Revista Habanera de Ciencias Médicas

YEAR 2021
20(6) NOVEMBER-DECEMBER
ISSN 1729-519X

ecimed

HISTORICAL

Modern Statistical Tendencies used in Investigations in the Medical Sciences

Tendencias estadísticas modernas utilizadas en investigaciones en las ciencias médicas

Fidel Cathcart Roca[1]*, Yeiny Terry Pena[2]

[1]Universidad de Ciencias Médicas de La Habana, Facultad de Ciencias Médicas "Comandante Manuel Fajardo". La Habana, Cuba.
[2]Caribbean School Medical Sciences. Kingston, Jamaica.

***Corresponding author:** cathcart@infomed.sld.cu

How to cite this article
Cathcart Roca F, Terry Pena Y . Modern Statistical Tendencies used in Investigations in the Medical Sciences. Rev haban cienc méd [Internet]. 2021 [cited]; 20(6):e4605. Available from: http://www.revhabanera.sld.cu/index.php/rhab/article/view/4605

Received: July 15, 2021
Approved: September 17, 2021

ABSTRACT

Introduction: The name Statistics was probably used in Egypt in 1549. The term Statistics, which was connected to the development of sovereign states, was coined in Germany in 1749 and used to designate the systematic collection of demographic and economic data by states.

Objective: To present the most relevant and current statistical techniques in clinical medicine and epidemiology, as well as to show examples of the use of these techniques.

Material and Methods: Review of the literature on the subject; presentation of some examples developed in class by professors of the subject Research Methodology and Statistics in Medical Sciences.

Development: From the very beginnig, Statistics was used to characterize data based on their properties, as well as to develop information summary measures.

In the modern era, procedures and techniques were designated to facilitate valid inferences to the universe from the sampling theory, whose foundation is the probability theory.

Conclusions: There is a need for the application of modern techniques and especially the multivariate ones used to explain biological phenomena, which cannot be explained by one or two variables. This makes possible that our universities, scientific research centers, and companies conduct studies using statistical techniques that involve many variables, which are supposed to be related to the variable under study.

Keywords:
Statistics, multivariate analysis, probabilities, sampling, correlation, regression.

RESUMEN

Introducción: El nombre de Estadística probablemente se usó en Egipto en 1549. El término Estadística se acuñó en Alemania en 1749, conectado con el desarrollo de estados soberanos y designando la recopilación sistemática de datos demográficos y económicos por estados.

Objetivo: Presentar las técnicas estadísticas más relevantes en el campo de la clínica médica y la epidemiología actualmente y mostrar ejemplos del uso de dichas técnicas.

Material y métodos: Revisión de la literatura sobre el tema, algunos ejemplos desarrollados en clases por profesores de la asignatura Metodología de la investigación y Estadística en carreras de Ciencias Médicas.

Desarrollo: En sus inicios la Estadística caracterizaba los datos basados en sus propiedades, desarrollaron medidas de resumen de información.

En la era moderna se diseñaron procedimientos y técnicas para realizar inferencias válidas al universo a partir de la teoría del muestreo, cuyo fundamento es la teoría de probabilidades.

Conclusiones: Necesidad de la aplicación de las técnicas modernas y especialmente las multivariadas para explicar los fenómenos biológicos, que no pueden ser explicados por una o dos variables, esto hace que nuestras universidades, centros de investigaciones científicas y empresas realicen sus estudios, utilizando técnicas estadísticas que envuelvan muchas variables que se suponen se relacionen con la variable objeto de estudio.

Palabras Clave:
Estadística, análisis multivariante, probabilidades, muestreo, correlación, regresión.

INTRODUCTION

The name Statistics was probably used in Egypt in 1549. The term was coined in Germany in 1749 and connected to the development of Sovereign States.

It was also used to designate the systematic collection of demographic and economic data by states.

In the early 19th century,[1,2] Statistics was the discipline concerned with the collection, summary, and analysis of data. Today, the data are collected and statistics are computed. It is widely distributed among the government, business, sports, and sciences.

The **objective** of this investigation is to present the most relevant statistical techniques used in the medical field.

DEVELOPMENT

Inception

Descriptive Statistics: Tables of frequency; Measures of Central Tendency; Measures of Dispersion and Variation; Measures of Position to characterize data based on their properties.

Modern era

Inferential Statistics: t-tests; ANOVA; Regression (Factorial analysis); Multiple discriminant analysis and logistic regression; Cluster analysis; others to predict outcomes from known predictor variable(s).

Statistics in Sciences

The basis of Inferential Statistics relies on the Theory of Probability.

Inferential Statistics allows us to infer or generalize the results of the data taken from a sample of a "probabilistic type".

Statistical techniques can show, whether and how, that strongly pairs of variables are related.

Simple: It uses an independent variable.

Multiple: It uses more than two independent variables.

Quantitative Variables: Pearson; Fisher; Kendall; Gosset.

Qualitative Variables: Spearman.

The development of Information and Communication Technologies has allowed and facilitated education, research, and business management as well as biomedical research during the last decades of the 20th century and the first decades of the 21st century.

Biomedical phenomena cannot be explained only by counting and classifying data, but once the data have been classified and specific techniques have been used, techniques to make predictions, estimate parameters, and establish relationships between various characteristics are needed. In general, many variables are required so that the phenomenon under study can be explained. Therefore, the use of statistical techniques that help in this regard such as multivariate analyzes is a must.

Important contributors and founders of statistics

Sir Ronald Aylmer Fisher (17 february 1890 - 29 july 1962) was a Bristish statistician and geneticist. He was the most significant statistician of the 20thcentury.[3] (**Figure 1**)

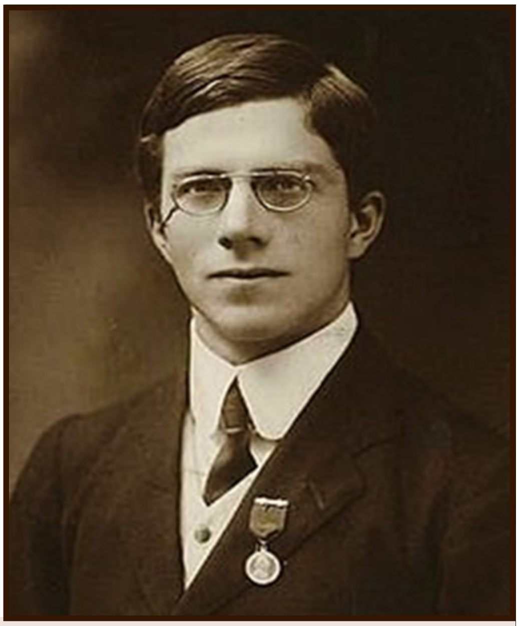


Figure 1- Sir Ronald Aylmer Fisher (17 february 1890 - 29 july 1962)

For his work in statistics, he has been described as a "genius" who almost single-handedly created the foundations for modern statistics.

Charles Edward Spearman (10 September 1863 – 17 September 1945), was an English psychologist who was known for his work in statistics as a pioneer of factor analysis, as well as for Spearman´s rank coefficient that is not parametric.[4] (**Figure 2**)

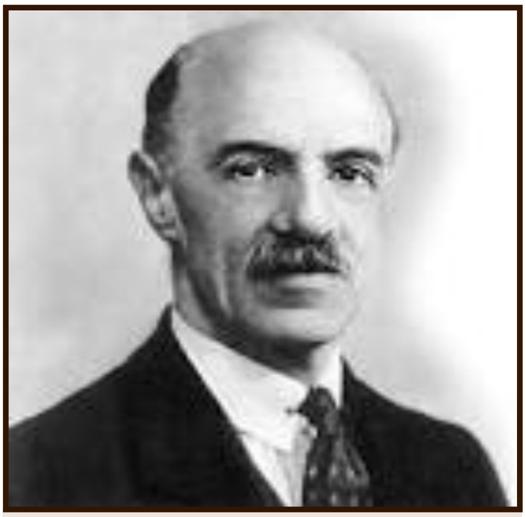


Figure 2- Charles Edward Spearman (10 September 1863 – 17 September 1945)

William Lealy Gosset

He designed the famous Student´s t Test. He was called "student" as statistician.[5]

Inferential statistics techniques most frequently used today

Logistic Regression

It is a widely used statistical model.

Dependent Variable: binomial, multinomial.

Dichotomous; Dummy; In Range

Predictor variables may be either continuous or categorical.

In the **Table 1**, you can see the data for applying simple linear correlation.

Pearson Simple Linear correlation (correlation) r= 0,791. [-1; +1]

Table 1- LDL-Cholesterol and total lipids

LDL-Cholesterol and total lipids	Total Lipids
6,8	7,2
6,5	7,0
3,9	5,0
5,7	6,5
7,3	9,5
6,5	7,0
6,5	6,8
4,7	5,2
5,7	7,0
7,0	9,0

Simple Lineal Correlation

If r is close to 1, a hypothesis test must be performed.

You must prove that there is correlation. Without Correlation, Regression must not be stated

Equation of Linear Regression

Y* = a + bX

Substituted values

Y*=-0,2+1,18x

Particular value=8

Y*= 9,2

Y = a + bX

Normal equations

∑ summatory

∑y= na+b∑x

∑yx= a∑x+b∑x^2

In this case, it allows to find the values of “a” and “b”

Y = a + bX

69,5= 10a + (60,6) (1,18)

69,5= 10a + 71,5

-2= 10a

a= -0,2

Linear Regression Equation

Y* = a + bX

Substituted values

Y*=-0.2+1.18x

Particular value=8

Y*= 9.2

Equation of Linear Regression

It is used to estimate the values or the presence of the variable in the study in any type of correlation (multiple, linear or logistic). From there, an equation is derived.

In **Table 2**, you can see the presentation of data for multiple correlation.

Table 2- Relationship between memory level, age, and cognitive function

X1	X2	X3
Memory Level	**Age**	**Cognitive function**
38,2	52	91,0
40,0	50	86,0
36,0	63	87,0
35,2	61	90,0
34,0	57	80,0
31,0	70	86.5
38,0	59	78,0
33,0	65	89,0
42,5	50	88,0
30,0	72	85,0

Pearson´s Multiple correlation

The Multiple Regression Equation is derived from a system of Normal equations.

X_1*= a1,23 + b12,3X2+b13,2X3

X_1*= 62,113-0,464X2+0,017X3

r= 0,918

r^2= 0,843

F= 18,827

p < 0,01

Pearson´s Multiple correlation[6,7]

X_1*= a1,23 + b12,3X2+b13,2X3

X_1*= 62,113-0,464X2+0,017X3

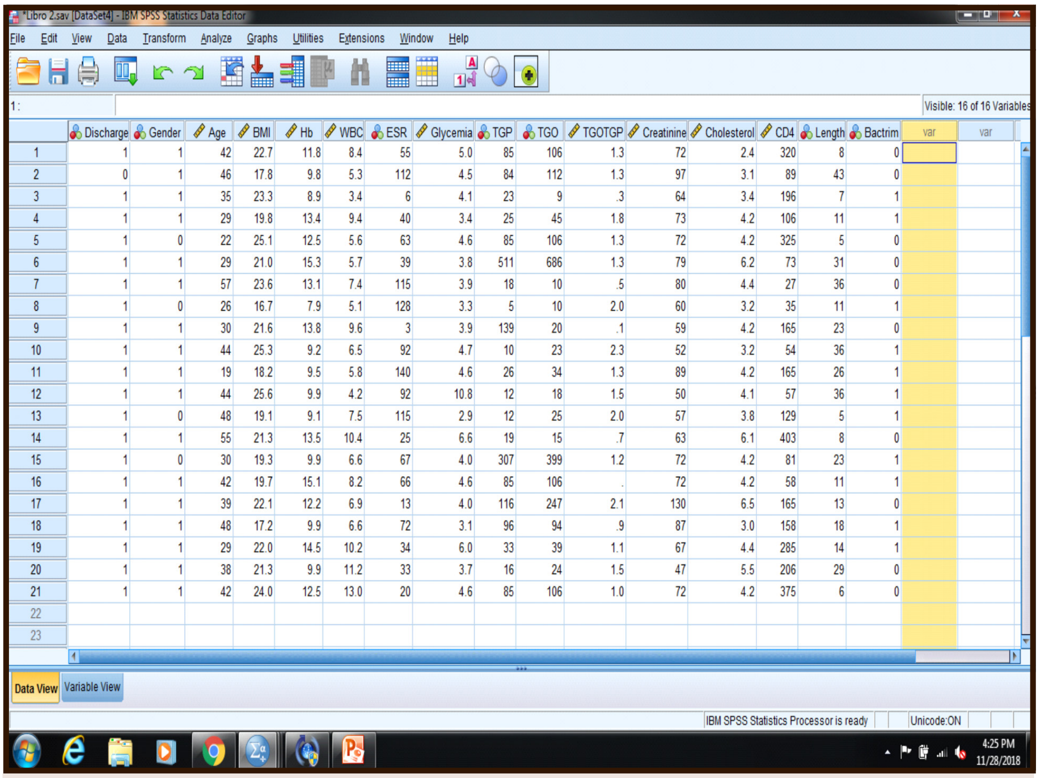
Estimation of X2=74 and X3 = 80

X_1*= 62,113-0,464 (74)+0,017(80)

X_1*= 62,113-34,336+1,36

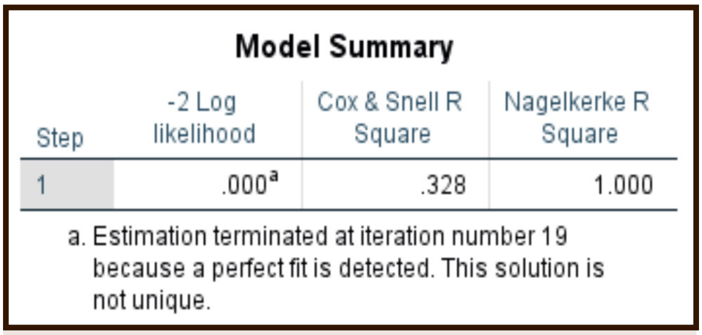
X_1* = 29,137

Figures 3 and **4** show an example of the application of logistic regression.


*Libro 2.sav [DataSet4] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Extensions Window Help
1 :
Visible: 16 of 16 Variables
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON
4:25 PM
11/28/2018

	Discharge	Gender	Age	BMI	Hb	WBC	ESR	Glycemia	TGP	TGO	TGOTGP	Creatinine	Cholesterol	CD4	Length	Bactrim	var	var
1	1	1	42	22.7	11.8	8.4	55	5.0	85	106	1.3	72	2.4	320	8	0		
2	0	1	46	17.8	9.8	5.3	112	4.5	84	112	1.3	97	3.1	89	43	0		
3	1	1	35	23.3	8.9	3.4	6	4.1	23	9	.3	64	3.4	196	7	1		
4	1	1	29	19.8	13.4	9.4	40	3.4	25	45	1.8	73	4.2	106	11	1		
5	1	0	22	25.1	12.5	5.6	63	4.6	85	106	1.3	72	4.2	325	5	0		
6	1	1	29	21.0	15.3	5.7	39	3.8	511	686	1.3	79	6.2	73	31	0		
7	1	1	57	23.6	13.1	7.4	115	3.9	18	10	.5	80	4.4	27	36	0		
8	1	0	26	16.7	7.9	5.1	128	3.3	5	10	2.0	60	3.2	35	11	1		
9	1	1	30	21.6	13.8	9.6	3	3.9	139	20	.1	59	4.2	165	23	0		
10	1	1	44	25.3	9.2	6.5	92	4.7	10	23	2.3	52	3.2	54	36	1		
11	1	1	19	18.2	9.5	5.8	140	4.6	26	34	1.3	89	4.2	165	26	1		
12	1	1	44	25.6	9.9	4.2	92	10.8	12	18	1.5	50	4.1	57	36	1		
13	1	0	48	19.1	9.1	7.5	115	2.9	12	25	2.0	57	3.8	129	5	1		
14	1	1	55	21.3	13.5	10.4	25	6.6	19	15	.7	63	6.1	403	8	0		
15	1	0	30	19.3	9.9	6.6	67	4.0	307	399	1.2	72	4.2	81	23	1		
16	1	1	42	19.7	15.1	8.2	66	4.6	85	106	.	72	4.2	58	11	1		
17	1	1	39	22.1	12.2	6.9	13	4.0	116	247	2.1	130	6.5	165	13	0		
18	1	1	48	17.2	9.9	6.6	72	3.1	96	94	.9	87	3.0	158	18	1		
19	1	1	29	22.0	14.5	10.2	34	6.0	33	39	1.1	67	4.4	285	14	1		
20	1	1	38	21.3	9.9	11.2	33	3.7	16	24	1.5	47	5.5	206	29	0		
21	1	1	42	24.0	12.5	13.0	20	4.6	85	106	1.0	72	4.2	375	6	0		
22																		
23																		

Figure 3- Data for logistic regression



Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	.000[a]	.328	1.000

a. Estimation terminated at iteration number 19 because a perfect fit is detected. This solution is not unique.

Perfect Association

Figure 4- Cox and Snell R square and Nagelkerke R square

There are variables that do not contribute, but their inclusion depends on the experience of the researcher. Inference is a theory to aid decision making but it does not determine the final decision.

CONCLUSIONS

The need for the application of modern techniques and especially multivariate ones to explain biological phenomena, which cannot be explained by one or two variables, makes our universities, scientific research centers, and companies carry out their studies using statistical techniques that involve many variables that are supposed to be related to the variable under study.

REFERENCES

1. Instituto Superior de Ciencias médicas de La Habana. Bioestadística y Computación. La Habana: Editorial Pueblo y Educación; 1987.

2. Instituto Superior de Ciencias médicas de La Habana. Cuaderno de ejercicios de Bioestadística. La Habana: Editorial Pueblo y Revolución; 1988.

3. Efron B. Fisher in the 21st Century. Invited Paper Presented at the 1996 R.A. Fisher Lecture. Statistical Science [Internet]. 1998 [Citado 02/06/2021];13(2):95-122. Disponible en:

https://www.mit.edu/~18.655/papers/efron1998.pdf

4. Aruquipa J. Historia de la Estadística. The American statistician [Internet]. 2021 [Citado 02/06/2021]; 49(2):121 2021. Disponible en: https://es.scribd.com/document/543874806/HISTORIA-DE-LA-ESTADISTICA

5. David HA. Prueba t de Student. The American statistician [Internet]. 2020 [Citado 02/06/2021];44(4):322-6. Disponible en: https://www.studocu.com/ec/document/analisis-de-datos/prueba-t-de-student/19716076

6. Joseph F. Análisis Multivariante [Internet]. Argentina: DocerArgentina; 2021 [Citado 02/06/2021]. Disponible en: https://docer.com.ar/doc/v1env

7. Pearson ES. *"Student'" as Statistician.* Biometrika [Internet].1939;30(3/4):210-50. Disponible en: https://doi.org/10.2307/2332648

Conflict of interest

The authors declare that there are not conflicts of interest.

Author´s contribution

Both authors participated in the discussion of results and have read, reviewed, and approved the final text of the article.