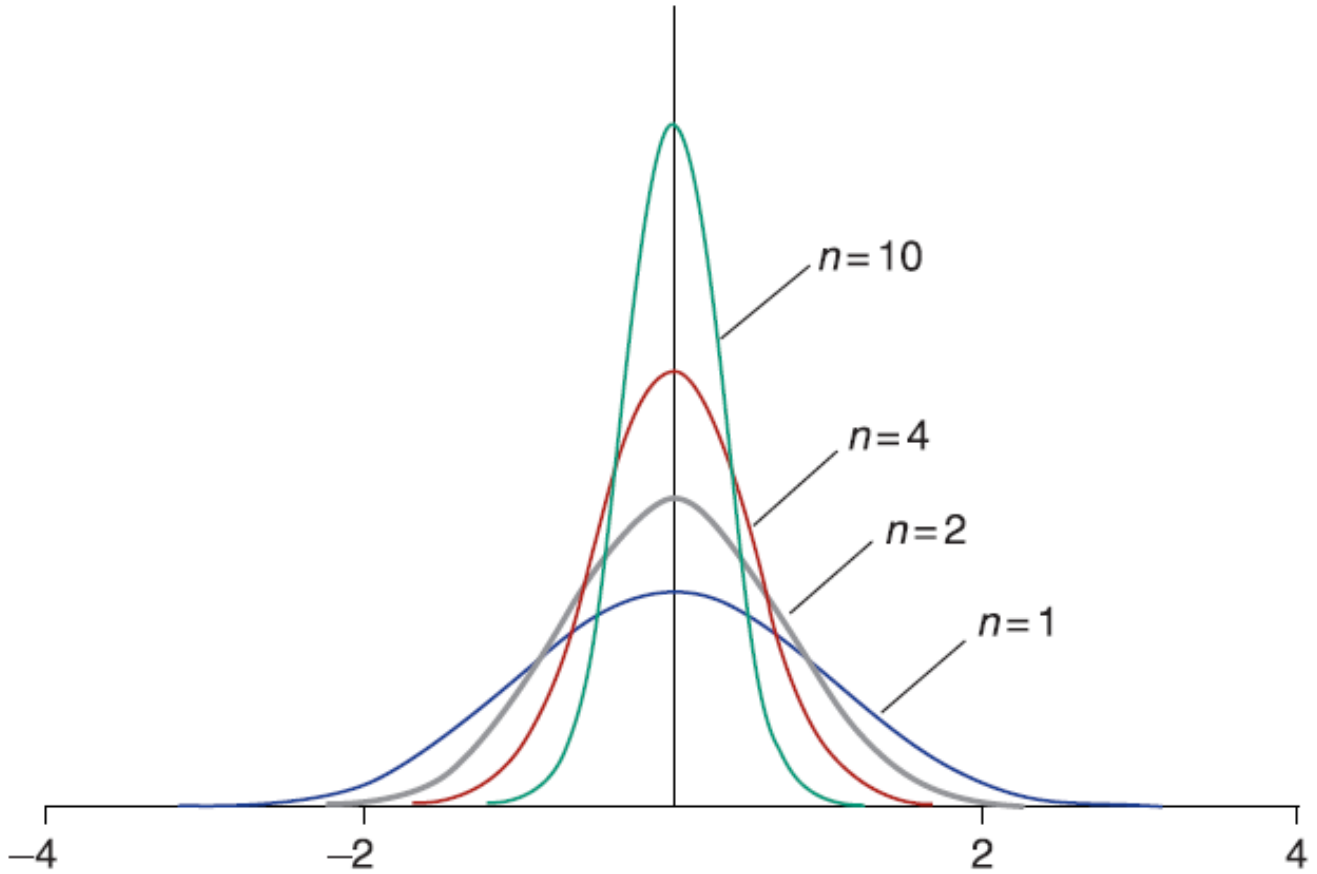
with distribution

$$\mathrm{N}\left(\frac{n\mu}{n}, \left(\frac{\sqrt{n}\sigma}{n}\right)^2\right) = \mathrm{N}\left(\mu + \frac{\sigma}{\sqrt{n}}0, \left(\frac{\sigma}{\sqrt{n}}1\right)^2\right) = \mathrm{N}\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right).$$

Thus, for n large, the values of the Sample Mean are concentrated around μ.

- If $X_1, X_2, \ldots, X_n$ are normally distributed, then, for any n, $S_n = Y_n$ and so $\overline{X} = \frac{Y_n}{n}$ is exactly distributed as $\mathrm{N}(\mu, (\frac{\sigma}{\sqrt{n}})^2)$.

In the figure we see, for several values of n, the pdfs of the Sample Mean $\overline{X}$ when $X_1, X_2, \ldots, X_n$ have the standard normal distribution.



In this case, $\overline{X}$ has the normal distribution $\mathrm{N}\left(0, \left(\frac{1}{\sqrt{n}}\right)^2\right)$.

- Since $\overline{X}$ is approximately distributed as $\frac{Y_n}{n}$, whose distribution is $\mathrm{N}\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$, we have, for n large and for $a, b \in \mathbb{R}$ with $a < b$,

$$\mathbb{P}\left(a < \overline{X} \leq b\right) \approx \mathbb{P}\left(a < \frac{Y_n}{n} \leq b\right) = \Phi\left(\frac{\sqrt{n}}{\sigma}(b-\mu)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(a-\mu)\right)$$

and also

$$\mathbb{P}\left(a < \overline{X} < b\right) \approx \mathbb{P}\left(a < \frac{Y_n}{n} < b\right) = \Phi\left(\frac{\sqrt{n}}{\sigma}(b-\mu)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(a-\mu)\right)$$

$$\mathbb{P}\left(a \leq \overline{X} < b\right) \approx \mathbb{P}\left(a \leq \frac{Y_n}{n} < b\right) = \Phi\left(\frac{\sqrt{n}}{\sigma}(b-\mu)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(a-\mu)\right)$$

$$\mathbb{P}\left(a \leq \overline{X} \leq b\right) \approx \mathbb{P}\left(a \leq \frac{Y_n}{n} \leq b\right) = \Phi\left(\frac{\sqrt{n}}{\sigma}(b-\mu)\right) - \Phi\left(\frac{\sqrt{n}}{\sigma}(a-\mu)\right).$$

In particular, for $c > 0$, we have, for n large,

$$\begin{aligned}
\mathbb{P}(\mu - c < \overline{X} \leq \mu + c) &\approx \Phi\left(\frac{\sqrt{n}}{\sigma}c\right) - \Phi\left(-\frac{\sqrt{n}}{\sigma}c\right) \\
&= \Phi\left(\frac{\sqrt{n}}{\sigma}c\right) - \left(1 - \Phi\left(\frac{\sqrt{n}}{\sigma}c\right)\right) \\
&= 2\Phi\left(\frac{\sqrt{n}}{\sigma}c\right) - 1.
\end{aligned}$$

and also

$$\begin{aligned}
\mathbb{P}(\mu - c < \overline{X} < \mu + c) &\approx 2\Phi\left(\frac{\sqrt{n}}{\sigma}c\right) - 1 \\
\mathbb{P}(\mu - c \leq \overline{X} < \mu + c) &\approx 2\Phi\left(\frac{\sqrt{n}}{\sigma}c\right) - 1 \\
\mathbb{P}(\mu - c \leq \overline{X} \leq \mu + c) &\approx 2\Phi\left(\frac{\sqrt{n}}{\sigma}c\right) - 1.
\end{aligned}$$

If $X_1, X_2, \ldots, X_n$ are normally distributed, then $\overline{X} = \frac{Y_n}{n}$ and all previous approximations become equalities valid for any n.

- Example. The blood cholesterol level of an individual randomly selected from a population of workers is a normal random variable X with mean $\mu = 202$ and standard deviation $\sigma = 14$.

  Suppose that a random sample of $n = 36$ or $n = 64$ workers is selected, what is the probability that the Sample Mean $\overline{X}$ of their blood cholesterol levels $X_1, X_2, \ldots, X_n$ (IID random variables distributes as X) will lie between 198 and 206?

  We have

$$\mathbb{P}\left(198 \leq \overline{X} \leq 206\right) = \mathbb{P}\left(\mu - 4 \leq \overline{X} \leq \mu + 4\right)$$

$$= 2\Phi\left(\frac{\sqrt{n}}{14} \cdot 4\right) - 1 = \begin{cases} 2\Phi\left(\frac{6}{14} \cdot 4\right) - 1 \text{ if } n = 36 \\ 2\Phi\left(\frac{8}{14} \cdot 4\right) - 1 \text{ if } n = 64 \end{cases}$$

$$= \begin{cases} 2\Phi\left(1.71\right) - 1 = 2 \cdot 0.9564 - 1 = 91.3\% \text{ if } n = 32 \\ 2\Phi\left(2.29\right) - 1 = 2 \cdot 0.9890 - 1 = 97.8\% \text{ if } n = 64 \end{cases}$$

  Observe that if the blood cholesterol level had not a normal distribution, then, since n is large, the previous estimates are still valid (but only approximately).

- Another example. An astronomer is interested in measuring the distance from the Earth to a distant star.

  However, due to differing atmospheric conditions and random errors, each time that a measurement is made, it will yield not the exact distance. The value of the measurement is a normal random variable X with mean μ the actual distance and standard deviation $\sigma = 3$ light-years.

  Thus, the astronomer plans a series of $n = 10$ measurements $X_1, X_2, \ldots, X_2$ (IID random variables distributed as X) and use their Sample Mean $\overline{X}$ as an estimated value for the actual distance.

  What is the probability that the estimated value will be within 0.5 light-years of the actual distance?

We have

$$\begin{aligned}\mathbb{P}\left(\mu - 0.5 \leq \overline{X} \leq \mu + 0.5\right) &= 2\Phi\left(\frac{\sqrt{n}}{3} \cdot 0.5\right) - 1 = 2\Phi\left(0.53\right) - 1 \\ &= 2 \cdot 0.7019 - 1 = 40.4\%.\end{aligned}$$

Observe the probability can be computed without to know the actual distance μ.

- Exercise. Consider a Bernoulli process of length n, where n is large, with outcomes α (probability p) and β (probability q) at any trial. For any $i \in \{1, \ldots, n\}$, let X_i be the random variable with value 1 if at the i-th trial the outcome is α and value 0 otherwise. By using the normal approximation, give an estimate of the probability

$$\mathbb{P}\left(\left|\frac{\overline{X}}{p} - 1\right| < \varepsilon\right),$$

where $\varepsilon > 0$. For the case where a regular coin is flipped, find how many flips are needed to have

$$\mathbb{P}\left(\left|\frac{\overline{X}}{\frac{1}{2}} - 1\right| < 0.1\right) \geq 95\%$$

and compare with the result previously found by the Chebyshev's inequality.

Law of Large Numbers

- Now, assume to have an infinite sequence $X_1, X_2, X_3, \ldots$ of IID random variables with common mean μ and common standard deviation σ. For any $n \in \{1, 2, 3, \ldots\}$, we denote by $\overline{X}_n$ the Sample Mean of $X_1, X_2, \ldots, X_n$.

  Since $\overline{X}_n$ has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, by the Chebyshev inequality, we have

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| < k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2} \text{ for any } k > 0.$$

  So, given $\varepsilon > 0$, with

$$k = \frac{\varepsilon}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}\varepsilon}{\sigma}$$

  we have

$$\mathbb{P}\left(\left|\overline{X}_n - \mu\right| < \varepsilon\right) \geq 1 - \frac{1}{\frac{n\varepsilon^2}{\sigma^2}} = 1 - \frac{\sigma^2}{n\varepsilon^2}.$$

By passing both sides to the limit as $n \to \infty$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| < \varepsilon\right) \geq 1$$

and so

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| < \varepsilon\right) = 1.$$

This fact is called the **Weak Law of Large Numbers**: for any given $\varepsilon > 0$ arbitrarily small, the probability to find the Sample Mean at a distance from μ smaller than ε tends to 1, as n tends to infinity.

- Indeed there is also the **Strong Law of Large Number**: we have

$$\mathbb{P}\left(\lim_{n\to\infty} \overline{X}_n = \mu\right) = 1.$$

This means that it is sure that the Sample Mean tends to μ, as n tends to infinity.

The proof of this result is outside the scope of the course. However, we remark an its important consequence.

- Consider a basic experiment with sample space Ω and a measure of probability $\mathbb{P}$.

Now, consider the super-experiment given by a sequence of independent repetitions of the basic experiment. The outcome of this super experiment is the sequence $\omega = (\omega_1, \omega_2, \omega_3, \ldots)$, where, for any $i \in \{1, 2, 3, \ldots\}$, ω_i is the outcome of the $i-$th repetition of the experiment. The sample space of the super experiment is

$$\Omega^\infty := \Omega \times \Omega \times \Omega \times \cdots$$

We assume that the measure of probability $\mathbb{P}^\infty$ for the super experiment is such that, for any event F relevant to the basic experiment and for any $i \in \{1, 2, 3, \ldots\}$, the event $\omega_i \in F$ relevant to the super experiment has probability

$$\mathbb{P}^\infty(\omega_i \in F) = \mathbb{P}(F).$$

This is the mathematical translation of the fact that the super experiment is a sequence of repetitions of the basic experiment.

Moreover, we assume that, for any event F relevant to the basic experiment, the random variables $X_1, X_2, X_3, \ldots : \Omega^\infty \to \mathbb{R}$ given by

$$X_i(\omega) = \begin{cases} 1 \text{ if } \omega_i \in F \\ 0 \text{ otherwise} \end{cases}, \ \omega \in \Omega^\infty \text{ and } i \in \{1, 2, 3, \ldots\},$$

are independent.

This is the mathematical translation of the fact that in the super experiment the repetitions are independent.

Observe that $X_1, X_2, X_3, \ldots$ are identically distributed. They have the same pmf: value 1 with probability $\mathbb{P}(F)$ and value 0 with probability $1 - \mathbb{P}(F)$. The common mean is

$$\mu = 1 \cdot \mathbb{P}(F) + 0 \cdot (1 - \mathbb{P}(F)) = \mathbb{P}(F).$$

Now, let F be a given event relevant to the basic experiment. Since the above defined random variables $X_1, X_2, X_3, \ldots$ corresponding to the event F are IID, the Strong Law of the Large Numbers says that

$$\mathbb{P}^{\infty}\left(\lim_{n\to\infty} \overline{X}_n = \mathbb{P}(F)\right) = 1,$$

i.e. it is sure that

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \text{frequency of the event } F \text{ over } n \text{ repetitions}$$

tends to $\mathbb{P}(F)$, as n tends to infinity.

So, for example, when a regular coin is flipped an infinite number of times, it is sure that the frequency of Heads tends to $\frac{1}{2}$ and when a regular die is rolled an infinite number of times, it is sure that the frequency of any score from 1 to 6 tends to $\frac{1}{6}$.

This is exactly the frequentist interpretation of the probability, but now it is not an interpretation where we have to assume the existence of the limit of the Long Term Relative Frequency of the events: now, it is a true fact, even if we adopt the bayesian interpretation.

- In the common language, one often invokes the Law of Large Numbers for saying that an event of very small nonzero probability surely happens if a large number of (independent) repetitions of the experiment are accomplished.

  Exercise. Prove that this is a consequence of $\mathbb{P}^{\infty}\left(\lim_{n\to\infty} \overline{X}_n = \mathbb{P}(F)\right) = 1$.

So, although the probability of winning the Superenalotto is quite small for a given individual, it is sure that someone will win if a sufficiently large number of individuals will play the Superenalotto.

Exercise. Suppose that Medicine will find the way for living without aging and diseases (this the goal of Medicine). Explain why it is sure that we will die anyway.

Exercise. A guy plays every week the lottery game. After one year without winning he says: "next year I will win, the Law of Large Numbers says this". Explain why he is saying a wrong thing.