

Ministry of Higher Education and
Scientific Research
Babylon University
College of Education for Pure Sciences



Experimental Distribution Function

Proposed research to the Council of the College of Education for Pure
Sciences /University of Babylon as a part of the requirements for a
Bachelor's degree in mathematics

By

Jaafar Hussein Ali Mohi

Supervision By

Dr. Kareema Abed Al-Kadim

2023 AD

1444 AH

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

وَقَالَ رَبِّ زِدْنِي عِلْمًا

صدق الله العلي العظيم

سورة طه - آية (١١٤)

Dedication

My university journey has come to an end after exhaustion and hardship.

and here I am, completing my graduation research with and energy.

I am grateful to everyone who has contributed to my career.

and help me a little bit

parents, family, friends, esteemed teachers...

I dedicate to you my graduation research.....

Acknowledgments

I thank Allah first and foremost for the great grace He has given me. Thank you to my dear parents who have not stopped making all their efforts from the moment I was born to these blessed moments. Thank you to all those who advised me, guided me, contributed to my education and I thank you in particular my dear teacher,

Dr. Kareema Abed Al-Kadim

support and guidance with advice, education, correction and follow-up. I also thank the members of the distinguished debate committee, all friends and those who accompanied me during my study trip.

List of Contents

1.	Chapter One Introduction	page
1.1	Introduction.....	2
1.2	Mathematical Definition.....	3
1.3	Empirical Cumulative Distribution Function.....	3
2.	Chapter Two Experimental Distribution Function	
2.1.	Experimental Distribution.....	7
2.2.	Empirical Distribution Function.....	9
2.3.	Graphical Representation of Empirical Distributions.....	12
	Examples.....	14
	References.....	.17

List of Figures

No		page
1.1.	Empirical cumulative distribution function of student grades for a hypothetical students class of 50.....	.4
1.2.	Distribution of students' scores in descending cdf form.....	4
1.3.	cdf for student grades.....	.5
2.1.	The graphic representation of the empirical distribution function	10
2.2.	Graphic representation of F_x11
2.3.	Relative frequency polygon.....	13

List of Tables

No		page
1.	The experimental distribution is shown tabularly ξ ,.....	.8
2.	Observed distribution of the random variable ξ	8
3.	Experimental distribution for ξ ,.....	10
4.	Results study table.....	12

Abstract

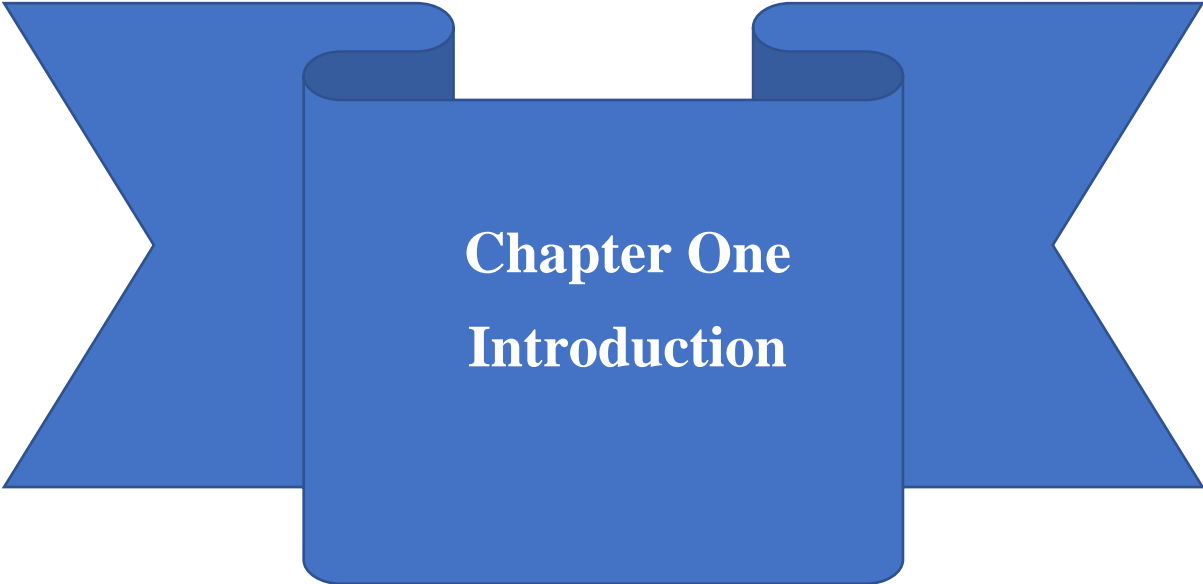
data points. Its value at any specified value of the measured variable is the fraction of observations of the measured variable that are less than or equal to the specified value.

Empirical distribution function

In statistics, an empirical distribution function (commonly also called an empirical Cumulative Distribution Function, eCDF) is the distribution function associated with the empirical measure of a sample. This cumulative distribution function is a step function that jumps up by $1/n$ at each of the n data points.

Its value at any specified value of the measured variable is the fraction of observations of the measured variable that are less than or equal to the specified value. The empirical distribution function is an estimate of the cumulative distribution function that generated the points in the sample. It converges with probability 1 to that underlying distribution, according to the Glivenko-Cantelli theorem.

The first chapter talks about the introduction and definition of the empirical distribution function, and the Empirical Cumulative Distribution Function. The second chapter talks about experimental distribution, Graphical Representation of Empirical Distributions, Example.



Chapter One
Introduction

1.1. Introduction

It is a specific mathematical function that links the values of a random variable with the probabilities of occurrence of these values. Method, evaluation and review of the program

In statistics, an empirical distribution function (commonly also called an empirical cumulative distribution function, (CDF) Is the distribution function associated with the empirical measure of a sample. This cumulative distribution function is a step function that jumps up by $1/n$ at each of the n data points. its value at any specified value of the measured variable is the fraction of observations of the measured variable that are less than or equal to the specified value .

The cumulative distribution function , or partition function in statistics and probability theory, is a function that determines what is the probability that the value of a random variable (x) is less than or equal to a certain value (d) Or in other words, it is a function that gives the probability distribution of a random variable, provided that its value is a real number . The cumulative distribution function should not be confused with the probability density function or the probability mass function of discrete random variables.

The ECDF is an unbiased and maximum likelihood estimator of the theoretical CDF of the process generating the data, so it plays an important role in many statistical methods.

In this paper, we tried to shed light on the concept and importance of the empirical distribution function.

In this research, we tried to shed light on the concept and importance of the empirical distribution function. The research included some definitions and examples on this topic.

1.2. Mathematical Definition[2]

The probabilistic space (Ω, F, P) where

Ω : represents a non-empty set, called the sample space.

F is the σ -algebra of a sample space in which we call each of its components:(event)

P : it is a probabilistic measure ,The true random variable is defined as follows: $X: \Omega \rightarrow R$ And the cumulative distribution

function is defined as follows : $F_X: R \rightarrow R$

$F_X(x) = P(X \leq x)$ that is for $x \in R$

Note :It is also possible to determine the probability that

the value of a random variable is greater than a and

b less than or equal to[2].

$p(a < x \leq b) = p(x \leq b) - p(x \leq a) = F(b) - F(a) \dots 1$

1.3. Empirical Cumulative Distribution Function[3]

Experimental: means that we are interested in observations, not theory.

Cumulative :we used the cumulative binomial and normal distributions, respectively, to calculate probabilities and visualize the distribution.

Distribution :However, in real life, the data we collect or monitor does not come from a theoretical distribution.

Function: We can do this using the ecdf function. CDF stands for "Empirical Cumulative Distribution Function".

CDF is an estimator for the cumulative distribution function. CDF basically allows you to plot a feature of your data in order from least to greatest and see the entire feature as if it were distributed across the dataset

To illustrate cumulative empirical distribution functions, It was started with a hypothetical example modeled closely on something like a data set of student grades. Suppose our hypothetical class has 50 students, and

the students have just completed an exam In which they can score between 0 and 100 points. How can it better visualize class performance, eg to define appropriate class boundaries ? If can plot the total number of students who got the most over a given number of points against all possible points.[3]

This plot will be an ascending function, starting at 0 for 0 points and ending at 50 for 100 points. A different way to think about this visualization Is as follows: we can arrange all the students by the number of points they got, In ascending order (so that the student with the lowest number of points gets the lowest rank and the student with the highest number of points Is the highest), and then draw ranking vs. actual points obtained. The result Is the empirical cumulative distribution function (cdf) or simply the cumulative distribution. [3]

Each point represents one student, and the lines depict the highest ranking student observed for any possible point value (Fig1)

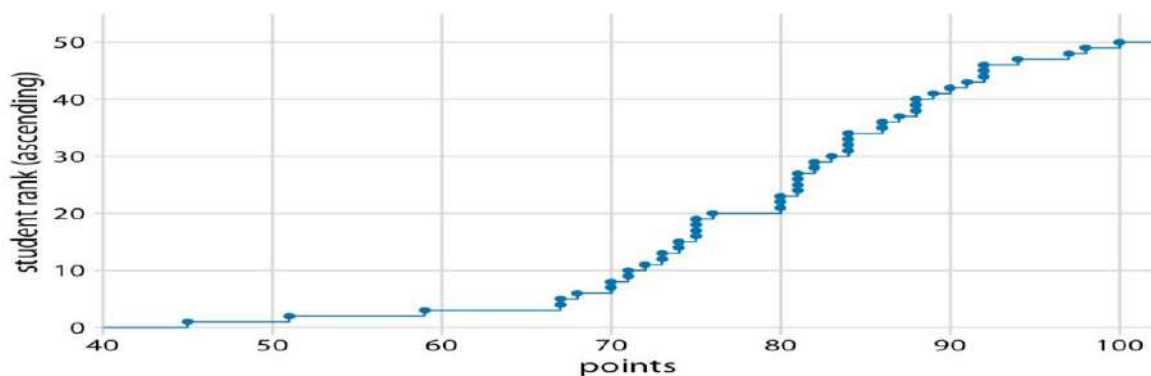
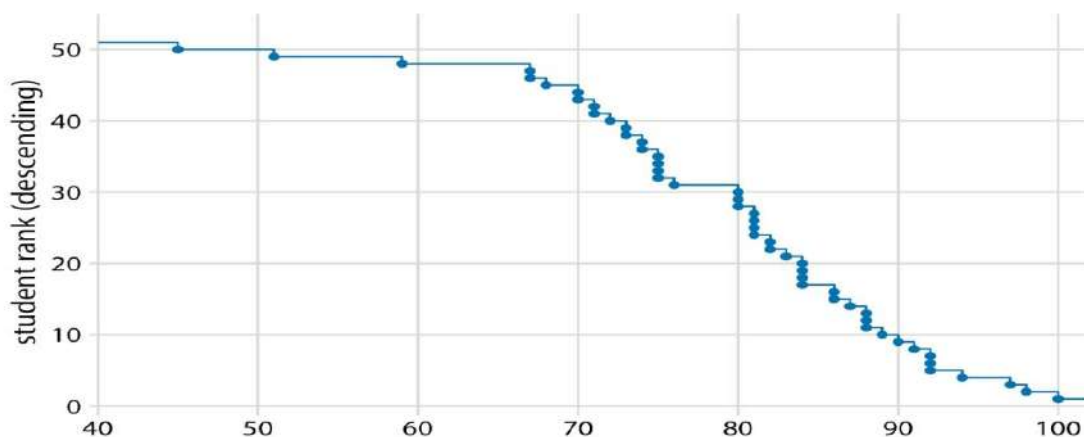


Figure 1. 1. Empirical cumulative distribution function of student grades for a hypothetical class of 50 students.

You might be wondering what happens if we rank the students In the opposite direction in descending order. This arrangement turns the job on its head. The score is still a function of the empirical cumulative distribution, but the lines now represent the lowest student rank observed for any possible point value (Fig.2)



Ascending cumulative distribution functions are more widely known and more commonly used than descending functions, but both have important applications. Downward cumulative distribution functions are important when we want to visualize highly skewed distributions

In practical applications, it is very common to plot the ecdf without highlighting the individual points and normalize the ranks in maximal order, with the y-axis representing the cumulative frequency (Fig.3).

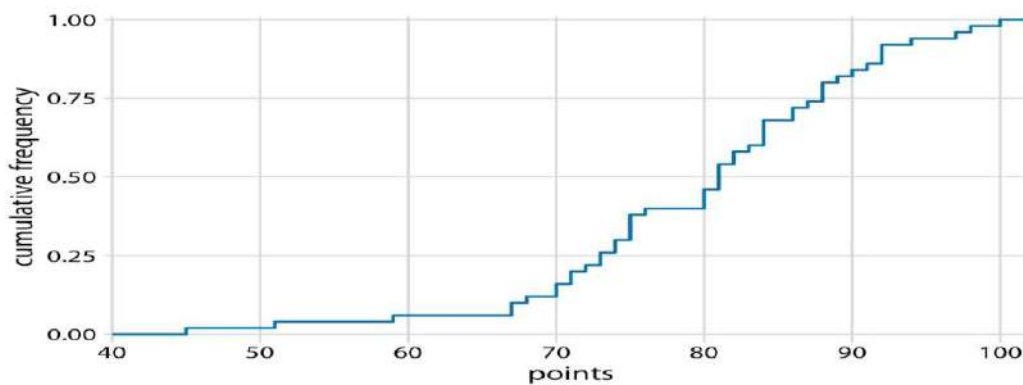
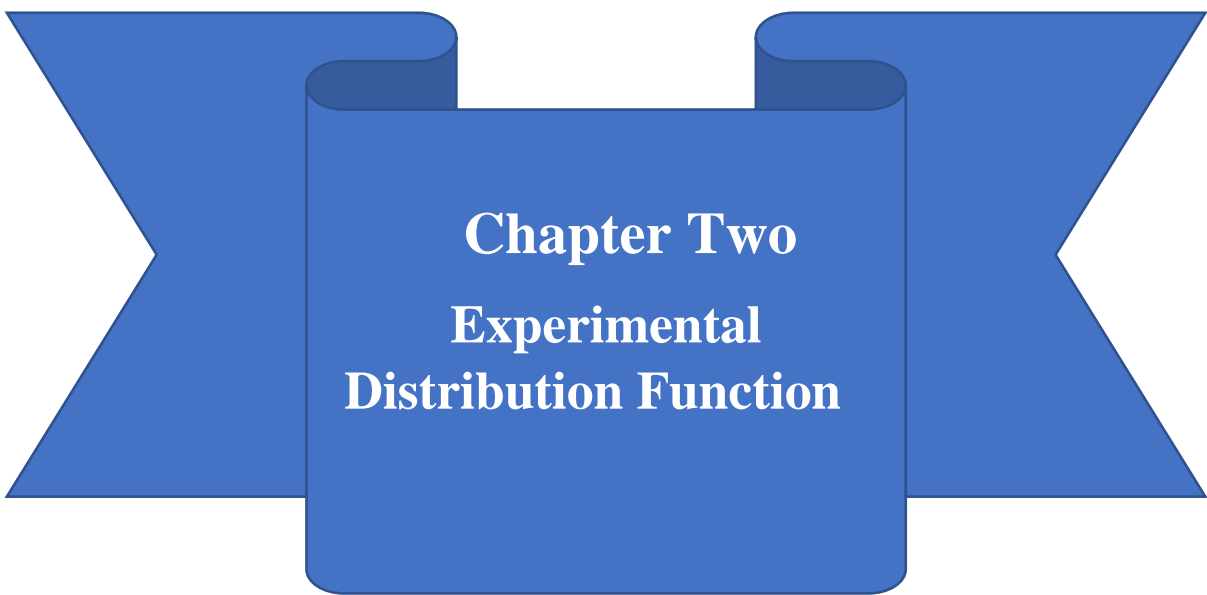


Figure 1. 3(cdf for student grades.)

The student ranks were normalized to the total number of students, such that the plotted y-values correspond to the fraction of students in the class with these many points the most.

We can directly read the main characteristics of the distribution of students' scores from this plot. For example, almost a quarter of students (25%) scored less than 75 points. The average point value (corresponding to a cumulative repeatability of 0.5) is 81. approximately 20% of students scored 90 points or more .



Chapter Two
Experimental
Distribution Function

2.1. Experimental Distribution[4]

The issue of inferring the probability distribution of the observed random variable ξ , or some of its characteristics, in the case ξ , of being unknown, requires in several cases tabulating the statistical data provided by a random sample observed[4]

$x = (x_1, \dots, x_n)$ of this population.

If the random variable ξ , is of the discrete type and the sample size n is small, then we take the different values in the observed sample x and arrange them ascending y , and then assign the corresponding relative frequencies, for example, if the observed values are

$x_i ; i = 1, \dots, n$ 1

include $k \leq n$ different value and denote it

$z_j ; j = 1, \dots, k$ 2

Where $z_1 < z_2 < \dots < z_k$ 3

Where $n_j/n ; j = 1, \dots, k$ 4

has corresponding relative frequencies, so it is called the binary class

$\{(z_j, n_j/n) ; j = 1, \dots, k\}$ 5

that is, the relative frequency distribution of the data of the observed sample, according to the experimental distribution of the observed random variable ξ ,

corresponding to the *sample* $x = (x_1, \dots, x_n)$ In order to distinguish , we sometimes use the term theoretical

$$\sum_{j=1}^k \frac{n_j}{n} = 1 \dots\dots\dots 6$$

distribution of ξ , to denote the probability distribution of it

(the distribution of the community). The experimental distribution is shown tabularly ξ , as the following Table (1)

Table 1 The experimental distribution is shown tabularly ξ ,

different note values z_j	z_1	z_2	z_k
Relative frequencies agree n_j/n	n_1/n	n_2/n	n_k/n

If we are studying a continuous random variable ξ , or if it is discontinuous but the sample size n is large, tabulating the statistical data requires dividing

$r = x(n) - x_{(1)}$ into K from non-intersecting subcategories, and the calculations are greatly simplified if the classes are equidistant $1 \approx r/k$, and this is what we will always adopt, then the relative frequencies are computed corresponding to those classes n_j/n , and here we indicate that the number of classes is k . Optional and usually not less than 5 and not more than 20, and this choice depends. About the nature of the phenomenon or the particular studied, so that it can be considered that the elements of the class are homogeneous for the studied trait ξ . As a result we get empirical.[4]

The observed distribution of the random variable ξ , as shown in Table (2).

Table 2 The observed distribution of the random variable ξ

Categories (c_{j-1}, c_j)	(c_0, c_1)	(c_1, c_2)	(c_{k-1}, c_k)
Relative Frequency(n_j/n)	n_1/n	n_2/n	n_k/n

where (c_{j-1}, c_j) is the category number

$$\sum_1^k \frac{n_j}{n} = 1, = 1, 2, \dots, k \dots \dots \dots 1$$

The relative frequency n_j/n of the event $B_j = (c_{j-1}, c_j)$ is randomly variable (because it changes in general from one observed sample to another) with an expected value equal to the probability of this event, that is:

$$E\left(\frac{n_j}{n}\right) = p(c_{j-1} < \xi, \leq c_j) = p(B_j) = p_j \dots \dots \dots 2$$

Because the random variable $Y = n_j$ follows the binomial distribution $B(n, p_j)$ as we know from the law of large numbers [Bernoulli's theorem] that if the imposed

experiment is repeated n times under the same conditions, the relative frequency of the occurrence of the event B_j ends up with the probability of this event,

That is when $n \rightarrow \infty$ that is, that

$$\lim_{n \rightarrow \infty} \frac{n_j}{n} = p_j \dots\dots\dots 3$$

This means, by observing tables (1) and (2), that the second line in both contains the estimated values for the probabilities

$$p_j ; j = 1, 2, \dots, k \dots\dots 4$$

$$\frac{n_j}{n} \approx p_j ; j = 1, \dots, k \dots\dots 5$$

2.2. Empirical Distribution Function [4]

Suppose for any real number x the variable z_i is defined as follows:

$$Z_i = \begin{cases} 1 & ; x_i < x \\ 0 & ; x_i \geq x \end{cases} , i = 1, 2, \dots, n \dots\dots\dots 1$$

Accordingly:

$$Z = \sum_{i=1}^n Z_i \dots\dots\dots 2$$

A random variable representing the number of observed sample elements $x = (x_1, \dots, x_n)$ younger than x .

If we denote by $F_n^*(x) = z/n$ then the function $F_n^*(x)$ is called the experimental distribution function ξ , (experimental distribution function) corresponding to the observed sample ξ , for discrimination, we call the distribution function $F(x)$ of the random variable ξ , by the theoretical distribution function (population distribution). In several cases where [4] there is no confusion. We drop the proof, n , that is, we denote $F^*(x)$ for the empirical distribution function of ξ

Example

If the experimental distribution of the random variable ξ , is given in Table (3). So find a function is the experimental distribution $F^*(x)$.

Table (3) Experimental distribution for ξ ,

observation values for ξ ,	2	3	5
relative frequency n_i/n	0.75	0.20	0.05

From the definition of the empirical distribution function $F_n^*(x) = Z/n$

$$\text{we find that: } F_n^*(x) = \begin{cases} 0 & ; x \leq 2 \\ 0.75 & ; 2 < x \leq 3 \\ 0.95 & ; 3 < x \leq 5 \\ 1 & ; x > 5 \end{cases}$$

The graphic representation of the empirical distribution function $F_n^*(x)$ is shown

by Figure (1).

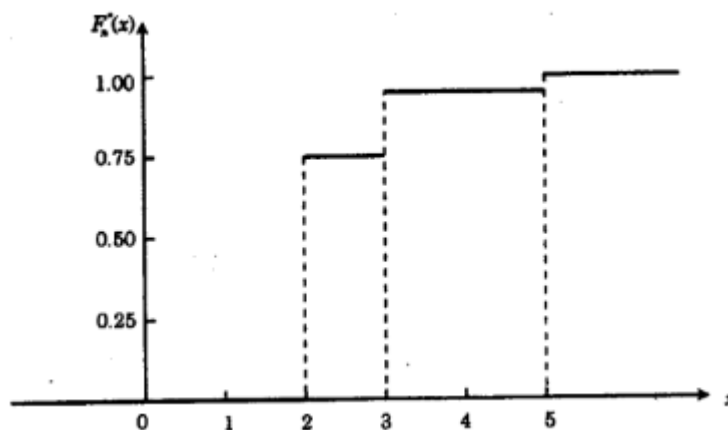


Figure (2.1) The graphic representation of the empirical distribution function

If the elements of the vector x with are different then the empirical distribution function $F_n^*(x)$ is given as follows:

$$F'_n(x) = \begin{cases} 0 & ; x \leq x_{(1)} \\ \frac{k}{n} & ; x_{(k)} < x \leq x_{(k+1)} \\ 1 & ; x > x_{(n)} \end{cases}, k = 1, 2, \dots, n-1 \dots\dots 1$$

In this case, the values of the hops are equal and equal to $1/n$. and this is The graphical representation of $F_n * (x)$ takes the form Figure (2).

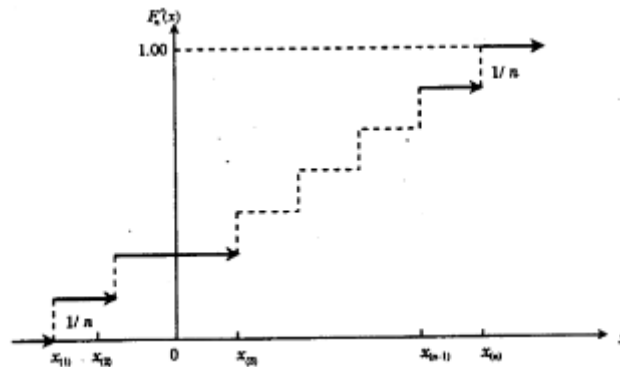


Figure (2.2) Graphic representation of F_x

In the general case, the empirical distribution function $F_n * (x)$ can be written as follows:

$$F'_n(x) = \frac{1}{n} \sum_{k=1}^n e(x - x_{(k)}) \dots\dots 2$$

Where $e(x)$ is nothing but the unitary function of the jumps (Khovsid function):

$$e(x) = \begin{cases} 0 & ; x \leq 0 \\ 1 & ; x > 0 \end{cases} \dots\dots 3$$

In the relationship the correlation of $F_n * (x)$ with the sample x appears clearly . The experimental distribution function has an important and fundamental role in mathematical statistics, and its special importance lies in the fact that with the increase in the number of observations of the random variable ξ , , the experimental distribution function $F_n * (x)$ approaches the theoretical distribution function $F(x)$

2. 3 Graphical Representation of Empirical Distributions[4]

From the foregoing, we note that the experimental distribution is a useful method for representing statistical data (observed sample data), which enables us to infer some results about the probability distribution of the observed random variable ξ , when it is fully or partially unknown. There are other ways to represent the statistical data, and one of these ways is to build a histogram and polygon for the empirical distribution of ξ , . The experimental distribution polygon is used when the observed random variable ξ , is continuous or intermittent.

As for the experimental distribution histogram, its use is limited to the case that the observed random variable is ξ , continuous, or the random sample X is of large size (the statistical data are classified). Let's show how to construct a histogram and polygon of an empirical distribution for a random variable ξ , by

next example.[4]

Example

Let us have the results of a study of the durability of 200 models of reinforced concrete, as shown in the following table(4):

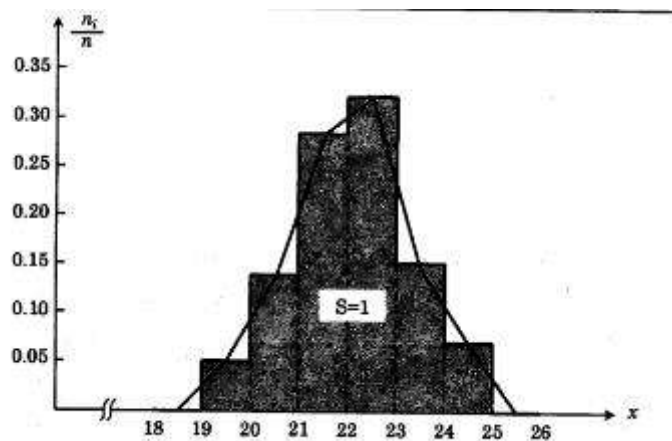
Table (4)Results study table

relative frequency ni/n	Durability field (MPa)
0.05	20-19
0.13	21-20
0.28	22-21
0.32	23-22
0.15	24-23
0.07	25-24

It is required to construct a histogram and polygon of the given experimental distribution. To construct the histogram for this experimental distribution, we draw two perpendicular axes, the horizontal

axis representing the partial periods and the vertical axis the relative repetition.

Then we determine the partial intervals of the observed values and the relative repetition corresponding to the horizontal and vertical transformer, respectively, and then we establish rectangles on those partial intervals with heights equal to the corresponding relative repetition, so we get the result on the histogram shown .to draw the polygon of the experimental distribution, we determine the midpoints of the upper sides of the rectangles shown in Figure (3), then we connect them, respectively, with straight lines, so we get the required. To close the polygon, we add a category before the first category and another after the last category, so that the relative frequency of each of them is zero, so we get the polygon shown in Figure (3).



Figure(2. 3)Relative frequency polygon

The histogram and polygon of the experimental distribution are used to infer the type of distribution model for ξ . For example, the model shown in Figure (3) reminds us of the general normal model, so it can be assumed that the type of distribution of the observed random variable is normal ξ .

Thus, the graph of the experimental distribution of a random variable ξ can be considered as a statistical counterpart to the form of its theoretical distribution.[4]

Examples [5]

Let X = amount of time (in minutes) a postal clerk spends with his or her customer. The time is known to have an exponential distribution with the average amount of time equal to **four minutes**. X is a continuous random variable since time is measured. It is given that $\mu = 4$ minutes. To do any calculations, you must know m , the decay parameter.

$$m = \frac{1}{\mu} \text{ therefore}$$

$$m = \frac{1}{4} = 0.25$$

The standard deviation, σ , is the same as the mean. $\mu = \sigma$.

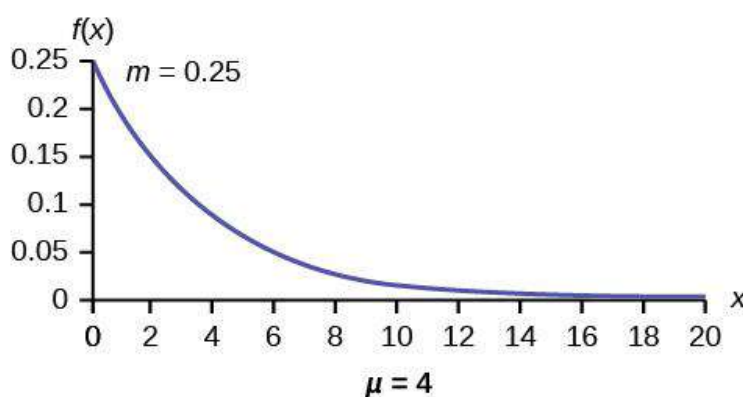
The distribution notation is $X \sim \text{Exp}(m)$. Therefore, $X \sim \text{Exp}(0.25)$.

The probability density function is $f(x) = me^{-mx}$. The number $e = 2.71828182846\dots$ It is a number that is used often in mathematics. Scientific calculators have the key " e^x ." If you enter one for x , the calculator will display the value e .

The curve is:

$$f(x) = 0.25e^{-0.25x} \text{ where } x \text{ is at least zero and } m = 0.25.$$

For example, $f(5) = 0.25e^{-(0.25)(5)} = 0.072$. The postal clerk spends **five** minutes with the customers. The graph is as follows:



Notice the graph is a declining curve. When $x = 0$,

$f(x) = 0.25e^{(-0.25)(0)} = (0.25)(1) = 0.25 = m$. The maximum value on the y - axis is m .

Using the information in example 1, find the probability that a clerk spends four to five minutes with a randomly selected customer.

The curve is: $X \sim \text{Exp}(0.125)$; $f(x) = 0.125e^{-0.125x}$

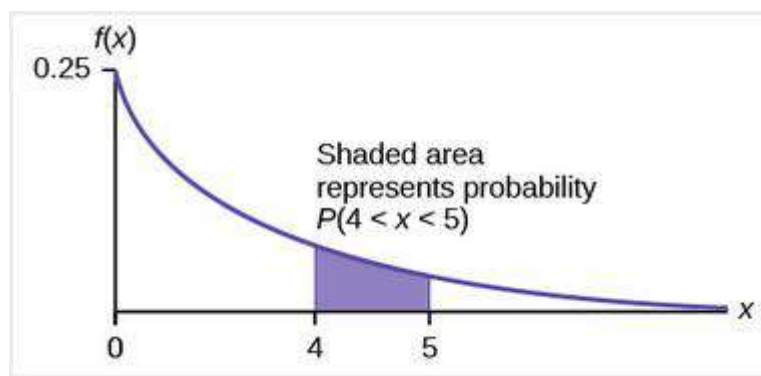
a) Find $P(4 < x < 5)$.

Solution:

The cumulative distribution function (CDF) gives the area to the left.

$$P(x < x) = 1 - e^{-mx}$$

$$P(x < 5) = 1 - e^{(-0.25)(5)} = 0.7135 \text{ and } P(x < 4) \\ = 1 - e^{(-0.25)(4)} = 0.6321$$



You can do these calculations easily on a calculator.

The probability that a postal clerk spends four to five minutes with a randomly selected customer is

$$P(4 < x < 5) = P(x < 5) - P(x < 4) = 0.7135 - 0.6321 \\ = 0.0814.$$

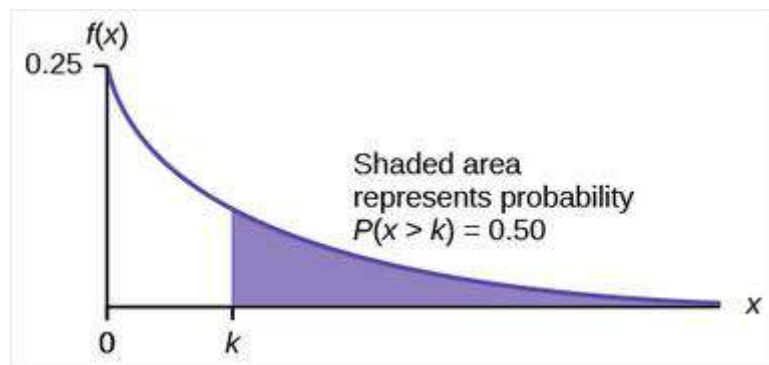
On the home screen, enter

$$(1 - e^{(-0.25 * 5)}) - (1 - e^{(-0.25 * 4)}) \text{ or enter } e^{(-0.25 \\ * 4)} - e^{(-0.25 * 5)}.$$

b) Half of all customers are finished within how long? (Find the 50th percentile)

Solution:

Find the 50th percentile



$$P(x < k) = 0.50, k = 2.8 \text{ minutes (calculator or computer)}$$

Half of all customers are finished within 2.8 minutes. You can also do the calculation as follows:

$$P(x < k) = 0.50 \text{ and } P(x < k) = 1 - e^{-0.25k}$$

$$\text{Therefore, } 0.50 = 1 - e^{-0.25k} \text{ and } e^{-0.25k} = 1 - 0.50 = 0.5$$

$$\text{Take natural logs: } \ln(e^{-0.25k}) = \ln(0.50). \text{ so, } -0.25k = \ln(0.50)$$

Solve for k :

$$k = \frac{\ln 0.50}{-0.25} = 2.8 = 2.8 \text{ minutes}$$

References

- [1] Clauset, A., C. R. Shalizi, and M. E. J. Newman. 2009. "Power-Law Distributions in Empirical Data." *SIAM Review* 51: 661–703.
- [2] Gentle, JE (2009). *Computational Statistics* . Springer . ISBN 978-0-387-98145-1. Archived from the original on January 24, 2020. Retrieved January 24, 2020 . View it on 08-06-2010 .
- [3] *A modern introduction to probability and statistics : understanding why and how*. Michel Dekking. London: Springer. 2005. p. 219. ISBN 978-1-85233-896-1. OCLC 262680588
- [4] Abdul Hafeez. M, *Statistical inference estimation theory*, Arab Nile Group , p.75
- [5] Zhou, Rick. "Exponential Distribution lecture slides." Available online at www.public.iastate.edu/~riczw/stat330s11/lecture/lec13.pdf (accessed June 11, 2013).