3 Correlation and Regression

So far we have been discussing a number of descriptive techniques for describing one variable only. However, a very important part of Statistics is describing the association between two (or more) variables, whether or not they are independent, and if they are not, what is the nature of their dependence. One of the most fundamental concepts in statistical research is the concept of correlation.

Correlation is a measure of the relationship between one dependent variable, called the *response* variable and one or more independent variables, called *predictor* variables (or, simply, predictors). If two variables are correlated, this means that one can use information about one variable to predict the values of the other variable. **Regression** is then the method or statistical procedure that is used to establish that relationship.

3.1 Correlation, Curves of Regression

We will restrict our discussion to the case of two characteristics, X and Y. If X and Y have the same length, we can get a first idea of the relationship between the two, by plotting them in a **scattergram**, or **scatterplot**, , which is a plot of the points with coordinates $(x_i, y_i)_{i=\overline{1,k}}$, $x_i \in X$, $y_i \in Y$, $i = \overline{1,k}$. We group the N primary data into mn classes and denote by (x_i, y_j) the class mark and by f_{ij} the absolute frequency of the class (i, j), $i = \overline{1,m}$, $j = \overline{1,n}$. Then we represent the two-dimensional characteristic (X, Y) in a *correlation table*, or *contingency table*, as shown in Table 1.

$X \setminus Y$	y_1	$\dots$	y_j	$\dots$	y_n	
x_1	f_{11}	$\dots$	f_{1j}	$\dots$	f_{1n}	$f_{1.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	f_{i1}	$\dots$	f_{ij}	$\dots$	f_{in}	$f_{i.}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_m	f_{m1}	$\dots$	f_{mj}	$\dots$	f_{mn}	$f_{m.}$
	$f_{.1}$	$\dots$	$f_{.j}$	$\dots$	$f_{.n}$	$f_{..} = N$

Table 1: Correlation Table

Notice that

$$\sum_{j=1}^{n} f_{ij} = f_{i.}, \ \sum_{i=1}^{m} f_{ij} = f_{.j}, \ \sum_{i=1}^{m} f_{i.} = \sum_{j=1}^{n} f_{.j} = f_{..} = N.$$

Now we can define numerical characteristics associated with (X, Y).

Definition 3.1. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 1 and let $k_1, k_2 \in \mathbb{N}$.*

(1) *The **(initial) moment of order (k_1, k_2)** of (X, Y) is the value*

$$\overline{\nu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} x_i^{k_1} y_j^{k_2}. \tag{3.1}$$

(2) *The **central moment of order (k_1, k_2)** of (X, Y) is the value*

$$\overline{\mu}_{k_1 k_2} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} (x_i - \overline{x})^{k_1} (y_j - \overline{y})^{k_2}, \tag{3.2}$$

where $\overline{x} = \overline{\nu}_{10} = \frac{1}{N} \sum_{i=1}^{m} f_{i.} x_i$ and $\overline{y} = \overline{\nu}_{01} = \frac{1}{N} \sum_{j=1}^{n} f_{.j} y_j$ are the means of X and Y, respectively.

Remark 3.2. Just as the means of the two characteristics X and Y can be expressed as moments of (X, Y), so can their variances:

$$\begin{aligned} \overline{\sigma}_X^2 &= \overline{\mu}_{20} = \overline{\nu}_{20} - \overline{\nu}_{10}^2, \\ \overline{\sigma}_Y^2 &= \overline{\mu}_{02} = \overline{\nu}_{02} - \overline{\nu}_{01}^2. \end{aligned}$$

Definition 3.3. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 1.*

(1) *The **covariance*** ($\boxed{\text{cov}}$) *of (X, Y) is the value*

$$\text{cov}(X, Y) = \overline{\mu}_{11} = \frac{1}{N} \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}(x_i - \overline{x})(y_j - \overline{y}). \tag{3.3}$$

(2) *The **correlation coefficient*** ($\boxed{\text{corrcoef}}$) *of (X, Y) is the value*

$$\overline{\rho} = \overline{\rho}_{XY} = \frac{\text{cov}(X, Y)}{\sqrt{\overline{\mu}_{20}}\sqrt{\overline{\mu}_{02}}} = \frac{\overline{\mu}_{11}}{\overline{\sigma}_X \overline{\sigma}_Y}. \tag{3.4}$$

These two notions have been mentioned before, for two random variables. They are defined similarly for sets of data and they have the same properties. The covariance gives a rough idea of the relationship between X and Y. As before, if X and Y are independent (so there is no relationship, no correlation between them), then the covariance is 0. If large values of X are associated with large values of Y, then the covariance will have a positive value, if, on the contrary, large values of X are associated with small values of Y, then the covariance will have a negative value. Also, an easier computational formula for the covariance is $\text{cov}(X, Y) = \overline{\nu}_{11} - \overline{x} \cdot \overline{y}$.

The correlation coefficient is then

$$\overline{\rho} = \frac{\overline{\nu}_{11} - \overline{x} \cdot \overline{y}}{\overline{\sigma}_X \overline{\sigma}_Y}$$

and, as before, it satisfies the inequality

$$-1 \leq \overline{\rho} \leq 1 \tag{3.5}$$

and, by its variation between -1 and 1, its value measures the linear relationship between X and Y. If $\overline{\rho}_{XY} = 1$, there is a *perfect positive correlation* between X and Y, if $\overline{\rho}_{XY} = -1$, there is a *perfect negative correlation* between X and Y. In both cases, the linearity is "perfect", i.e there exist $a, b \in \mathbb{R},\ a \neq 0$, such that $Y = aX + b$. If $\overline{\rho}_{XY} = 0$, then there is no linear correlation between X and Y, they are said to be *(linearly) uncorrelated*. However, in this case, they may not be independent, some other type of relationship (not linear) may exist between them.

In our task of finding a relationship between X and Y, we may go the following path: knowing the value of one of the characteristics, try to find a probable, an "expected" value for the other. If the two characteristics are related in any way, then there should be a pattern developing, that is the expected

value of one of them, *conditioned* by the other one taking a certain value, should be a function of that value that the other variable assumes. In other words, we should consider *conditional means*, defined similarly to regular means, only taking into account the condition.

Definition 3.4. *Let (X, Y) be a two-dimensional characteristic whose distribution is given by Table 1.*

(1) *The **conditional mean** of Y, given $X = x_i$, is the value*

$$\overline{y}_i = \overline{y}(x_i) = \frac{1}{f_{i\cdot}} \sum_{j=1}^{n} f_{ij} y_j,\ i = \overline{1, m}. \tag{3.6}$$

(2) *The **conditional mean** of X, given $Y = y_j$, is the value*

$$\overline{x}_j = \overline{x}(y_j) = \frac{1}{f_{\cdot j}} \sum_{i=1}^{m} f_{ij} x_i,\ j = \overline{1, n}. \tag{3.7}$$

Definition 3.5. *Let (X, Y) be a two-dimensional characteristic.*

(1) *The curve $y = f(x)$ formed by the points with coordinates $(x_i, \overline{y}_i)$, $i = \overline{1, m}$, is called the **curve of regression** of Y on X.*

(2) *The curve $x = g(y)$ formed by the points with coordinates $(y_j, \overline{x}_j)$, $j = \overline{1, n}$, is called the **curve of regression** of X on Y.*

Remark 3.6. The curve of regression of a characteristic Y with respect to another characteristic X is then the mean value of Y, $\overline{y}(x)$, given $X = x$. The curve of regression is determined so that it approximates best the scatterplot of (X, Y).

3.2 Least Squares Estimation, Linear Regression

One of the most popular ways of finding curves of regression is the *least squares method*.
Assume the curve of regression of Y on X is of the form

$$y = y(x) = f(x; a_1, \ldots, a_s).$$

We determine the unknown parameters $a_1, \ldots, a_s$ so that the *sum of squares error* (SSE) (the sum of the squares of the differences between the responses y_j and their fitted values $y(x_i)$, each counted

with the corresponding frequency)

$$S = SSE = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - y(x_i)\Big)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - f(x_i; a_1, \ldots, a_s)\Big)^2$$

is minimum (hence, the name of the method).
We find the point of minimum $(\overline{a}_1, \ldots, \overline{a}_s)$ of S by solving the system

$$\frac{\partial S}{\partial a_k} = 0,\ k = \overline{1, s},$$

i.e.

$$-2\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - f(x_i; a_1, \ldots, a_s)\Big)\frac{\partial f(x_i; a_1, \ldots, a_s)}{\partial a_k} = 0, \tag{3.8}$$

for every $k = \overline{1, s}$.
Then the equation of the curve of regression of Y on X is

$$y = f\left(x; \overline{a}_1, \ldots, \overline{a}_s\right).$$

Linear regression

Let us consider the case of *linear regression* and find the equation of the *line of regression* of Y on X.

We are finding a curve

$$y = ax + b,$$

for which

$$S(a, b) = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - ax_i - b\Big)^2$$

is minimum. We have to solve the 2×2 system

$$\begin{aligned} \frac{\partial S(a, b)}{\partial a} &= 0 \\ \frac{\partial S(a, b)}{\partial b} &= 0, \end{aligned}$$

i.e.

$$\begin{aligned} -2\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - ax_i - b\Big)x_i &= 0 \\ -2\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\Big(y_j - ax_i - b\Big) &= 0, \end{aligned}$$

which becomes

$$\begin{cases} \left(\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_i^2\right)a + \left(\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_i\right)b = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_iy_j \\ \left(\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}x_i\right)a + \left(\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\right)b = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}y_j \end{cases}$$

and after dividing both equations by N,

$$\begin{cases} \overline{\nu}_{20}a + \overline{\nu}_{10}b = \overline{\nu}_{11} \\ \overline{\nu}_{10}a + \overline{\nu}_{00}b = \overline{\nu}_{01}. \end{cases}$$

Its solution is

$$\overline{a} = \frac{\overline{\nu}_{11} - \overline{\nu}_{10}\overline{\nu}_{01}}{\overline{\nu}_{20} - \overline{\nu}_{10}^2} = \frac{\overline{\nu}_{11} - \overline{x}\cdot\overline{y}}{\overline{\sigma}_X^2} = \frac{\overline{\nu}_{11} - \overline{x}\cdot\overline{y}}{\overline{\sigma}_X\overline{\sigma}_Y}\cdot\frac{\overline{\sigma}_Y}{\overline{\sigma}_X} = \overline{\rho}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X},$$

$$\overline{b} = \overline{\nu}_{01} - \overline{\nu}_{10}\overline{a} = \overline{y} - \overline{a}\cdot\overline{x}.$$

So the equation of the line of regression of Y on X is

$$y - \overline{y} = \overline{\rho}\,\frac{\overline{\sigma}_Y}{\overline{\sigma}_X}\,(x - \overline{x}) \tag{3.9}$$

and, by analogy, the equation of the line of regression of X on Y is

$$x - \overline{x} = \overline{\rho}\,\frac{\overline{\sigma}_X}{\overline{\sigma}_Y}\,(y - \overline{y})\,. \tag{3.10}$$

Remark 3.7.

1. The point of intersection of the two lines of regression, $(\overline{x}, \overline{y})$, is called the *centroid* of the distribution of the characteristic (X, Y).

2. The slope $\overline{a}_{Y|X} = \overline{\rho}\,\dfrac{\overline{\sigma}_Y}{\overline{\sigma}_X}$ of the line of regression of Y on X is called the *coefficient of regression* of Y on X. Similarly, $\overline{a}_{X|Y} = \overline{\rho}\,\dfrac{\overline{\sigma}_X}{\overline{\sigma}_Y}$ is the coefficient of regression of X on Y and

$$\overline{\rho}^2 = \overline{a}_{Y|X}\;\overline{a}_{X|Y}.$$

3. For the angle α between the two lines of regression, we have

$$\tan\alpha \;=\; \frac{1-\overline{\rho}^2}{\overline{\rho}^2}\cdot\frac{\overline{\sigma}_X\overline{\sigma}_Y}{\overline{\sigma}_X^2+\overline{\sigma}_Y^2}.$$

So, if $|\,\overline{\rho}\,| = 1$, then $\alpha = 0$, i.e. the two lines coincide. If $|\,\overline{\rho}\,| = 0$ (for instance, if X and Y are independent), then $\alpha = \dfrac{\pi}{2}$, i.e. the two lines are perpendicular.

Example 3.8. Let us examine the situations graphed in Figure 1.

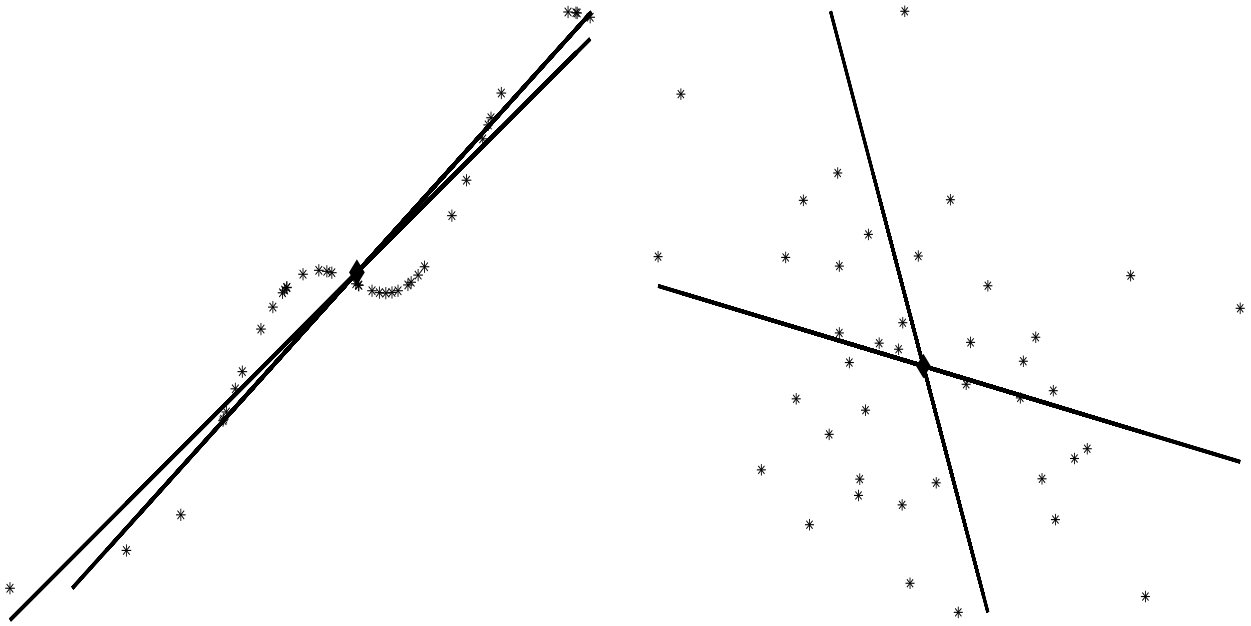
- In Figure 1(a) $\overline{\rho} = 0.95$, positive and very close to 1, suggesting a strong positive linear trend. Indeed, most of the points are on or very close to the line of regression of Y on X. The positivity indicates that large values of X are associated with large values of Y. Also, since the correlation coefficient is so close to 1, the two lines of regression almost coincide.

- In Figure 1(b) $\overline{\rho} = -0.28$, negative and fairly small, close to 0. If a relationship exists between X and Y, it does not seem to be linear. In fact, they are very close to being independent, since the points are scattered around the plane, no pattern being visible. The two lines of regression are very distinct and both have negative slopes, suggesting that large values of X are associated with small values of Y.

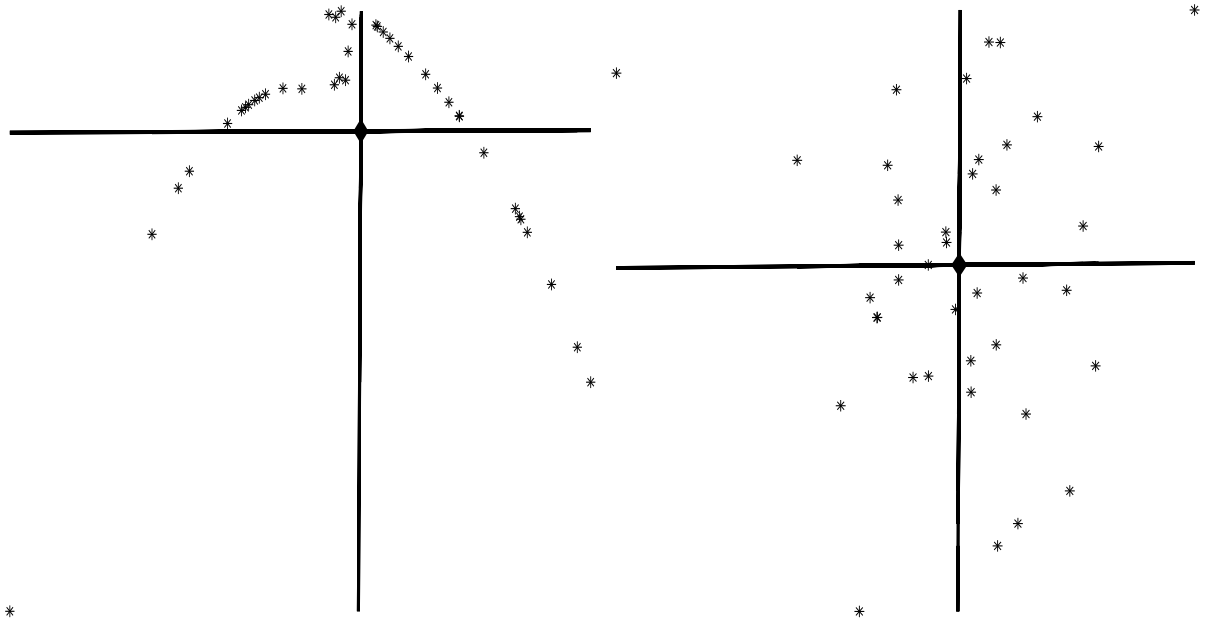
- In Figure 1(c) $\overline{\rho} = 0$, so the two characteristics are uncorrelated, no linear relationship exists between them. However they are not independent, they were chosen so that $Y = -X^2 + \sin\left(\dfrac{1}{X}\right)$. Notice also, that the two lines of regression are perpendicular.

- Finally, in Figure 1(d) $\overline{\rho} = 0$, again, so no linear relationship exists. In fact the two characteristics are independent, which is suggested by their random scatter inside the plane.

Remark 3.9. Other types of curves of regression that are fairly frequently used are
– *exponential* regression $y = ab^x$,
– *logarithmic* regression $y = a\log x + b$,
– *logistic* regression $y = \dfrac{1}{ae^{-x}+b}$,
– *hyperbolic* regression $y = \dfrac{a}{x} + b$.



(a) $\overline{\rho} = 0.95$ (b) $\overline{\rho} = -0.28$



(c) $\overline{\rho} = 0$ (d) $\overline{\rho} = 0$

Fig. 1: Scattergram, Lines of Regression and Centroid

Chapter 3. Sample Theory

In inferential Statistics, we will have the following situation: we are interested in studying a characteristic (a random variable) X, relative to a population P of (known or unknown) size N. The difficulty or even the impossibility of studying the entire population, as well as the merits of choosing and studying a random sample from which to make inferences about the population of interest, have already been discussed in the previous chapter. Now, we want to give a more rigorous and precise definition of a random sample, in the framework of random variables, one that can then employ probability theory techniques for making inferences.

1 Random Samples and Sample Functions

We choose n objects from the population and actually study $X_i,\ i = \overline{1,n}$, the characteristic of interest *for the i^{th} object selected*. Since the n objects were randomly selected, it makes sense that for $i = \overline{1,n}$, X_i is a random variable, one that has *the same* distribution (pdf) as X, the characteristic relative to the entire population. Furthermore, these random variables are independent, since the value assumed by one of them has no effect on the values assumed by the others. Once the n objects have been selected, we will have n numerical values available, $x_1, \ldots, x_n$, the observed values of $X_1, \ldots, X_n$.

Definition 1.1. *A **random sample of size n** from the distribution of X, a characteristic relative to a population P, is a collection of n independent random variables $X_1, \ldots, X_n$, having the same distribution as X. The variables $X_1, \ldots, X_n$, are called **sample variables** and their observed values $x_1, \ldots, x_n$, are called **sample data**.*

Remark 1.2. The term *random sample* may refer to the objects selected, to the sample variables, or to the sample data. It is usually clear from the context which meaning is intended. In general, we use capital letters to denote sample variables and corresponding lowercase letters for their observed values, the sample data.

We are able now to define sample functions, or statistics, in the more precise context of random variables.

Definition 1.3. *A **sample function** or **statistic** is a random variable*

$$Z_n = h_n(X_1, \ldots, X_n),$$

where $h_n : \mathbb{R}^n \to \mathbb{R}$ *is a measurable function. The value of the sample function* Z_n *is* $z_n = h_n(x_1, \ldots, x_n)$.

We will revisit now some sample numerical characteristics discussed in the previous chapter and define them as sample functions. That means they will have a pdf, a cdf, a mean value, variance, standard deviation, etc. A sample function will, in general, be an approximation for the corresponding population characteristic. In that context, the standard deviation of the sample function is usually referred to as the **standard error**.

In what follows, $\{X_1, \ldots, X_n\}$ denotes a sample of size n drawn from the distribution of some population characteristic X.

2 Sample Mean

Definition 2.1. *The* ***sample mean*** *is the sample function defined by*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.1}$$

and its value is $\overline{x}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_i$.

Now that the sample mean is defined as a random variable, we can discuss its distribution and its numerical characteristics.

Proposition 2.2. *Let* X *be a characteristic with* $E(X) = \mu$ *and* $V(X) = \sigma^2$. *Then*

$$E\left(\overline{X}\right) = \mu \ \text{ and } \ V\left(\overline{X}\right) = \frac{\sigma^2}{n}. \tag{2.2}$$

Moreover, if $X \in N(\mu, \sigma)$, *then* $\overline{X} \in N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$.

Proof. Since $X_1, \ldots, X_n$ are identically distributed, with the same distribution as X, $E(X_i) = E(X) = \mu$ and $V(X_i) = V(X) = \sigma^2$, $\forall i = \overline{1, n}$. Then, by the usual properties of expectation, we have

$$E\left(\overline{X}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\, n\mu = \mu.$$

Further, since $X_1, \ldots, X_n$ are also independent, by the properties of variance, it follows that

$$V\left(\overline{X}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(X_i) = \frac{1}{n^2}\, n\sigma^2 = \frac{\sigma^2}{n}.$$

The last part follows from the fact that $\overline{X}$ is a linear combination of independent, normally distributed random variables.

□

Corollary 2.3. *Let X be a characteristic with $E(X) = \mu$ and $V(X) = \sigma^2$ and for $n \in \mathbb{N}$ let*

$$Z_n = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

Then the variable Z_n converges in distribution to a Standard Normal variable, as $n \to \infty$, i.e. $F_{Z_n} \overset{n\to\infty}{\longrightarrow} F_Z = \Phi$.
Moreover, if $X \in N(\mu, \sigma)$, then the statement is true for every $n \in \mathbb{N}$.

Proof. This is a direct consequence of the Central Limit Theorem (CLT).

□