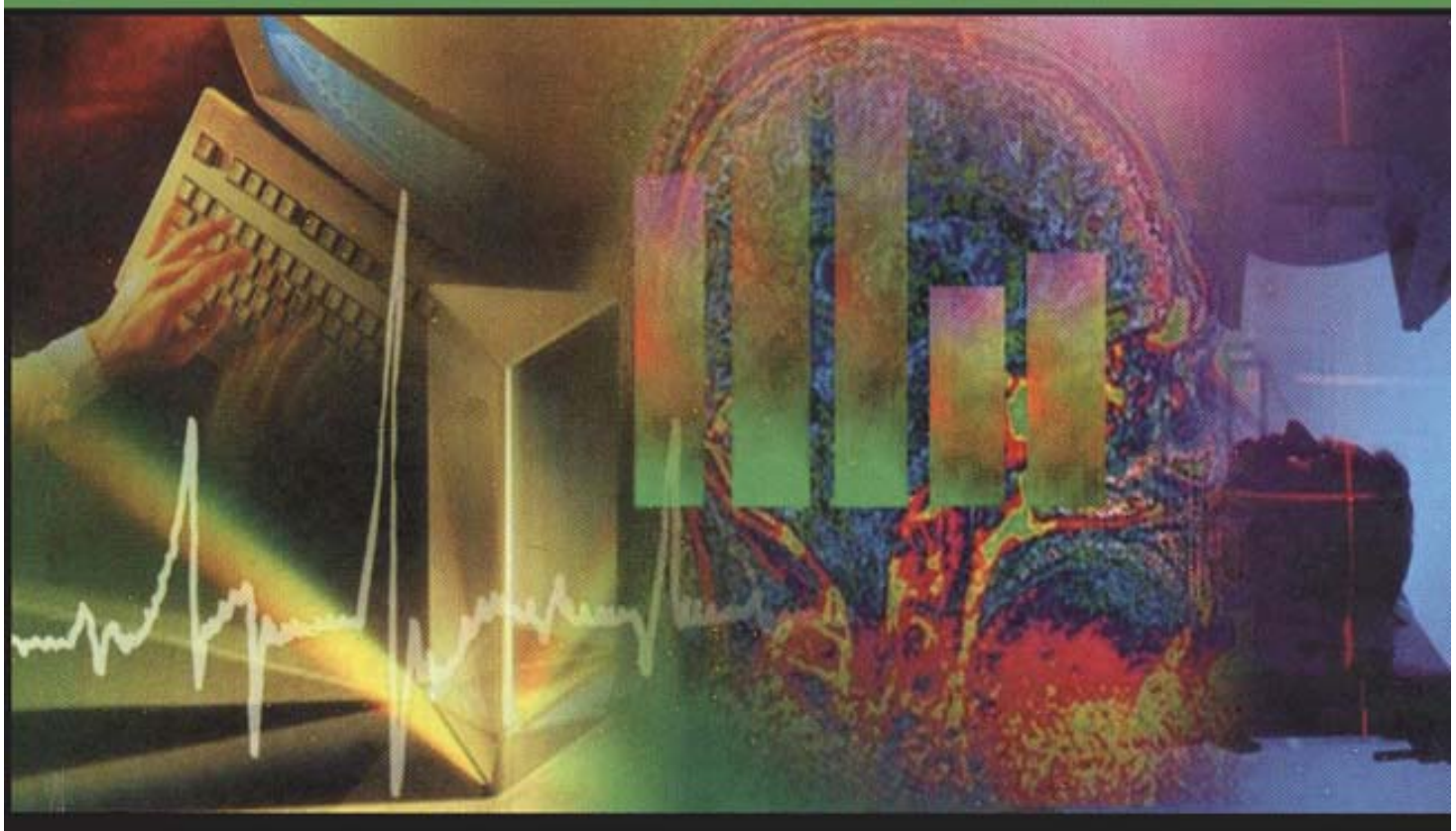
Pedro José Salinas

ESTADÍSTICA PARA INVESTIGADORES



Pedro José Salinas

Ingeniero Agrónomo, Universidad Central de Venezuela 1962. Investigador, extensionista y docente, en la Fundación Shell (Servicio Shell para el Agricultor) de enero 1963 a agosto 1968. Diploma del Imperial College of Science, Technology and Medicine, London, 1966, MSc Univ. of London 1966, PhD Univ. of London 1972. Desde 1968 es Profesor de la Universidad de Los Andes. Actualmente es Profesor Titular en la Facultad de Ciencias Forestales y Ambientales, en la Facultad de Medicina y en la Facultad de Odontología. Ha tutorado y asesorado tesis de Postgrado y Trabajos de Ascenso en varias especialidades ambientales, médicas y odontológicas. Ha publicado trabajos especializados en revistas científicas nacionales e internacionales, arbitradas e indizadas. Fue Director de la "Revista Forestal Venezolana". Fue seleccionado e invitado especialmente para organizar y fundar "MedULA, Revista de la Facultad de Medicina de la Universidad de Los Andes" y es su Editor Jefe desde su creación y de la Revista Odontológica de los Andes (Facultad de Odontología, Universidad de Los Andes), es su Editor Asociado desde su creación Es miembro de la Asociación de Editores de Revistas Biomédicas Venezolanas (ASEREME) y de la World Association of Medical Editors (WAME). Fue Miembro (durante su existencia) del Grupo de Ecología, posteriormente del Ambiente, del CONICIT, que elaboró el Plan Nacional de Ciencia y Tecnología en Ecología, los Proyectos de Leyes: Orgánica del Ambiente; Conservación, Defensa y Mejoramiento de la Cuenca Hidrográfica del Lago de Maracaibo; Penal del Ambiente; de Aguas; y de Playas, consultorías ambientales a instituciones públicas y privadas,

ESTADÍSTICA PARA INVESTIGADORES

ÍNDICE

CAPÍTULO 1.

UNIVERSO. POBLACIÓN. MUESTRA. OBSERVACIÓN.

INTRODUCIÓN.

La estadística es una de las más importantes disciplinas de la ciencia. La estadística juega un papel muy importante en todas las actividades del hombre. El nombre estadística se origina en el término *Status* que significa *Estado*, porque inicialmente se refería a todo lo concerniente al manejo y la administración de los estados o naciones. El censo de las personas, sus actividades económicas, sus propiedades, sus labores, sus nacimientos, muertes y otras cifras demográficas, sus ingresos y egresos, etc. Hoy día la estadística es parte integral de todas las actividades del hombre y es uno de los principales instrumentos, generalmente esencial e indispensable, de la mayoría de ellas, por ejemplo, las actividades científicas, administrativas, políticas, comerciales, de comunicaciones, de producción, industriales, etc.

Hay algunas definiciones humorísticas o cínicas de la estadística, que no deben ser tomadas estrictamente en serio, por ejemplo: 1) La ciencia que al apretarla por el cuello dice lo que uno quiere que diga. 2) La estadística la usan los políticos como los borrachos usan los postes de luz: para sostenerse y no para iluminarse. 3) La estadística es como el bikini que muestra mucho pero oculta lo esencial. 4) Según el Primer Ministro inglés Disraeli: Hay tres tipos de mentiras: las mentiras, las malditas mentiras y la estadística.

En lo que a este texto concierne, la idea es presentar los métodos estadísticos más comúnmente usados en investigación científica y explicarlos de una manera sencilla y fácil, sin necesidad de conocimientos especiales en matemáticas, para que sea entendido por la mayoría de aquellos que se inician en la investigación científica.

La estadística, en nuestro caso, se refiere a los métodos científicos de colectar, organizar, resumir, presentar y analizar datos, generalmente cifras, así como sacar conclusiones válidas y tomar decisiones sobre la base de esos análisis.

La estadística se puede considerar como la mejor forma de estudiar y analizar datos numéricos.

La estadística se puede dividir en estadística descriptiva o deductiva y estadística analítica o inductiva. La estadística descriptiva es aquella que se encarga de describir y analizar las características de un grupo de observaciones o muestra, sin sacar conclusiones o inferencias acerca del grupo mayor o población. La estadística analítica se encarga de analizar la muestra y sacar conclusiones, es decir, inferir la interpretación de los datos analizados. Como la inferencia o las conclusiones no son basadas sobre el total de la población, generalmente se refiere dicha inferencia a la probabilidad.

Para información detallada sobre los diferentes aspectos de la estadística, especialmente de la estadística aplicada, se sugiere consultar las siguientes referencias: Fisher y Yates (1963), Panse y Sukhatme (1959), Snedecor (1956), Snedecor y Cochran (1989), Sokal y Rohlf (1994), Spiegel (1991).

Universo: Se denomina universo o población al grupo total de individuos, sujetos u objetos que componen ese grupo. Así, el conjunto de todas las personas del mundo, será el universo o población mundial de personas. Igualmente, el conjunto de personas de Venezuela o del Estado Mérida, serán los universos o poblaciones de personas de Venezuela o del Estado Mérida. De la misma manera, el conjunto total de pupitres de una escuela o el conjunto de caballos de una hacienda, o el conjunto de plantas de maíz de una región, o el conjunto de pacientes que asisten a un hospital, serán los universos o poblaciones de esos objetos o individuos en sus respectivos ámbitos.

Población: es el conjunto de elementos, individuos, observaciones con características similares que se encuentran en un sitio y tiempo determinados, es decir, están delimitados en espacio y tiempo. Generalmente se le identifica como N.

Las poblaciones pueden ser finitas o infinitas. Una población es finita cuando tiene un número determinado de individuos, por ejemplo el conjunto de panes que produce una panadería en un día. Mientras que una población es infinita cuando no tiene un límite definido, por ejemplo el número de hojas que pueden caer de los árboles de un bosque tropical.

Observación: Se le denomina observación a cada unidad de estudio, por ejemplo, cada individuo, planta, insecto, roca, casa, vaca, bombillo, paciente, lápiz, carro, etc. También se le denomina **dato.**

Los datos pueden ser de dos tipos: **datos categóricos** y **datos numéricos**.

Los **datos categóricos**, como su nombre indica se refieren a categorías y donde cada individuo u observación es uno de un número de clases mutuamente excluyentes. Los datos categóricos, a su vez, son de dos tipos: **datos nominales** y **datos ordinales.**
Los datos nominales son aquellos que se refieren a condiciones o características que solo se pueden separar por nombres (de allí su denominación), por otra parte, no pueden ser ordenados por jerarquía, ya que no hay ninguno mayor que otro. Ejemplos de datos nominales son: referidos al sexo, hombre, mujer o varón, hembra, o macho, hembra; referidos a la ocupación: comerciante, abogado, limpiador, ingeniero, cocinero, taxista, profesor; referidos al estado civil: soltero, casado, divorciado, viudo; referidos a la procedencia: urbano, sub-urbano, rural.
Los datos ordinales son aquellos que se refieren a pueden ordenarse (de allí su nombre) de mayor a menor, de inferior a superior, de mejor a peor, etc. o viceversa. Ejemplos de datos ordinales son: referidos a la aceptación de un candidato a un puesto público: eficiente, regular, malo, muy malo; referidos a las notas de alumnos: excelente, sobresaliente, bueno, regular, aplazado; referidos al efecto de un medicamento: eficaz, medianamente eficaz, ineficaz; referidos a la opinión sobre algún tema: de acuerdo, medianamente de acuerdo, en desacuerdo. Por supuesto que estas escalas las debe establecer el investigador, es decir, que pueden ser más o menos tipos en cada escala.

Los **datos numéricos** son aquellos que pueden presentarse o expresarse con cifras o números, por ejemplo, la población de una ciudad o país, la estatura de los estudiantes en una escuela, los reprobados en un examen, los matrimonios en una iglesia, las vacas de una finca, las picadas de hormigas por persona en un día de campo, las casas en una urbanización, la edad de los pacientes en un hospital, etc.
Los datos numéricos son de dos tipos de acuerdo con su configuración, es decir, si no admiten división de la unidad, se denominan **discretos**, por ejemplo, el número de personas en una familia, las ranas en un charco, los árboles de un bosque, los accidentes de tránsito (no se puede decir que en una familia hay siete personas y tres cuartos personas, o en un charco hay 523.75 ranas, etc. Mientras que si admiten división de la unidad se denominan **continuos**, como ocurre con la edad, la estatura, la temperatura, etc, que pueden dividirse, prácticamente hasta el infinito, asi podemos decir que la estatura de una persona es de 1.45398102 metros o que su edad es de 36.4902461832 años y su temperatura corporal es de 36.703485 grados Celsius. Sin embargo, los datos continuos pueden agruparse y ordenarse como datos categóricos, por ejemplo, las edades de un grupo de personas pueden agruparse en categorías, así se tendría 1-5 años, 6-10 años, 11-15 años, etc. Debe tenerse en cuenta que no deben sobreponerse, es decir, no pueden agruparse como 1-5, 5-10 años, 10-15 años, etc.

Los datos, para fines de cálculos estadísticos, generalmente se les identifican con x. Cada observación será: x_1, x_2, x_3, x_4, … x_n. Cuando son más de dos serán y, z, etc.

Muestra. Generalmente no es posible o es impráctico observar o estudiar todo el conjunto de individuos u objetos, es decir, toda la población o universo, porque es muy costoso, toma mucho tiempo, es muy difícil, etc. En ese caso se estudia solamente una parte que se llama muestra. La muestra es el conjunto de observaciones que se toman de una población y que se supone representa todas las características generales de la población de estudio. Generalmente se le identifica como n.
La muestra debe tener todas o la mayoría de las características de la población. La muestra se considera **representativa** cuando es igual o mayor al 10% de la población, pero como esto es muchas veces difícil o imposible de lograr por lo costoso, por el tiempo que toma o por los recursos humanos necesarios, se usa la denominada muestra **significativa**, que es aquella que tiene una probabilidad del 95% de representar las características de la población. Esta muestra no tiene un número predeterminado de observaciones, por lo cual se toma la muestra (el mayor número posible de observaciones, nunca menos de tres) y se hace el análisis estadístico para calcular su probabilidad.
Lo ideal en cualquier investigación es que, como muestra, se tomen todos los individuos de una población, en este caso la muestra será, obviamente, 100% representativa y 100% significativa de la población.

CAPÍTULO 2.

VARIABLES. CONSTANTE. FRECUENCIA.

Variable: Por variable se entiende a una característica, cualidad, fenómeno, etc. que varía, se modifica o cambia, es decir, que puede tomar cualquiera de los valores de un grupo determinado.
Una variable que puede tener cualquier valor entre dos valores dados se denomina **variable continua**. Por ejemplo, la estatura de una persona puede ser 168 cm, 168.5 cm, 168.53 cm, 168.53749 cm, 168.53749062 cm y así hasta el infinito.
Una variable que solo puede tener valores fijos entre dos valores dados se denomina **variable discreta**. Por ejemplo, el número de hijos en una familia puede ser 0, 1, 2, 3, 4, etc., pero no puede ser 2.5 ni 4.67, ni 5.873, etc., es decir, no puede ser la fracción de la unidad.

Variables dicotómicas, de modalidad, binarias o booleanas: Las variables categóricas (aquellas que representan categorías) pueden ser dicotómicas, es decir, que pueden tener solo dos alternativas o posibilidades, o dos grados, que se generalizan como presencia/ausencia de la característica en estudio, por ejemplo, sexo masculino/femenino, fumador/no fumador, VIH positivo/negativo, despierto/dormido, vivo/muerto, embarazada/no embarazada, día/noche, claro/oscuro, bueno/malo, aprobado/reprobado, etc. Se les llama binarias por tener solo dos alternativas y booleanas por haber sido estudiadas y documentadas por el inglés Bool.

En los casos donde una condición depende de otra condición, es decir, una situación es causada por otra, o un efecto es ocasionado por una causa, las condiciones se denominan: **variable independiente** a la que genera el efecto y este se llama **variable dependiente**. Por ejemplo, la aplicación de diferentes dosis del fertilizante nitrato de amonio (variable independiente) a una siembra de maíz causará aumento en el crecimiento y por ende en el rendimiento en kilogramos por hectárea (variable dependiente). A medida que aumenta la dosis de fertilizante aumentará el rendimiento del maíz. La relación puede ser inversa, es decir, a medida que aumenta la independiente, disminuye la dependiente. Por ejemplo, la aplicación de diferentes dosis de insecticidas (variable independiente) a una población de mosquitos (variable dependiente) causará su disminución.
En algunos casos, especialmente relacionados con la epidemiología, se denominan **variables desconcertantes** (en inglés: *confounding variables*) a aquellas que están asociadas tanto a la causa como al efecto que se investiga. Por ejemplo, si en un grupo de personas entre 0 y 20 años de edad se relaciona la edad con el peso, se encontrará que cuando la edad (variable independiente o causa) aumenta, también aumenta el peso (variable dependiente o efecto), pero si existen otros elementos asociados, tal como el consumo de vitaminas, esto estará asociado a la causa y al efecto, por lo cual se le llama variable desconcertante. Otro término para denominar este tipo de variables es de **variables intervinientes**.

Constante: Si la variable solo puede tomar un solo valor, se le denomina constante, por ejemplo, la temperatura de ebullición del agua, en el nivel del mar, a temperatura ambiente de 20 ºC y presión atmosférica de 760 mm de Hg, es de 100 ºC.

Datos crudos se denomina a los datos tomados directamente del experimento, investigación o trabajo que se realiza o se ha realizado, pero que no están arreglados u ordenados. Por ejemplo, los datos de las edades, las estaturas o pesos de los estudiantes de una escuela tomados de acuerdo con el orden alfabético. En ese caso se consideran datos crudos, pues no están arreglados en ningún orden numérico.

Datos ordenados son aquellos datos que se han ordenado para su estudio o análisis. El orden puede ser ascendente o descendente, de acuerdo con el tipo de análisis que se va a realizar. Generalmente se usa el orden ascendente.

Frecuencia (frecuencia absoluta): Se llama frecuencia o frecuencia absoluta al número de veces que una observación se repite. Por ejemplo, en un plantel educativo las edades de los alumnos y sus frecuencias son: 10 años 20 alumnos, 11 años 14 alumnos, 12 años 18 alumnos, etc. Las frecuencias serán los números de alumnos que se repiten en cada edad. Generalmente, se llama clase a cada grupo de valores repetidos, así para la edad 11 años, esa será la clase, y luego serán la clase 14 años, 12 años, etc.

Clase y Frecuencia de clase: Cuando existen grandes cantidades de datos, de los cuales muchos son repetitivos, se deben resumir en grupos similares. Estos grupos son llamados clases o categorías. Se debe determinar el número de individuos que se incluirá en cada clase, esta cifra se denomina la frecuencia de clase. Luego se elabora una tabla con los valores ordenados por frecuencia, la cual se denomina distribución de frecuencia o tabla de frecuencia. Por ejemplo, las edades de los estudiantes de una escuela se ordenan en la siguiente tabla de frecuencia:

Tabla 1. Edades en años de 100 estudiantes de la escuela X.

Edad (años)	Número de estudiantes
5 - 7	5
8 – 10	18
11 – 14	42
15 - 17	27
18 - 20	8
Total	100

Las cifras que definen una clase, tal como podría ser 11 – 14 en la tabla anterior, se denominan intervalo de clase. La primera y última cifra se denominan los límites de clase, en este caso serían 11 y 14, donde 11 es el límite inferior de clase y 14 es el límite superior de clase. Los términos clase e intervalo de clase se usan indistintamente, aunque en realidad el intervalo de clase es el valor expresado, por ejemplo, 11 – 14. Cuando el intervalo de clase no tiene un límite superior de clase o inferior de clase se le denomina intervalo de clase abierto, por ejemplo, en una ciudad se pueden contar los adultos mayores de 65 años, sin especificar cada edad, entonces se dice 65 y más años, lo cual es un intervalo de clase abierto.

Los verdaderos **límites de clase** se determinan sumando el límite superior de una clase al límite inferior de la clase inmediatamente superior y dividiendo esta suma entre dos. Por ejemplo, en la tabla 1, los verdaderos límites serían: superior 14 + 15 = 29 y 29 ÷ 2 = 14.5; inferior 10 + 11 = 21 y 21 ÷ 2 = 10.5. En algunos casos, se utilizan estas cifras, pero el problema está en que hay indefinición en dónde ubicar un valor que sea precisamente el verdadero límite, por ejemplo, si tenemos un alumno cuya edad es de 14 años y seis meses exactos, no sabríamos si ubicarlo en la clase 10.5 – 14.5 años o en la clase 14.5 – 17.5. Por eso, se recomienda, en los casos donde exista un dato que quede en esa posición, no usar los verdaderos límites, sino los aproximados, por ejemplo 11 – 14; 15 – 17.

El tamaño de una clase, es decir, la diferencia entre el valor superior y el valor inferior de la clase, se denomina tamaño, amplitud o longitud de clase, aunque se prefiere usar el término amplitud de clase.

El **centro de clase** es el punto medio de un intervalo de clase y se obtiene sumando los límites superior e inferior de una clase y dividiendo la suma entre dos, por ejemplo, en la tabla 1, el centro de clase de 11 – 14 años, sería: 11 + 14 = 25 y 25 ÷ 2 = 12.5. Para fines prácticos en los cálculos matemáticos, se considera que todas las observaciones de una clase tienen el mismo valor del centro de clase, por ejemplo, los 42 estudiantes tienen 12.5 puntos.

Para formar las distribuciones de frecuencia se recomienda lo siguiente: Determinar los valores máximo y mínimo del grupo de los valores crudos y obtener el rango (diferencia entre esos valores). Dividir el rango en un número de intervalos de clase que tengan la misma amplitud, si esto no es posible, entonces usar intervalos de clase de diferente amplitud o usar intervalos de clase abiertos. Generalmente, se usan entre 5 y 20 intervalos de clase, claro está que esto depende del número de datos que se tienen. Luego determinar el número de observaciones que caen en cada intervalo de clase, es decir, determine las frecuencias de clase.

Un **histograma de frecuencias**, llamado simplemente histograma, es la representación gráfica, mediante barras verticales, de las diferentes frecuencias de las clases. En este caso los valores se deben ordenar en forma ascendente. Son varios rectángulos que tienen su base en un eje horizontal, cada uno con centro en el centro de clase y longitud igual a la amplitud de clase. Las áreas de los rectángulos son proporcionales a las frecuencias de clase. Por lo tanto, hay que tener en cuenta que las amplitudes de clase bien sean iguales o diferentes, las alturas de los rectángulos serán proporcionales a las frecuencias de clase (Fig 1).

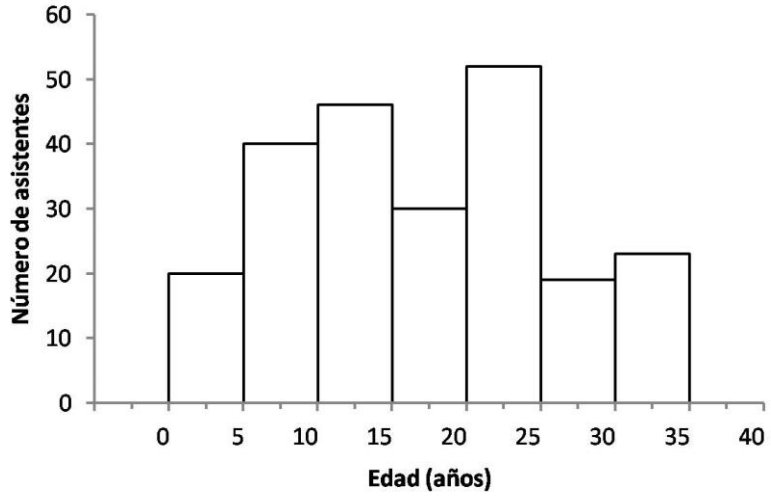


Fig. 1. Histograma de frecuencias.

Un **polígono de frecuencias** se forma cuando se unen con una línea continua los puntos medios de los topes de los rectángulos o barras de un histograma de frecuencia.). Por lo general, el polígono se cierra, uniendo el primer y último centro de clase con el centro de clase inferior y superior, es decir, los valores inicial y final, se estiman ser de valor 0, por lo tanto línea de unión será discontinua (Fig. 2).

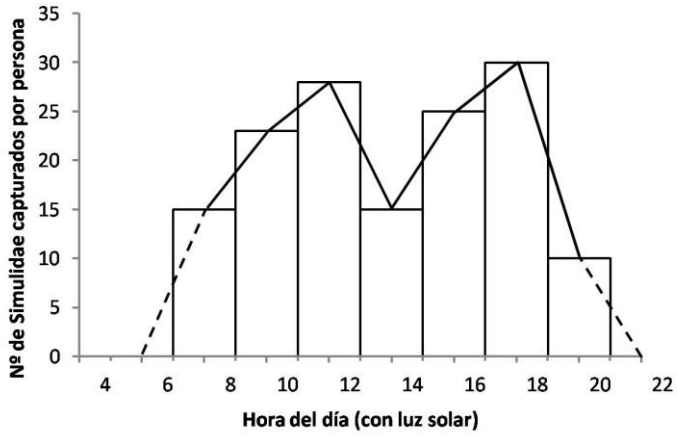


Fig. 2. Polígono de fecuencias.

La **frecuencia relativa** de una clase es la frecuencia de la clase divida entre el total de todas las clases. Generalmente se expresa como porcentaje. En el caso de los estudiantes, la clase 11 – 14 años es 42 ÷ 100 = 42%. Obviamente, la suma de todas las frecuencias relativas es 100%. La tabla con estos valores se denomina distribución de frecuencia relativa, distribución de porcentajes o tabla de frecuencia relativa. Las representaciones gráficas se denominan histogramas de frecuencias relativas, histogramas de porcentajes, polígonos de frecuencias relativas o polígonos de porcentajes.

La **frecuencia (frecuencia absoluta) acumulativa o acumulada** es el total de todos los valores menos el límite verdadero superior de un intervalo de clase dado. Por ejemplo, en la tabla 1, la frecuencia acumulada hasta e incluyendo el intervalo 11 – 14 años es 5 + 18 + 42 = 65, es decir, que hay 65 estudiantes que tienen menos de 14.5 años. La tabla que representa estos valores se denomina distribución de frecuencias acumuladas o tabla de frecuencias acumuladas (ver tabla 2), o puede usarse el término “Hasta …” o “Menos de …”.

Tabla 2. Edad en años de los estudiantes de la escuela XX.

Edad (años)	Número de estudiantes
Hasta 6.5	0
Hasta 7.5	5
Hasta 10.5	23
Hasta 14.5	65
Hasta 17.5	92
Hasta 20.5	100

Sin embargo, es de notar que en la mayoría de las publicaciones solo se indica el centro de clase superior y no el límite superior verdadero del intervalo de clase, en este caso sería como la tabla 3.

Tabla 3. Edad en años de los estudiantes de la escuela XX.

Edad (años)	Estudiantes
7	5
10	23
14	65
17	92
20	100

La **frecuencia relativa acumulada o frecuencia acumulada de porcentajes** es la frecuencia acumulada dividida entre la frecuencia total. Por ejemplo, la frecuencia acumulada relativa de las edades menores de 14 años es 65 ÷ 100 = 65%, que indica que 65% de los estudiantes tienen edades menores de 14 años. Con estos datos se pueden construir las tablas de distribución de frecuencias relativas acumuladas o distribución acumulada de porcentajes, así mismo como los polígonos de frecuencia relativa acumulada, también llamados ojivas de porcentajes por la forma ojival que toman.

Tanto los polígonos de frecuencia como los de frecuencia acumulada, de frecuencia relativa y de frecuencia relativa acumulada son formados por segmentos de líneas rectas (Fig. 3 y Fig. 4), pero para fines de representación en informes, reportes o publicaciones, generalmente se "suavizan" haciéndolos aparecer como líneas curvas continuas (no unión de segmentos rectos) (Fig. 5).

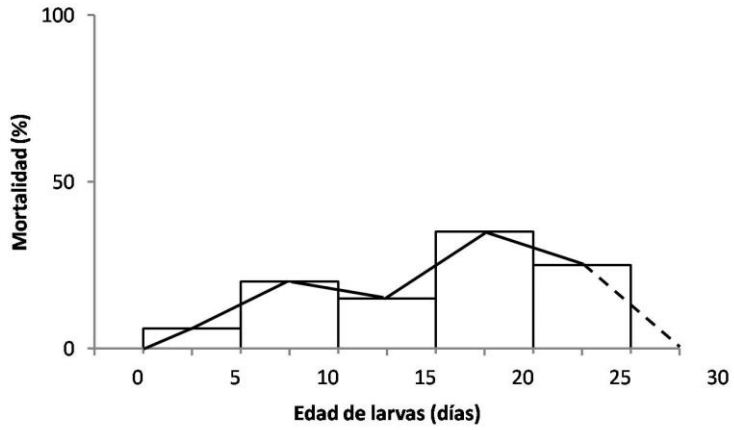


Fig. 3. Polígono de frecuencia relativa.

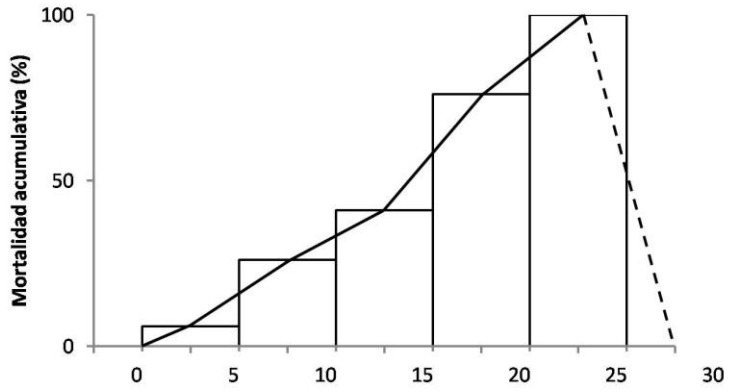


Fig. 4. Polígono de frecuencia relativa acumulada.

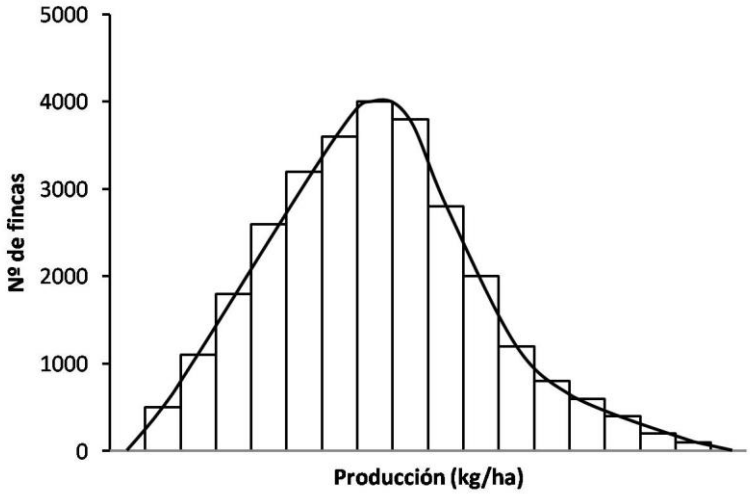


Fig. 5. Polígono "suavizado" para aparecer como un curva, en vez de una serie de segmentos rectos.

CAPÍTULO 3.

MEDIDAS DE TENDENCIA CENTRAL. MEDIA. MEDIANA. MODA.

Medidas de tendencia central.

Dentro de un grupo de datos hay, generalmente, valores que representan al grupo total. Esos valores están hacia el centro del grupo de datos por lo que se les denomina medidas de tendencia central. Las medidas de tendencia central más comunes son la media aritmética, la mediana y la moda. Aunque hay varias medias (media aritmética, media geométrica, media harmónica, media ponderada, etc.) nos referiremos solo a la media aritmética que es la más común y por tanto la más usada. Hay otras medidas de tendencia central que no trataremos por ser, generalmente, más especializadas, tal como la media cuadrática, los cuartiles, deciles y percentiles (en su conjunto llamados cuantiles). A los percentiles también se les llama porcientos o porcentajes.

Para facilitar la comprensión de las siguientes fórmulas se usará la notación convencional de: x para cada valor, observación o dato con un sufijo para indicar su orden, por ejemplo si son 5 observaciones, será: x_1, x_2, x_3, x_4, x_5. se usa la notación hasta *n* observaciones en la muestra, aunque se usa N para denotar la totalidad de la población.

Media.

Media: Hay diferentes tipos de medias o promedios, de las cuales las más usadas en estadísticas son la media aritmética, la media geométrica, la media ponderada y la media harmónica, entre otras.

La media aritmética: También llamada promedio y en inglés "average", se simboliza por medio de una x con una barra arriba. Se calcula dividiendo los valores de todos los elementos o individuos, generalmente llamados en estadística, "observaciones", entre el número de dichas observaciones:

La suma o sumatoria de todos los valores se simboliza con la letra griega sigma mayúscula: Σ, que significa la suma de todos los elementos al que se le antepone.

Donde la letra *j* en X_j indica que la X puede tomar cualquier valor numérico, por ejemplo, 1, 2, 3, 4, 5, … etc. hasta N.

El símbolo $\sum_{j=1}^{N} N$ significa que es la suma de todos los valores X_j, desde $j = 1$ hasta $j = N$, es decir, $x_1 + x_2 + x_3 + x_4 + \ldots \ldots + x_N$

La media se simboliza de manera abreviada como sigue:

$$\overline{x} = \frac{\sum_{j=1}^{N} x_j}{N}$$

En forma muy simplificada se expresa el cálculo de la media como:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + \ldots\ldots + x_N}{N}$$

Por ejemplo, en un grupo de estudiantes, las notas finales son: 12, 17, 20, 7, 15, 18, 10, 13 y 11, la media será (12 + 17 + 20 + 7 + 15 + 18 + 10 + 13 + 11) / 9 = 13.66.

Mediana.

La mediana, $\overline{m}$, en grupo de datos numéricos ordenados de acuerdo con su magnitud (ascendente o descendente) es el valor medio, es decir el valor que divide en dos partes iguales el grupo. Cuando es un número impar de datos u observaciones, será el valor que deja a cada lado igual número de datos. Cuando el número de datos es par será la media aritmética de los dos valores medios. Ejemplos:

a) Número de datos impar: 4, 7, 8, 9, 11, 14, 18, 23, 31. La mediana es 11 porque siendo 9 datos el quinto dato (11) es el que deja igual número de observaciones a cada lado.

b) Número de datos par: 2, 5, 6, 8, 9, 12, 15, 19, 22, 34. la mediana es 10.5, es decir la media aritmética de los dos valores medios (9 + 12 = 21, entre 2 = 10.5).

Cuando se trata de un histograma de frecuencias, la mediana es el valor de x (en el eje de las abscisas) correspondiente a la línea vertical que separa al histograma en dos partes de igual área.

Moda.

La moda, $\ddot{M}$, de un grupo de datos numéricos es el valor que ocurre con mayor frecuencia, es decir, el valor que más se repite. Tal como ocurre en la vida diaria, normal, de cualquier grupo o sociedad la moda es aquello que más se repite, sea una canción, una prenda de vestir, un modelo de carro, una comida, un destino turístico, etc., se dice que es lo que está de moda, pero en estadística, puede que un grupo de datos numéricos no tenga moda, es decir, cuando no hay un valor que se repite más que los otros. Se le llama amodal. Si tiene una moda se le denomina unimodal Por otra parte, puede haber un grupo de datos numéricos que tenga más de una moda, en este caso de le llama bimodal si son dos las modas como ocurre con la precipitación pluvial anual de muchos países tropicales o con la insolación diaria en la ciudad de Mérida, Venezuela. Si tiene más de dos modas se les denomina, genéricamente, polimodales, porque tienen muchas modas (los llamados "picos"), como serían la tensión arterial (sistólica o diastólica), la precipitación pluvial mes por mes de alguna región húmeda, la representación gráfica de un sismógrafo o de un electrocardiograma, etc. Ejemplos son:

Amodal: 1, 3, 4, 7, 9, 11, 14, 17, 20, 25, 37, 43. No tiene moda.

Unimodal: 2, 4, 6, 7, 9, 12, 12, 12, 16, 22, 35. 66. La moda es 12.

Bimodal: 3, 6, 8, 9, 12, 14, 14, 14, 16, 19, 22, 26, 28, 28, 28, 28, 39, 43. Las modas son 14 y 28.

Polimodal: 3, 6, 8, 9, 11, 11, 11, 15, 15, 15, 15, 18, 23, 26, 29, 29, 29, 31. En esta caso tiene tres modas que son 11, 15 y 29.

En una distribución normal, simétrica, hay solo una moda, no hay sesgo y la curva tiene forma de campana simétrica. En esta curva la media, la moda y la mediana coinciden en el mismo punto o valor central.

CAPÍTULO 4.

MEDIDAS DE DISPERSIÓN. RANGO. VARIANZA. DESVIACIÓN STANDARD O TÍPICA. ERROR STANDARD O TÍPICO. COEFICIENTE DE VARIACIÓN. SESGO. KURTOSIS O CURTOSIS

Medidas de dispersión.
El grado en el cual los datos numéricos tienden a esparcirse alrededor de un valor general se denomina variación o dispersión de los datos. Hay varias formas de medir esa variación o dispersión. Entre las más comúnmente usadas están el rango, la desviación media, el rango semi-intercuartil, el rango del percentil 10-90, el coeficiente de variación, la varianza y la desviación standard. Dentro de estos nos referiremos a los más usados en las publicaciones sobre investigación científica.

Rango.
Es el intervalo o diferencia de valores entre el valor máximo y el valor mínimo. Rango = Valor máximo – valor mínimo. Por ejemplo: el rango de edades de una muestra o población entre 45 años y 15 años será: 45 – 15 = 30 años.
El rango de un grupo de datos es la diferencia que existe entre el valor mayor y el valor menor de dichos datos. Por ejemplo, en un grupo de notas de un examen el valor mayor fue 19 puntos y el valor menor fue 5 puntos, entonces el rango de las notas es: 19 – 5 = 14 puntos.
El rango de un grupo de datos numéricos, como ya se ha dicho, es la diferencia entre el valor máximo y el valor mínimo. Por ejemplo, el rango de 2, 5, 8, 10, 13, 16, 22, 27, es 27 – 2 = 25. Algunas veces el rango se indica citando los valores mínimo y máximo. En este caso el rango se citaría como 2 a 27 ó 2 – 27.

Varianza, variancia o variación.
La varianza es la medida de mayor utilidad en todos los análisis estadísticos. Aunque la varianza misma no es de utilidad directa, ella es la base de todos los cálculos en estadística, es decir, nos sirve para calcular la desviación standard, el error standard, el coeficiente de regresión, el coeficiente de correlación y, por supuesto, para hacer el análisis de varianza, entre otros.
La varianza también es llamada media cuadrática, media cuadrada o cuadrado de la media (en inglés Variance y Mean Square) es una medida de la dispersión de los valores alrededor de la media. Mientras más dispersos los valores, mayor será la varianza y por el contrario cuando los valores son muy similares, es decir, no varían mucho, la varianza será muy pequeña hasta llegar a cero cuando todos los valores son exactamente iguales. Así, cuando se desea que algo sea uniforme, por ejemplo, la cantidad de un producto en cada envase, el tamaño de una pieza de algún material o la estatura de algunas personas, etc., se trata de que la varianza sea lo más pequeña posible. Por el contrario, si se desea que la variación sea grande, por ejemplo, la edad de las personas que se someterán a la prueba de una vacuna, o el sueldo de las personas que optarán por casa, etc. Se trata de que la varianza sea lo más grande posible para que los datos abarque al mayor número de personas. La varianza, que se simboliza por sigma minúscula al cuadrado, σ^2, para la población y s^2 para la muestra, se calcula de varias formas (no es solamente elevar al cuadrado la media, como su nombre pareciera indicar) a partir de la suma de los valores elevados al cuadrado $\sum x^2$, luego restado un valor denominado Factor de Corrección, $FC = \frac{\left(\sum x\right)^2}{n}$, y finalmente promediados.

$$s^2 = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}$$

Como un ejemplo sencillo, vamos a calcular la varianza de las notas de algunos estudiantes:
$x_1 = 15$, $x_2 = 7$, $x_3 = 12$, $x_4 = 10$, $x_5 = 17$, $x_6 = 13$, $x_7 = 10$, $x_8 = 20$, $x_9 = 14$, $x_{10} = 15$, $x_{11} = 11$, $x_{12} = 5$. Primero re-agruparemos los valores en orden ascendente, así que ahora serán:
$X_1 = 5$, $x_2 = 7$, $x_3 = 10$, $x_4 = 10$, $x_5 = 11$, $x_6 = 12$, $x_7 = 13$, $x_8 = 14$, $x_9 = 15$, $x_{10} = 15$, $x_{11} = 17$, $x_{12} = 20$.

Luego calculamos su media: en este caso es de 12.4166 que redondearemos a 12.42.

Ahora elevamos al cuadrado cada valor y los sumamos, lo cual será:
$\Sigma x^2 = 25+49+100+100+121+144+169+196+225+225+289+400$
$\Sigma x^2 = 2043$
Ahora calculamos el factor de corrección (FC) que es la suma de los valores originales (149) elevada al cuadrado 149^2 y dividida entre el número de valores n (en este caso 12), lo cual será: FC = 22201/12, FC = 1850.08. Importante: Este valor **nunca** puede ser mayor que la suma anterior o de los valores al cuadrado (2043). Puede ser igual, cuando no hay variación entre los valores, es decir, son exactamente iguales, pero no mayor, pues daría un valor negativo de la varianza y sabemos que cualquier número al cuadrado es positivo, NO (nunca) negativo. Así que este paso nos indica si nuestros cálculos hasta ahora han sido bien hechos. Si este valor da negativo, debemos rehacer todos los cálculos y corregir el error..
Ahora, a la suma de los valores cuadrados (2043) le restamos el factor de corrección: 2043 – 1850.08 = 192.9166 que promediamos, es decir, dividimos entre el número de observaciones (n) aunque para muestras de iguales o menores a 30 observaciones es mejor dividir entre el número de grados de libertad (n – 1), en este caso 12 – 1 = 11, lo que nos da: 192.9166 ÷ 11 = 17.901509 que redondearemos a 17.90. Como vemos es un valor positivo. Este es el valor de la varianza (redondeada en este caso para facilitar la comprensión del lector), pero que en la realidad debe ser calculada al valor más exacto posible).
Otro método fácil de calcular la varianza de un grupo de datos numéricos es sumando la diferencia al cuadrado, entre cada valor y la media de la muestra y luego dividiendo esa suma entre el número de grados de libertad:

$$s^2 = \frac{\sum\left(x-\bar{x}\right)^2}{n-1}$$

En el ejemplo anterior, tenemos como media 12.42 al cual le restaremos cada uno de los valores (por ejemplo: 15 – 12.42 = 2.58; 7 – 12.42 = -5.42 y así sucesivamente hasta llegar a los 12 datos), luego cada uno de esos valores se eleva al cuadrado (por ejemplo: 2.58^2 = 6.6564; -5.42^2 = 29.3764 y así sucesivamente), luego se suman todos estos valores cuadrados y la suma (en este caso 192.9166) se divide entre el número de grados de libertad (11) que resulta en 17.901509 que redondeamos a 17.90. Lógicamente es el mismo resultado que el obtenido por el método anterior. Existen otros métodos para calcular la varianza, pero estos son los más comúnmente usados.

Desviación standard o típica.
Desviación Standard (o Estándar): este es el componente estadístico más usado en todos los experimentos e investigaciones científicas y es una medida de la media de las variaciones de los valores. La desviación standard es la variación o dispersión verdadera y absoluta de los valores estudiados. Se simboliza con sigma minúscula, σ, para la población y con s para la muestra. Su cálculo es muy sencillo una vez que se tiene la varianza, pues es solo la raíz cuadrada de dicha varianza y por ser el producto de una raíz cuadrada, su valor es positivo y negativo y se le prepone el símbolo ±.

$$s = \sqrt{s^2} \qquad s = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1}}$$

En nuestro caso será la raíz cuadrada de 17.90, es decir, $s = \sqrt{17.90}$ es ±4.2308392 que redondearemos a ±4.23. Es importante destacar que las unidades de medición de la desviación standard son, lógicamente, las mismas que las unidades de los datos originales, es decir, si estamos midiendo las edades en años, las unidades de la desviación standard estarán expresadas en años. En las publicaciones, informes, reportes, etc., generalmente, se expresan los valores de la media y las desviaciones standard como sigue: la media ± la desviación standard y las unidades de medición, por ejemplo, $\bar{x} \pm s$, en el caso de las edades en años sería: 12.42 ± 4.23 años
En una población o muestra con distribución normal cuya curva tiene forma de campana simétrica, el 68.27% de los valores caen dentro de la media más o menos una desviación standard, es decir que si en nuestro caso la distribución fuese normal, 68.27% de los valores estarían entre 12.42±4.23, o sea entre 8.19 y

16.65. Esta operación nos indica que las notas de los estudiantes están muy dispersas y que no siguen la distribución normal.
En la distribución normal el 95.45% de los valores caen dentro de la media más o menos dos desviaciones standards y el 99.73% de los valores caen dentro de la media más o menos tres desviaciones standards.

Error standard o típico.
El error standard o típico también llamado de la media, es una medida estadística de la probabilidad de que lo encontrado en la muestra refleje lo que se encuentra en la población. Por lo tanto, el error standard depende de dos factores: el tamaño de la muestra y la variación de las observaciones medida como desviación standard. El cálculo se realiza dividiendo la desviación standard entre la raíz cuadrada del número de observaciones:

$$s_{\ddot{x}} = \frac{s}{\sqrt{n}}$$

El error standard en sí mismo no tiene mucha utilidad, pero se usa para calcular los límites de o intervalo de confianza que si es de gran utilidad.

En el ejemplo anterior, el error standard o típico será: $\frac{4.23}{\sqrt{12}} = 1.22113$

Intervalo de confianza.
El intervalo de confianza es un rango de valores que incluye los parámetros de la población en un nivel determinado de probabilidad, generalmente el 95%.
El intervalo de confianza al 95% se calcula añadiendo y restando a la media de la muestra 1.96 veces (para fines prácticos: 2 veces) su error standard:

$$IC = \bar{x} \pm 2s_x$$

En el ejemplo anterior, el intervalo de confianza será: 12.42 ± 2 (1.22) = 12.42 ± 2.44
Esto significa que el 95% de los valores de la muestra deberían estar entre 9.98 y 14.86.

Diferencia entre desviación standard y error standard.
El error standard debe ser claramente diferenciado de la desviación standard en su interpretación. La desviación standard es una medida de la variabilidad de los datos en la muestra, mientras que el error standard es una medida de la certeza o de la incertidumbre en una muestra estadística.

Coeficiente de variación.
La desviación standard es una medida de la variación o dispersión absoluta de los valores de un grupo de datos numéricos, pero a veces se necesita conocer cuán importante es la variación o dispersión en relación con el conjunto de datos que se analiza, es decir, la relación que hay entre la desviación standard y el promedio de dichos datos. A esta relación se le denomina variación o dispersión relativa, pero es mejor conocida como coeficiente de variación y se simboliza por V o por CV. Se calcula dividiendo la desviación standard entre la media:

$$CV = \frac{s}{\bar{x}}$$

Aún cuando el coeficiente de variación es una fracción o parte de la unidad, por ejemplo, en el caso de las edades será 4.23 entre 12.42 = 0.3405797 o redondeando a 0.3406, esta fracción se transforma a porcentaje (1 = 100%), por lo tanto en el ejemplo será 34.06%. Se describe como un grupo de datos (edades) con una variación de 34.06%. Es de notar que el coeficiente de variación es independiente de las unidades usadas. Por esta razón, el coeficiente de variación es especialmente útil para comparar grupos da datos de diferentes unidades, por ejemplo si se quisiera comparar las edades del ejemplo anterior con las notas (unidad: puntos) que obtuvieron los estudiantes o con sus estaturas (unidad: centímetros), etc. Sin embargo, la principal desventaja del coeficiente de variación es que pierde su utilidad cuando la media es cercana a cero, pues cualquier cifra de la desviación standard dará por resultado un muy alto coeficiente de dispersión (porcentaje de variación), .lo cual puede no ser cierto.

Sesgo. Se denomina sesgo al grado de asimetría que tiene la curva de una distribución de datos numéricos. Cuando la curva de frecuencia, graficada a partir de la modificación ("suavización") del polígono de frecuencia, tiene la cola hacia la derecha del máximo valor, más larga que hacia la izquierda, se dice que la curva está sesgada a la derecha, que tiene sesgo a la derecha o que tiene sesgo positivo (Fig. 6). Si la cola hacia la izquierda es más larga que hacia la derecha se dice que está sesgada a la izquierda, tiene sesgo a la izquierda o tiene sesgo negativo (Fig. 7). Si las colas son iguales, se dice que la curva no tiene sesgo (Fig. 8).

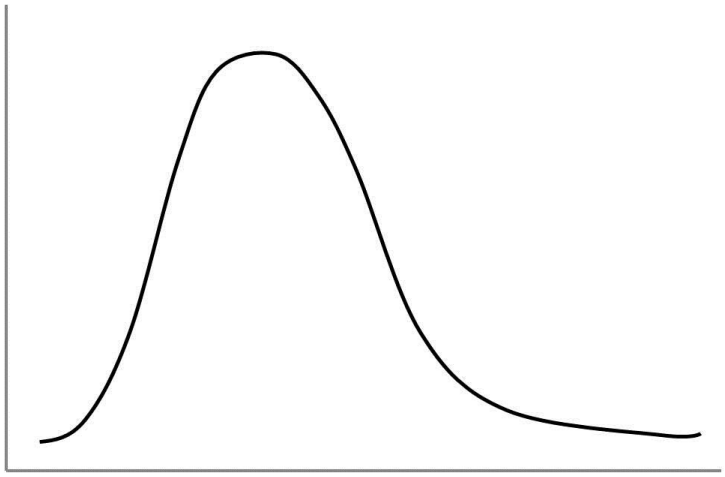


Fig. 6. Sesgo a la derecha,

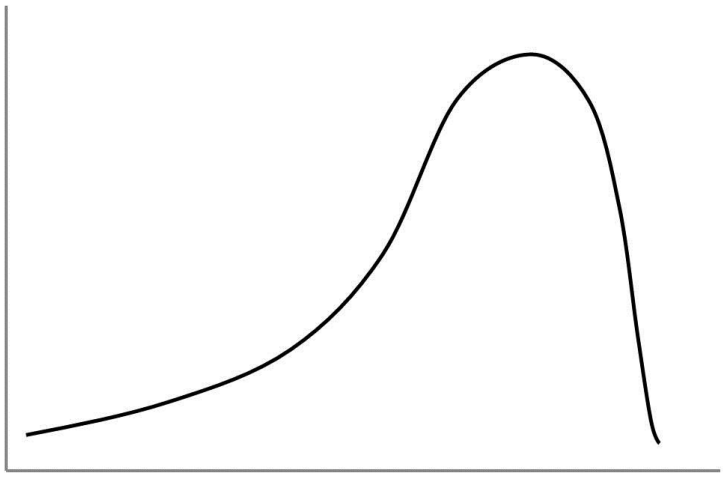


Fig. 7. Sesgo a la izquierda.

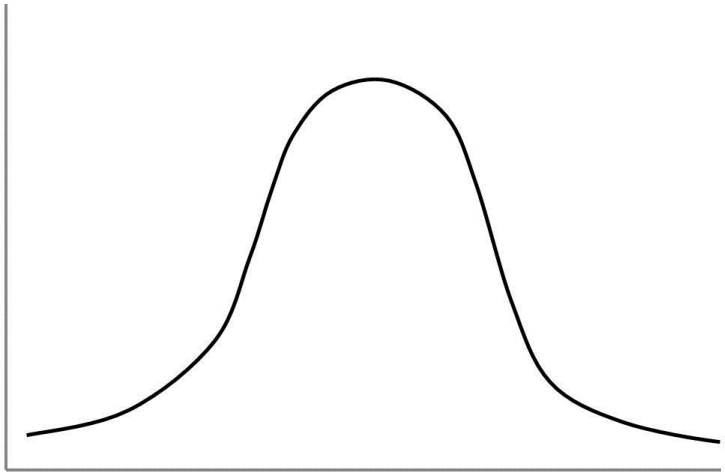


Fig. 8. Sin sesgo.

En una distribución normal, simétrica, hay solo una moda, no hay sesgo y la curva tiene forma de campana simétrica. En esta curva la media, la moda y la mediana coinciden en el mismo punto o valor central (Fig. 9).

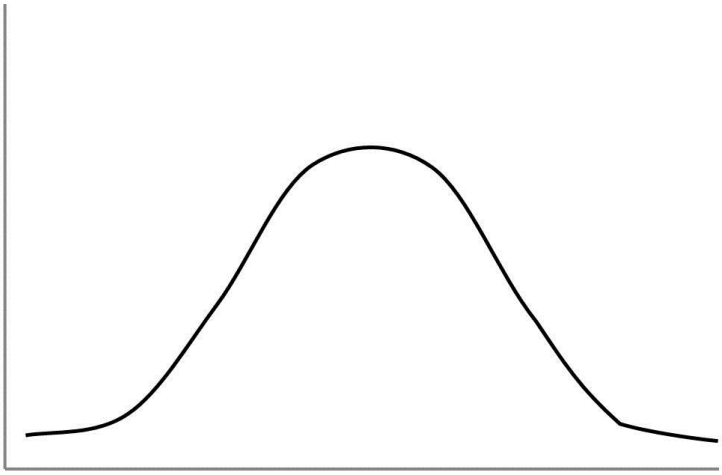


Fig. 9. Curva normal.

Kurtosis o curtosis. Se denomina curtosis al grado de "pico" o de "apuntamiento" que tiene la curva de una distribución de datos numéricos. Generalmente se refiere a una curva de distribución de tipo normal, es decir simétrica. Cuando la curtosis que tiene la curva es en forma de pico elevado, se le denomina leptocurtosis (Fig. 10). Si la curtosis de la curva es poco elevada, se le denomina mesocurtosis (Fig. 11) y cuando la curtosis es más bien aplanada que en forma de pico se le nombra platicurtosis (Fig. 12).

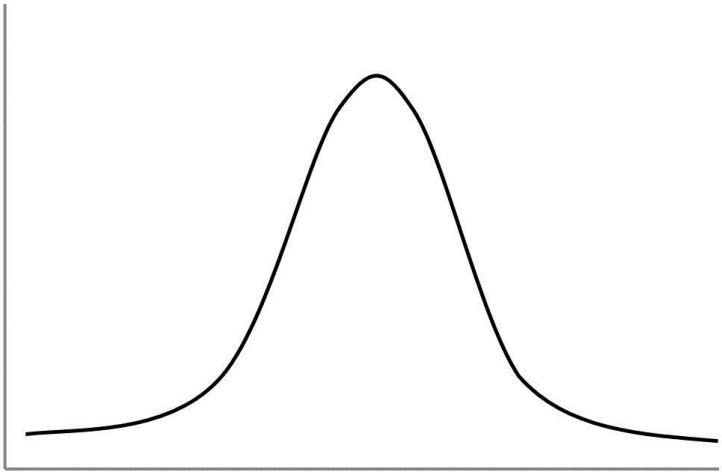


Fig. 10. Leptocurtosis.

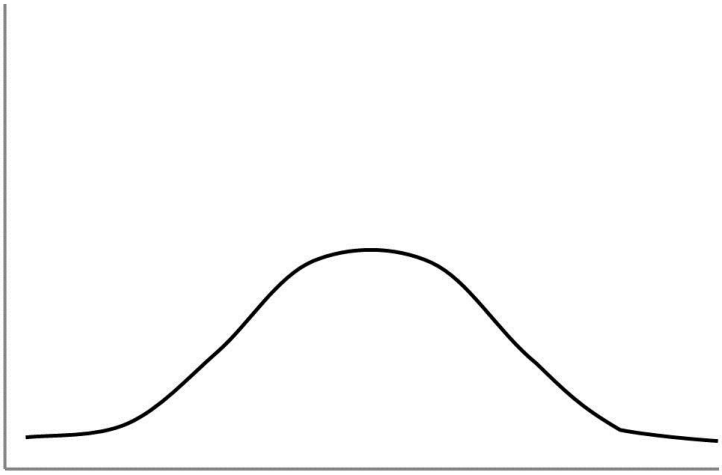


Fig. 11. Mesocurtosis.

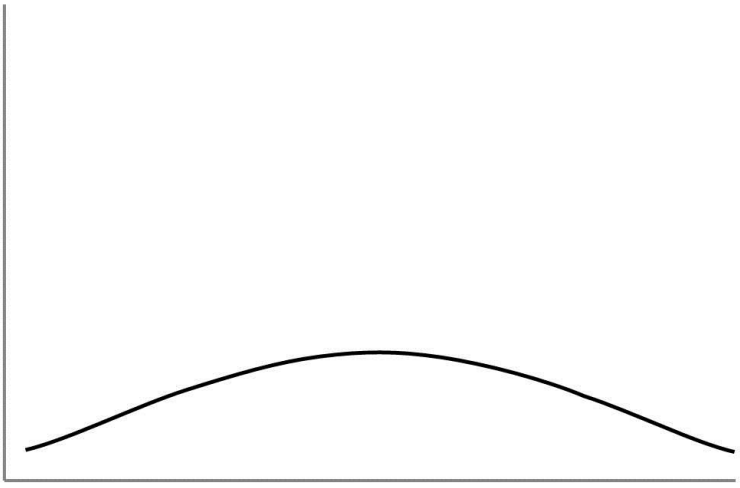


Fig. 12. Platicurtosis.

CAPÍTULO 5.

PROBABILIDAD. PROBABILIDAD CONDICIONAL. SIGNIFICANCIA ESTADÍSTICA

Probabilidad. La definición de probabilidad, según el Diccionario de la Real Academia Española es: 1. f. Verosimilitud o fundada apariencia de verdad. 2. f. Cualidad de probable, que puede suceder. 3. f. *Mat.* En un proceso aleatorio, razón entre el número de casos favorables y el número de casos posibles.
Posibilidad. 1. f. Aptitud, potencia u ocasión para ser o existir algo. 2. f. Aptitud o facultad para hacer o no hacer algo. 3. f. Medios disponibles, hacienda propia. U. m. en pl.
Por lo anterior vemos que desde el punto de vista general, la probabilidad es la cualidad de que algo pueda suceder u ocurrir. Mientras que desde el punto de vista matemático o estadístico por lo que aquí concierne, es un proceso aleatorio, razón entre el número de casos favorables, es decir, aciertos en lo que se estudia, trabaja o actividad cualquiera, y el número de casos posibles, es decir, las diferentes alternativas sobre ese estudio, trabajo o actividad. Quizá una forma más fácil de entenderlo es resumiéndolo al decir que probabilidad es la posibilidad de que algo ocurra. Véase arriba la definición de posibilidad del Diccionario de la Real Academia Española.
Debe entenderse que en ciencia no hay certezas, solo hay probabilidades. Aunque sea paradójico, la única certeza acerca de la ciencia es su incerteza (o incertidumbre).
En la investigación científica se trata de minimizar la posibilidad de encontrar alguna asociación cuando en realidad no exista, o de minimizar la probabilidad de no encontrar alguna asociación cuando en realidad sí exista.
No se puede eliminar la probabilidad de un error, pero con el uso de la estadística analítica sí se puede estimar la magnitud del error. La probabilidad de cometer un error depende del tamaño de la muestra. Mientras mayor el tamaño de la muestra, menor será la probabilidad de cometer el error.
Para mejor entender el término Probabilidad, supongamos que un evento *A* puede ocurrir o suceder en *l* maneras de un total de *n* igualmente posibles maneras. Lógicamente *l* forma parte de *n*. La probabilidad de ocurrencia del evento, denominada *éxito*, se expresa de la siguiente forma:

$$p = \Pr(E) = \frac{l}{n}$$

Pero, también hay la probabilidad o posibilidad de que el evento no ocurra. La no-ocurrencia del evento que se denomina *falla* o *fracaso*, se expresa como sigue:

$$q = \Pr(\text{no } E) = \frac{n-l}{n} = 1 - \frac{l}{n} = 1 - p = 1 - \Pr(E)$$

Por lo tanto, $p + q = 1$, es decir, $\Pr(E) + \Pr(\text{no } E) = 1$.
En otras palabras, la probabilidad de algo ocurra más la probabilidad de que ese algo no ocurra es 1, es decir, la probabilidad total siempre es 1. Esta probabilidad total o simplemente probabilidad, generalmente se expresa en términos porcentuales, es decir, 1 = 100%.

Pongamos algunos ejemplos. En todos los ejemplos se supone que las monedas o dados no están cargados. La probabilidad de que salga cara al lanzar una moneda es de 1 entre 2, ya que la moneda tiene solo dos posibilidades. p = ½ = 0.5 de la unidad, o sea 50%. De la misma manera, la probabilidad de no obtener cara es de q = 1-½ = ½ o sea 50%. En el caso de lanzar dos monedas, la probabilidad de que salgan dos caras es dos veces la posibilidad de cada una, es decir, p = ½ x ½ = ¼, o sea 0.5 x 0.5 = 0.25 ó 25%, por lo tanto la probabilidad de no obtener dos caras en un lanzamiento de dos monedas será de q = 1 – ¼ = ¾ o sea 0.75 ó 75%. Pero si lanzamos un dado (que tiene seis caras) para lograr el as (1) tendremos una posibilidad en seis, es decir p = 1/6 ó sea 0.166666… que redondeamos a 0.167 ó 16.7%. La probabilidad de no obtener el as será de q = 1 – 1/6 = 5/6 = 0.833333… que redondeamos a 0.833 ó 83.3%. Si lanzamos un dado deseando que salga el 1 ó el 6, tendremos dos posibilidades de las seis totales posibles, es decir, p = 2/6 = 0.333333… que redondeamos a 0.333 ó 33.3%. La probabilidad de no obtener el 1 ni el 6 es de q = 1 – 2/6 = 4/6 = 0.666666… que redondeamos a 0.667 ó 66.7%. Si lanzamos dos dados, deseando que salgan dos ases (1 y 1), la probabilidad será 1/6 x 1/6 ó 0.167 x 0.167, es decir, 1/36 = 0.027777… que redondeamos a 0.028 ó 2.8%. La probabilidad de no obtener los dos ases es de q = 1 – 1/36 = 35/36 = 0.9722222… que redondeamos a 0.972 ó sea 97.2%.

Veamos cuáles son las posibilidades en los juegos de azar. Supondremos que no hay trampas en dichos juegos. Comenzaremos por las carreras de caballos, mejor conocidas por juego del 5 y 6. Para ganar el mejor premio debemos "meter" seis caballos ganadores. En cada carrera hay 12 caballos, así que en cada carrera tenemos una probabilidad en doce posibilidades de ganar, o sea $p = 1/12 = 0.083333\ldots$ que redondeamos a 0.083 ó 8.3% y tenemos once posibilidades en doce de perder, o sea $q = 11/12 = 0.916666\ldots$que redondeamos a 0.917 ó 91.7%. Pero son seis carreras, es decir que la probabilidad de anotar los seis ganadores es de:
1/12 x 1/12 x 1/12 x 1/12 x 1/12 x 1/12, es decir, 1/2985984, ó 0.083 x 0.083 x 0.083 x 0.083 x 0.083 x 0.083 que resulta en la probabilidad de $p = 0.000000334897976680384087791495198902 61$, es decir, 0. 000033% que equivale a 3.3 aciertos en 10 millones de posibilidades o una posibilidad de acierto en 3 millones de posibilidades, lo que quiere decir que para tener completa seguridad de "pegar" los seis ganadores tenemos que jugar 3 millones de "cuadros" (como se llama al formulario de las apuestas) con diferentes combinaciones de caballos en cada carrera. En el juego de loterías (en Venezuela uno de los más populares es el llamado Kino), la situación es peor, pues el premio máximo es lograr 15 aciertos de 25 posibilidades que trae la papeleta o billete, pero sobre 100 posibilidades, es decir, números del 1 al 100, o sea $p = 25/100$ x 24/99 x 23/98 x 22/97 x 21/96 x 20/95 x 19/94 x 18/93 x 17/92 x 16/91 x 15/90 x 14/89 x 13/88 x 12/87 x 11/86, ya que si se acierta uno de los 25 números del billete en primer lugar de las 100 posibilidades, nos quedan 24 posibilidades en el billete para acertar el segundo número de las 99 posibilidades restantes y luego nos quedan 23 números en el billete para acertar como tercer número, alguna de las 98 posibilidades restantes y así sucesivamente hasta acertar los quince números, lo que resulta en $p = 0.000000000012902738311246017452166235133313$ ó 0.0000000012%, en otras palabras que para tener completa seguridad de acertar los quince números tendríamos que jugar cien mil millones de billetes con diferente combinación cada uno.

Probabilidad condicional. Se denomina probabilidad condicional a aquella que condiciona la ocurrencia de un evento a que otro evento ocurra o haya ocurrido. A dichos eventos se les llama eventos dependientes, ya que la ocurrencia de uno depende de la ocurrencia del otro. Por ejemplo, la probabilidad de un estudiante de obtener mejores notas en una asignatura está dada por la condición del número de páginas que sobre el tema estudie o haya estudiado el estudiante. La probabilidad de que una persona viaje de un lugar a otro está dada por la disponibilidad de asientos desocupados en el autobús. También puede ser que un evento depende de que otro no ocurra, por ejemplo la posibilidad de viajar en avión de una persona en lista de espera dependerá de que algún pasajero no llegue a tiempo para chequearse. Cuando la probabilidad de que un evento ocurra no depende de la ocurrencia o no ocurrencia de otro, se les denomina eventos independientes. Por ejemplo, la probabilidad de un estudiante de sacar 20 puntos en un examen de matemáticas es independiente de las páginas de literatura clásica que haya leído.
Las probabilidades pueden ser mutuamente excluyentes, es decir, que para que una condición se cumpla debe ser excluida la otra, por ejemplo, la probabilidad de una persona de viajar del sitio A al sitio B excluye toda posibilidad de que esa persona viaje al mismo tiempo al sitio C, entonces viajar a B y viajar a C son probabilidades mutuamente excluyentes. La probabilidad de una mujer estar en la menopausia excluye toda posibilidad de salir embarazada, entonces menopausia y embarazo son probabilidades mutuamente excluyentes.

Significancia o significación estadística.
Una prueba de significancia o significación estadística calcula la probabilidad de que los resultados de algún tipo de análisis, por ejemplo la diferencia entre las medias de dos o más grupos, o la relación de dependencia entre dos o más grupos se deba a la casualidad o al azar y por tanto no se puede inferir nada de ello.
Las pruebas de significancia estadística se basan en la lógica y el sentido común. El hecho de que una diferencia sea real y no por azar, se basa en tres criterios. Primero, en la magnitud de la diferencia observada, mientras mayor sea la diferencia, más probabilidades hay de que sea real y no por azar. Segundo, el grado de variación en los valores obtenidos en la investigación, es decir, si los valores caen en un rango muy amplio, las diferencias de las medias pueden deberse al azar y no realmente al estudio. Tercero, el tamaño de la muestra estudiada. Mientras mayor es el tamaño de la muestra, mayor es la probabilidad de que los resultados sean un fiel reflejo de los resultados de la población entera.
Se considera que el resultado de una investigación no es debido al azar , es decir, es estadísticamente significativo si el valor de P es menor de 5%, ($P < 0.05$), y se dice que es altamente significativo si el valor de P es menor de 1% ($P < 0.01$). Estos valores que se pueden considerar arbitrarios fueron establecidos por el gran estadístico inglés Sir Ronald Fisher, quien desarrolló la prueba de la "Razón de la varianzas", como veremos más adelante.

CAPÍTULO 6.

LA PRUEBA DE *t* DE STUDENT. GRADOS DE LIBERTAD

La prueba de *t* de Student.
La prueba de *t* de Student es una de la pruebas más comúnmente usada en investigación científica. Con ella podremos saber si una muestra es homogénea o heterogénea, es decir, si sus datos están estadísticamente distribuidos de manera uniforme o desuniforme. En este caso la forma más fácil de determinar su valor es dividiendo la media entre el error standard, mediante la siguiente fórmula.

$$t = \frac{\bar{x}}{s_{\bar{x}}}$$

Este valor se compara con un valor previamente calculado para una distribución teórica de los valores de *t*, que llamaremos valor de la tabla o valor "tabulado", para mejor comprensión del lector. La tabla de *t* de Student es una tabla de doble entrada, es decir, se puede entrar por las hileras donde están los grados de libertad o se puede entrar por las columnas donde están los niveles de probabilidad (**Anexo 1).**

Los **grados de libertad** para determinar su nivel de probabilidad en la tabla se determina por el número de observaciones o datos menos uno, es decir, GL = n – 1.

Sin embargo, su uso más frecuente es para determinar la significancia de las diferencias que hay entre dos medias. Para esto hay varias formas de calcular el valor de la *t* de Student. Veremos las más comunes y fáciles.

En primer término están las diferencias entre medias de dos grupos de igual tamaño, es decir, $n_1 = n_2$, por lo que se usa solamente el término n.
El valor de *t*, por ser producto de una raíz cuadrada, tiene signo positivo y negativo. El orden en que se utilicen las medias no es importante; sin embargo, para facilitar los cálculos, conviene utilizar como $\bar{x}_1$ a la media que tenga el mayor valor, de manera que al restarla de la de menor valor ($\bar{x}_2$) dará un resultado positivo.

$$t = \left(\bar{x}_1 - \bar{x}_2\right)\sqrt{\frac{n(n-1)}{\left(\sum n_1^2 - \frac{\left(\sum x_1\right)^2}{n}\right) + \left(\sum n_2^2 - \frac{\left(\sum x_2\right)^2}{n}\right)}}$$

Si las muestras son de igual tamaño, es decir, de igual número de observaciones, los grados de libertad serán 2n - 2, por ejemplo, si ambas muestras tienen 15 observaciones, los grados de libertad son GL = (2 x 15) – 2 = 28.

En segundo lugar, están las diferencias entre medias de dos grupos de diferente tamaño, es decir, $n_1 \neq n_2$.

$$t = \left(\bar{x}_1 - \bar{x}_2\right)\sqrt{\frac{n_1.n_2(n_1 + n_2 - 2)}{(n_1 + n_2)\left[\left(\sum x_1^2 - \frac{\left(\sum x_1\right)^2}{n}\right) + \left(\sum x_2^2 - \frac{\left(\sum x_2\right)^2}{n}\right)\right]}}$$

En este caso los grados de libertad son GL = $n_1 + n_2 - 2$, ya que a cada muestra corresponden n - 1 grados de libertad, o sea $(n_1 - 1) + (n_2 - 1)$, por ejemplo, si una muestra tenía 15 observaciones y la otra 12 observaciones, los grados de libertad serán 15 + 12 – 2 = 25
La prueba de significancia se realiza comparando el valor de t obtenido en el experimento o investigación con un valor establecido en una tabla de los valores correspondientes a una distribución ideal de *t*. Esta es una tabla de doble entrada (tabla 4) donde se puede observar en el encabezado de las columnas el grado de probabilidad (algunas veces expresado como fracción de la unidad y otras veces como porcentaje) y en el

inicio de las filas el número de grados de libertad. Si deseamos saber si la diferencia entre dos muestras es estadísticamente significativa tendremos que buscar en la fila o hilera correspondiente a los grados de libertad del experimento (por ejemplo, 28) el valor en columna de $P = 0.05$ ó 5%. Si el valor calculado en el experimento es igual o mayor que el observado en la tabla se dice que hay diferencias estadísticamente significativas entre las dos medias o que los dos grupos son estadísticamente diferentes, ya que hay 95% o más de probabilidades de que las diferencias se deban al experimento y 5% o menos de que se deban al azar. Si es así podremos continuar buscando en las columnas correspondientes a probabilidades menores, tal como 0.01, 0.005, 0.001 o aún menores, aunque un valor de 0.001 (uno en mil) se considera, generalmente suficiente. Si el valor calculado es menor que el valor observado en la tabla se dice que no hay diferencias estadísticamente significativas entre las dos medias o que los dos grupos son estadísticamente iguales. Mientras mayor el valor de *P*, menor será la diferencia entre los dos grupos. Por esta razón si estamos probando la hipótesis de que dos medicamentos o dos fertilizantes son iguales, nos interesa que el valor de la *t* calculada sea lo menor posible para que la *P* sea lo mayor posible y por tanto las diferencias sean insignificantes desde el punto de vista estadístico.
Por ejemplo, si comparamos las notas de los estudiantes de dos escuelas secundarias, las cuales son (máxima puntuación 20 puntos).

Tabla 4. Notas de los alumnos de dos liceos, con grupos de igual tamaño.

Alumno	Liceo A (x_1)	x_1^2	Liceo B (x_2)	x_2^2
1	14	196	7	49
2	13	169	18	324
3	15	225	19	361
4	16	256	10	100
5	14	196	12	144
6	18	324	20	400
7	16	256	9	81
8	15	225	13	169
9	19	361	10	100
10	14	196	20	400
11	16	256	18	324
12	17	289	10	100
13	15	225	20	400
14	18	324	8	64
Total	220	3498	194	3016
Media	15.71		13.86	

$$t = (15.71 - 13.86)\sqrt{\frac{14(14-1)}{\left(3498 - \frac{220^2}{14}\right) + \left(3016 - \frac{194^2}{14}\right)}}$$

$$t = 1.85\sqrt{\frac{182}{(3498 - 3457.14) + (3016 - 2688.29)}}$$

$$t = 1.85\sqrt{\frac{182}{40.86 + 327.71}}$$

$$t = 1.85\sqrt{\frac{182}{368.71}}$$

$$t = 1.85\sqrt{0.49361}$$

$t = 1.85 \text{ x } 0.7026$

t = 1.2998

Valor de t en la tabla de t para (2 x 14) – 2 = 26 grados de libertad y $P_{.05}$ es t = 2.056, es decir, que el valor de la t del estudio es menor que la de la tabla, por lo tanto no hay diferencias estadísticamente significativas entre las dos medias o que las notas de los estudiantes de los dos liceos son estadísticamente iguales.

Un ejemplo de dos grupos de diferente tamaño sería, las notas del primer liceo de comparadas con la de otro liceo con 16 alumnos.

Tabla 5. Notas de alumnos de dos liceos, con grupos de diferente tamaño.

Alumno	Liceo A (x_1)	$x,^2$	Liceo B (x_2)	x_2^2
1	14	196	8	64
2	13	169	12	144
3	15	225	11	121
4	16	256	17	289
5	14	196	10	100
6	18	324	8	64
7	16	256	6	36
8	15	225	13	169
9	19	361	18	324
10	14	196	7	49
11	16	256	12	144
12	17	289	14	196
1	15	225	9	81
3	18	324	10	100
14	14	196	13	169
15	-	-	8	64
16	-	-	11	121
Total	220	3498	187	2235
Media	15.712		11.69	139.69

$$t = (15.71-11.69)\sqrt{\frac{14x16(14+16-2)}{(14+16)\left[(3498-3457.14)+\left(2235-\frac{187^2}{16}\right)\right]}}$$

$$t = (15.71-11.69)\sqrt{\frac{14x16(14+16-2)}{(14+16)[(3498-3457.14)+(2235-2185.56)]}}$$

$$t = 4.02\sqrt{\frac{224x28}{30\left[(3498-3457.14)+\left(2235-\frac{187^2}{16}\right)\right]}}$$

$$t = 4.02\sqrt{\frac{224x28}{30[40.86+49.44]}}$$

$$t = 4.02\sqrt{\frac{6272}{30x90.30}}$$

$$t = 4.02\sqrt{\frac{6272}{2709}}$$

$$t = 4.02\sqrt{2.3152}$$

t = 402 x 1.5216

t = 6.1167

Valor de t en la tabla de t para (14 + 16) – 2 = 28 grados de libertad y $P_{.05}$ es t = 2.048, $P_{.01}$ es t = 2.763 y para $P_{.001}$ es t = 3.674, es decir, que el valor de la t del estudio es mayor que el de la tabla, para probabilidad de 95% por lo tanto hay diferencias estadísticamente significativas entre las dos medias o que las notas de los estudiantes de los dos liceos son desde el punto de vista estadístico significativamente diferentes; también para probabilidad de 99%, es decir que las diferencias son estadísticamente *altamente significativas*, e igualmente para probabilidad de 99.9%, lo que indica que las diferencias son estadísticamente *muy altamente significativas.*

CAPÍTULO 7.

LA PRUEBA DE JI CUADRADO. TABLA DE CONTINGENCIA.

La prueba de ji cuadrado, χ^2.

Ji es el nombre en castellano de la letra griega χ, su nombre en inglés es chi, de allí que se haya difundido esta forma de llamar a la prueba de ji cuadrado. La prueba de ji cuadrado o ji cuadrada, es una de las pruebas de hipótesis más comúnmente usadas en investigación científica, al igual que la *t* de Student. La prueba de ji cuadrado se usa, generalmente, para variables categóricas.

La prueba sirve para detectar si las diferencias encontradas en algunas condiciones o características de algunos grupos son estadísticamente significativas.

La prueba consiste en comparar los valores observados en la investigación con aquellos esperados que deberían ocurrir o con los valores de otros grupos de comparación. También sirve para determinar si hay dependencia entre dos grupos de variables observadas, es decir, si la presencia o ausencia de uno depende de la presencia o ausencia del otro.

Hay varias fórmulas para calcular el valor de ji cuadrado que se simboliza así: χ^2.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Lo cual significa que el valor del ji cuadrado es igual a la sumatoria de cada uno de los valores observados (O) menos cada valor esperado (E) elevado al cuadrado y dividido entre cada uno de los valores esperados.

Este valor de χ^2 calculado a partir de nuestros datos se compara con el valor de χ^2 en una tabla de valores predeterminados **(Anexo 2)**. Se busca de acuerdo con los grados de libertad. Es una tabla de doble entrada como la d et de Student. Los niveles de probabilidad (P) están en las columnas y los grados de libertad están en las filas o hileras. Cuando un valor de χ^2 calculado es igual o mayor al del a tabla, indica que hay diferencias estadísticamente significativas entre los valores observados y los esperados. Un valor de χ^2 menor al del a tabla indica que los valores observados no son estadísticamente diferentes de los esperados.

Los grados de libertad están dados por el número de operaciones que se llevan a cabo menos uno, es decir, por el número de datos observados menos uno (n – 1), por ejemplo, si tenemos un grupo de 8 niños en un concurso y deseamos saber si conforman la edad en la que deberían teóricamente estar, para evitar que sean mayores o menores a la dificultad del concurso, podemos usar la prueba de ji cuadrado, así:

Edad observada: 5, 7, 7, 8, 9, 10, 10 y 11 años.
Edad para entrar al concurso: 10 años.

Estos datos podemos incluirlos en una tabla donde en una hilera van los datos observados y en otra hilera van los valores esperados:

Tabla 6. Edad de niños para entrar a un concurso.

Observados	5	7	7	8	9	10	10	11
Esperados	10	10	10	10’	10	10	10	10

Esta se llama una **tabla de contingencia** de 1 hilera (los valores observados) x 8 columnas o simplemente de 1 x 8.

$$\chi^2 = \frac{(5-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(7-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(11-10)^2}{10}$$

χ^2 = (25/10)+(9/10)+(9/10)+(4/10)+(1/10)+(0/10)+(0/10)+(1/10)

χ^2 = 2.5+0.9+0.9+0.4+0.1+0.0+0.0+0.1

χ^2 = 4.9

Grados de libertad: 8-1 = 7

Al comparar nuestros valores con los valores de la tabla de χ^2 que son para :

$P_{.95}$ = 14.1
$P_{.99}$ = 18.5
$P_{.995}$ = 20.3

Se observa que las edades de los niños no son estadísticamente diferentes a los esperados para entrar al concurso, es decir, son similares desde el punto de vista estadístico.
El valor 4.9 cae ente $P_{.50}$ = 6.35 y $P_{.25}$ = 4.25, es decir que apenas muy cerca de 25% ($P_{.25}$ = 4.25) es la probabilidad de que sean diferentes las edades observadas con respecto a las esperadas. Por lo tanto, 75% de probabilidad de que sean similares.

Hay ocasiones donde los valores observados ocupan varias hileras, por ejemplo, si en el caso anterior (tabla **A** de 1 x 2) tuviésemos el grupo de niños separados por sexo serían dos hileras (2 x 8), pero podríamos tenerlos clasificados según su procedencia y serían varias, quizá muchas, las hileras (*h* x 8 en el ejemplo), pero podrían ser más las condiciones de la columnas, por ejemplo, comparar las procedencias de un grupo de pacientes con sus diferentes patologías, en ese caso sería una tabla de contingencia de *h* hileras por *c* columnas o *h* x *c*. Por ejemplo, en un caso de una tabla de 7 hileras por 3 columnas, o de 7 x 3, la notación sería:

Tabla 7. Ejemplo de tabla de contingencia de 7 x 3.

	Columna 1	Columna 2	Columna 3	Total hileras
Hilera 1	h1c1	h1c2	h1c3	Total h1
Hilera 2	h2c1	h2c2	h2c3	Total h2
Hilera 3	h3c1	h3c2	h3c3	Total h3
Hilera 4	h4c1	h4c2	h4c3	Total h4
Hilera 5	h5c1	h5c2	h5c3	Total h5
Hilera 6	h6c1	h6c2	h6c3	Total h6
Hilera 7	h7c1	h7c2	h7c3	Total h7
Total columnas	Total c1	Total c2	Total c3	TOTAL

Los grados de libertad se calculan multiplicando los grados de libertad h – 1 para las hileras por los *c* – 1 para las columnas, GL = (*h*-1) (*c*-1).

Hay casos donde no se conoce el valor de los datos esperados, es decir, no hay valores prefijados, por lo cual hay que calcularlos. Por ejemplo, se quiere saber la relación que hay entre la profesión de un grupo de mujeres y sus disfunciones sexuales cuyos valores observados están en la tabla siguiente. Como es lógico no hay valores prefijados, es decir, esperados para este grupo de mujeres, por lo cual se deben calcular. La tabla contingencia del ejemplo sería:

Tabla 8. Ejemplo del cálculo de valores esperados en tabla de contingencia.

	Vaginismo	Dispareunia	Anorgasmia	Total hileras
Abogadas	12	3	7	22
Arquitectas	15	5	8	28
Bioanalistas	7	2	9	18
Historiadoras	10	6	6	22
Ingenieras	20	4	10	34
Médicas	10	8	12	30
Odontólogas	5	5	11	21
Total columnas	79	33	63	175

El cálculo de los valores esperados se realiza de la siguiente manera:

Para cada valor observado, el cálculo del valor esperado se realiza multiplicando el total de la hilera correspondiente al valor observado por el total de la columna correspondiente a ese valor y dividiendo el resultado entre el total general. Supongamos que queremos calcular el valor esperado para médicas con dispareunia, cuyo valor observado es 8. Multiplicamos el total de esa hilera (Total h6) por el total de esa

columna (Total c2) y ese resultado lo dividimos entre el total general (TOTAL). En este ejemplo, el valor esperado se calcula así: 30 (Total h6) x 33 (Total c2) = 990, cifra que dividimos entre 175 (TOTAL), 990/175 = 5.657 que redondeamos a 5.7. Así obtenemos los valores esperados para cada uno de los valores observados. Luego procedemos en la forma antes explicada de observado menos esperado al cuadrado, etc. hasta obtener el valor calculado de χ^2 y lo comparamos con el valor de la tabla de χ^2 y decidimos si hay o no hay diferencias estadísticamente significativas entre los valores observados y los esperados.

Otro de los usos muy comunes de la prueba de χ^2 es la llamada Prueba de Independencia, donde se presenta la hipótesis nula de que los datos confrontados de dos condiciones son independientes, para demostrarla (cuando en realidad son independientes las condiciones) o rechazarla (cuando son dependientes las condiciones).

Se construye la tabla de contingencia de 2 x 2 para realizar la prueba de independencia, donde cada condición se le atribuye presencia y no presencia u ocurrencia y no ocurrencia.

Tabla 9. Tabla de contingencia de 2 x 2 para Prueba de Independencia.

	Condición B presente (+)	Condición B ausente (-)	Total hileras
Condición A presente (+)	Condiciones A y B presentes (+ +) a	Condición A presente y B ausente (+ -) b	a + b
Condición A ausente (-)	Condición A ausente y B presente (- +) c	Condiciones A y B ausentes (- -) d	c + d
Total columnas	a + c	b + d	a + b + c +d = n

La fórmula para calcular el ji cuadrado en este caso es:

$$\chi^2 = \frac{n\left(|ad-bc|-\frac{1}{2}n\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Donde el símbolo $|...|$ significa el valor absoluto de la cifra incluida, es decir, se toma como positivo sin importar si es negativo o positivo, así que si *ad* es mayor que *bc* será positivo, pero si es menor será negativo y aún así se le considera positivo para el cálculo.
Los grados de libertad serán, como para cualquier tabla de contingencia: (*h*-1)(*c*-1): en este caso son: (2-1)(2-1) = 1 x 1 = 1 grado de libertad.
Si el valor de χ^2 calculado con nuestros datos es igual o mayor que el de la tabla, rechazamos la hipótesis nula de que ambas condiciones son independientes y aceptamos que son dependientes. Si el valor calculado es menor que el de la tabla, aceptamos la hipótesis de que ambas condiciones son independientes y rechazamos la hipótesis de que son dependientes.
Veamos el siguiente ejemplo: deseamos saber si el consumo de alcohol está asociado a los accidentes, es decir, si hay dependencia del consumo sobre los accidentes de tránsito, es decir, si hay dependencia del consumo sobre los accidentes. La hipótesis nula es que ambas condiciones son independientes. Se entrevistaron a 300 personas en total. El número de personas que consumió alcohol y tuvo accidentes es de 120, el número de personas que consumió alcohol y no tuvo accidentes es de 70, el número de personas que no consumió y tuvo accidentes es de 20 y los que no consumieron alcohol y no tuvieron accidentes es de 90.

Tabla 10. Tabla de contingencia de 2 x 2 para Prueba de Independencia.

	Consumieron alcohol (+)	No consumieron alcohol (-)	Total hileras
Tuvieron accidente (+)	120	20	140
No tuvieron accidente (-)	70	90	160
Total columnas	190	110	300

El resultado del ji cuadrado es:

$$\chi^2 = \frac{300\left(|10800-1400|-\frac{1}{2}300\right)^2}{(140)(160)(190)(110)}$$

$$\chi^2 = \frac{25668750000}{468160000}$$

χ^2 = 54.829017

Que redondeamos a 54.829 y comparamos con los valores en la tabla de χ^2, para 1 grado de libertad, que son:

$P_{.95}$ = 3.84
P_{99} = 6.63
P_{995}= 7.88

Como el χ^2 calculado es mayor que el de la tabla, incluso al 99.5% de probabilidad, rechazamos la hipótesis nula de que las condiciones son independientes y aceptamos la hipótesis alternativa de que en el grupo estudiado, los accidentes de tránsito dependen del consumo de alcohol, es decir, concluimos que hay una gran dependencia entre el consumo de alcohol y los accidentes de tránsito.

Veamos otro ejemplo donde deseamos saber si la presencia de un parásito en las heces de un grupo de personas depende del agua de una fuente sin tratar, de donde se nutren dichas personas. Supongamos que se examinaron las heces de 300 personas en total, de las cuales 165 tomaron agua de la fuente y 80 presentaron parásitos, mientras que de las 135 que no tomaron agua de la fuente, 65 no presentaron parásitos.

Tabla 11. Tabla de contingencia de 2 x 2 para Prueba de Independencia.

	Tomaron agua de la fuente (+)	No tomaron agua de la fuente (-)	Total hileras
Presentaron parásitos (+)	++ 80	+- 70	150
No presentaron parásitos (-)	-+ 85	-- 65	150
Total columnas	165	135	300

El resultado del ji cuadrado es:

$$\chi^2 = \frac{300\left(|5200-5950|-\frac{1}{2}300\right)^2}{(150)(150)(165)(135)}$$

En este caso 5200-5950 = - 750 es negativo pero se toma como positivo (valor absoluto) para los cálculos.

$$\chi^2 = \frac{108000000}{501187500}$$

χ^2 = 0.215488

Que redondeamos a 0.215 y comparamos con los valores en la tabla de χ^2, para 1 grado de libertad, que son:
$P_{.95}$ = 3.84
P_{99} = 6.63
P_{995}= 7.88
Como el χ^2 calculado es menor que el de la tabla, incluso al 95% de probabilidad, concluimos que no hay ningún tipo de dependencia entre el consumo de agua de la fuente sin tratar y la presencia de parásitos en las heces, es decir, que el agua no es la fuente del parasitismo intestinal de las personas, en este caso.

CAPÍTULO 8.

ANÁLISIS DE REGRESIÓN. COEFICIENTE DE REGRESIÓN. INTERCEPTO. COVARIANZA DE *x.y*. PRUEBA DE SIGNIFICANCIA DE *b*.

Análisis de regresión

En muchas situaciones, investigaciones, etc., se observa que hay cierta relación de dependencia entre dos o más variables. Algunas veces esa dependencia es una forma de causa-efecto, es decir, una variable causa un efecto en la otra variable. Esa relación puede ser de dependencia directa, es decir, que cuando aumenta una variable, la otra también aumenta, o puede ser indirecta o inversa que quiere decir que cuando una variable aumenta la otra disminuye. Pero la relación también tiene un grado de intensidad, es decir, que puede ser una relación muy estrecha con fuerte dependencia de una variable sobre la otra, o puede ser muy débil, es decir, que la variación de una casi no afecta a la otra, hasta llegar al extremo de que no la afecta en absoluto, es decir, que en vez de dependencia hay independencia absoluta entre las dos variables. Esta relación es conveniente medirla para poder expresarla y analizarla, para lo cual se usa la medición matemática a través de lo que se denomina análisis de regresión.

El nombre de análisis de regresión se estableció cuando un investigador norteamericano realizó un estudio sobre la herencia de ciertas características y medidas corporales, por ejemplo, estatura, expansión entre brazos abiertos, longitud del radio y del cúbito, etc., entre padres e hijos, con la hipótesis de que con el tiempo de una generación a la siguiente los caracteres iban en aumento. Para lo cual tomó varias decenas de miles de datos, confrontó cada característica del padre con su homóloga en los hijos. Los analizó mediante la aplicación de la ecuación matemática de la línea recta a cada característica estudiada. En general se demostró la hipótesis, pero en algunas medidas la hipótesis no se corroboró sino por el contrario se demostró que las características disminuían, iban en retroceso, es decir, iban en regresión, de allí el nombre del análisis.

El análisis de regresión es un método estadístico usado para determinar el grado de dependencia que existe entre dos variables. Se parte del supuesto de que las dos variables tienen una relación de dependencia. Se usa la aplicación de la ecuación matemática de la línea recta que relaciona las dos variables en estudio. Conviene destacar aquí que el análisis de regresión también puede realizarse para más de una variable independiente, en este caso se le denomina análisis de regresión múltiple.

Las dos variables se denominan: variable dependiente, simbolizada por *y*, que es aquella que se modifica cuando la otra o variable independiente, simbolizada por x, se modifica o varía. De su análisis se obtiene un valor o parámetro estadístico denominado el coeficiente de regresión, simbolizado por *b*. Con esos datos se determina el valor estimado ("teórico"), simbolizado por $\hat{Y}$ que debe corresponder a cada valor realmente observado de *y*.

La ecuación de la recta es: $\hat{Y} = a + bx$

Donde a es el **intercepto** o valor sobre el eje de las ordenadas (*y*) desde el origen (0) hasta la interceptación de la línea recta o de regresión con dicho eje, en la representación gráfica de los datos. Más adelante desarrollaremos las explicaciones sobre cada uno de los símbolos usados.

El **coeficiente de regresión** *b* indica el grado de inclinación que tendrá la línea recta en relación con la línea horizontal o eje de las abscisas. También se le llama pendiente de la recta. Geométricamente, es la tangente (cateto opuesto sobre cateto adyacente) del ángulo que forma la línea recta con el eje de las abscisas.

Su cálculo estadístico se realiza mediante la ecuación siguiente

$$b = \frac{Covarianza\ de\ x.y}{Varianza\ de\ x}$$

La **covarianza de *x.y*** se calcula así:

Suma de cuadrados de *x.y* dividido entre los grados de libertad. Los grados de libertad son el número de observaciones menos uno, GL = n – 1.

La suma de cuadrados para *x.y* se calcula así:

Se suman todos los datos de x, se le llama sumatoria de x y se simboliza por Σx. Se suman todos los datos de y. Se le llama sumatoria de y, se simboliza por Σy.

Se ordenan los datos de la variable independiente en orden ascendente, con su correspondiente valor de la variable dependiente a un lado, es decir, una columna para los datos de x y otra para los datos de y. Se calcula la media de la variable independiente ($\bar{x}$).
Luego se multiplica, para cada observación, su valor de *x* por su respectivo valor de *y*. Se suman todos esos resultados que llamaremos sumatoria de *x.y*. La simbolizaremos por Σ*x.y*.
Luego obtenemos el factor de corrección (FC) para *x.y*, así:

$$FC = \frac{\left(\sum x.\sum y\right)}{n}$$

Luego se resta el FC de la Σ*x.y*. Ese es la suma de cuadrados para x.y. Se simboliza así:

$$SCx.y = \sum x.y - \frac{\left(\sum x.\sum y\right)}{n}$$

La varianza de x se calcula así:
s^2 = Suma de cuadrados (corregida) dividido entre los grados de libertad (n – 1).
La suma de cuadrados para x se calcula así:
Se eleva al cuadrado cada valor de x. Se suman los valores. Se le llama sumatoria de x^2 y se simboliza así: $\sum x^2$. A este valor se le resta el FC para x, que es

$$FC = \frac{\left(\sum x\right)^2}{n}$$

Así obtenemos la suma de cuadrados para x:

$$SCx = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

Para el cálculo del coeficiente de regresión, se puede obviar el paso de dividir cada suma de cuadrados entre los grados de libertad, pues como son igual para el dividendo y el divisor, no es necesario efectuar la operación y el coeficiente *b* se obtiene directamente dividiendo la suma de cuadrados de *x.y* entre la suma de cuadrados de x, como sigue:

b = Suma de cuadrados de x.y / Suma de cuadrados de x

$$b = \frac{\sum x.y - \frac{\sum x.\sum y}{n}}{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}$$

Hay que destacar que el valor de *b* puede tomar cualquier valor desde -∞ hasta +∞, pasando por 0. Un valor de 0 nos indica que no hay dependencia entre las variables, es decir, son totalmente independientes. Mientras *b* aumenta se hacen más dependientes. La línea de regresión se hace más inclinada (aumentas u pendiente).
En la ecuación de la recta: $\hat{Y} = a + bx$, ya conocemos el valor de *b* y para calcular el valor (desconocido ahora) de a, podemos sustituir en dicha ecuación, los valores de de $\hat{Y}$ y de *x*, desconocidos, por valores conocidos como son las medias de las variables dependiente $\bar{y}$ e independiente $\bar{x}$. Así tendremos:

$\bar{y}$ = a + b $\bar{x}$. Despejamos a y obtenemos: a = $\bar{y}$ - b $\bar{x}$.
Así hemos resuelto la ecuación de la regresión.

Para obtener gráficamente los resultados del cálculo de la regresión, debemos señalar ("plotear") en un papel, preferiblemente cuadriculado, los valores de x y de y de cada dato observado. Así obtendremos lo que se denomina la nube de dispersión de los datos u observaciones. Ahora solo queda calcular el valor de $\hat{Y}$ para el dato menor y para el dato mayor de las observaciones y sustituir los valores observados por esos puntos en el papel y unirlos con una recta, por ejemplo, con una regla, lo que nos dará la línea de regresión para el

juego de datos numéricos que hemos tomado de nuestra investigación o estudio. Las figuras 13 y 14 muestran líneas de regresión positiva y negativa, respectivamente y nube de dispersión de los datos.

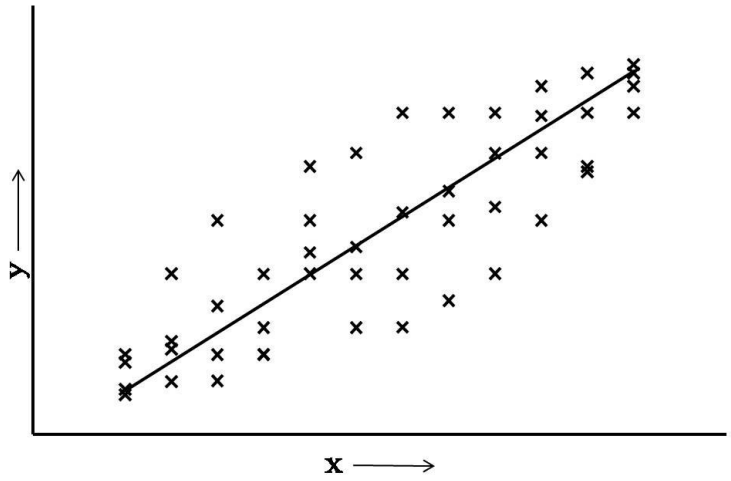


Fig. 13. Línea de regrresión (positiva) y nube de dispersión de los datos.

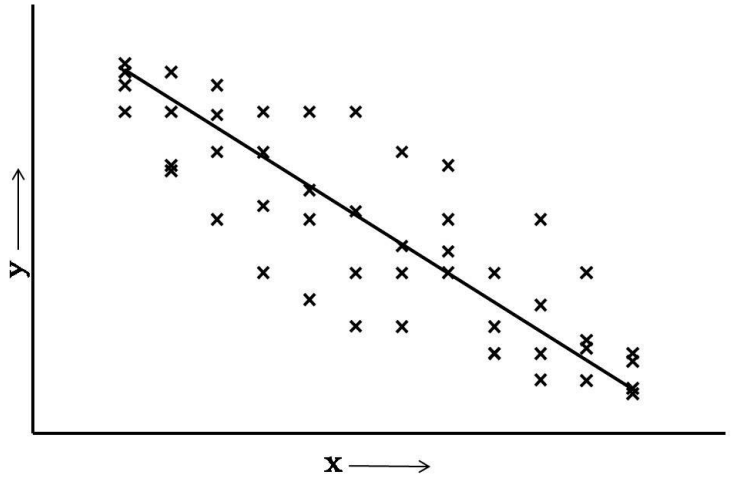


Fig. 14. Línea de regrresión (negativa) y nube de dispersión de los datos.

Ejemplo:

En un grupo de niños y adolescentes se registra la edad en comparación con las páginas de un libro que puede leer en un día. Los datos directamente registrados son:

Edad (x)	Nº pág. leídas/día (y)
7	30
3	20
9	33
12	36
5	25
14	40
8	31
13	39
17	45
4	24
18	47
16	42
5	27
10	35
6	28

Se ordenan en orden ascendente de la edad o variable independiente y se eleva al cuadrado cada dato de la variable independiete (x) y se multiplica cada juego de datos para cada individuo (x.y):

Edad (x)	Nº pag. leídas/día (y)	x^2	x.y
3	20	9	60
4	24	16	96
5	25	25	125
5	27	25	135
6	28	36	168
7	30	49	210
8	31	64	248
9	33	81	297
10	35	100	350
12	36	144	432
13	39	169	507
14	40	196	560
16	41	256	656
17	45	289	765
18	47	324	846

$\sum x = 147$ $\bar{x} = 9.8$ $\sum x^2 = 1783$ $(\sum x)^2 = 21609$ FC = 21609/15 FC = 1440.6
GL (n-1) = 14 $s^2(x) = (1783 - 1440.6)/14$ $s^2 = 24.457$ $s = 4.945$
$\sum y = 501$ $\bar{y} = 33.4$ $\sum y^2 = 17621$ $s^2 = 63.400$ $s = 7.9622$
$\sum x \sum y = 73647$
FC = 73647/15 FC = 4909.8
$\sum x.y = 5455$
$(\sum x.y - FC)/GL = (5455 - 4909.9)/14$ 545.2/14 = 38.942

$$b = \frac{\sum xy - FC}{s^2} \quad b = \frac{(5455 - 4909.8)/14}{24.457} \quad b = \frac{545.2/14}{24.457} \quad b = \frac{38.942}{24.457} \quad b = 1.592$$

El cálculo de a es: 33.4 = a + 1.592 x 9.8 33.4 = a + 15.6016 a = 33.4 – 15.6016 a = 17.7984

Prueba de significancia de *b*.
El obtener el valor del coeficiente de regresión *b* para un grupo de datos no indica que ese coeficiente sea realmente debido a una dependencia cierta de las variables, es decir, puede deberse al azar. Para probar su certeza estadística se debe aplicar la prueba de significancia de t de Student. Esta prueba se realiza como sigue.

Suma de cuadrados para desviaciones de *x.y* (x por y):

$$SCd_{y.x} = SCy - \frac{(SCx.y)^2}{SCx}$$

$$SCd_{x.y} = \left(\sum y^2 - \frac{(\sum y)^2}{n} \right) - \left[\frac{\left(\sum x.y - \frac{\sum x \sum y}{n} \right)}{\sum x^2 - \frac{(\sum x)^2}{n}} \right]$$

Varianza de las desviaciones de la regresión:

$$s^2_{y.x} = \frac{SCd_{y.x}}{n-2}$$

Varianza del coeficiente de la regresión:

$$s_b^2 = \frac{s_{y.x}^2}{SCx}$$

Desviación standard del coeficiente de la regresión:

$$s_b = \sqrt{s_b^2}$$

t de Student para la regresión:

$$t = \frac{b}{s_b}$$ Grados de libertad = n-2

En el ejemplo de arriba, el valor calculado de t es de 24.057 con $p \leq 0.000$, es decir, la prueba de t indica que es muy poca la probabilidad (menos de 1 en 10000) de que los resultados de b se deban al azar.

Otras formas de análisis de regresión.

El análisis de regresión antes expuesto se refiere a lo que se ha denominado análisis de regresión simple, determinado por la ecuación de la línea recta ($y = a + bx$), pero hay muchas otras formas de realizar el análisis de regresión, dependiendo de la distribución de los datos, tanto de la variable dependiente como de la o las variables independientes.

Cuando se tiene varias variables independientes, pero que se suponen influencia en forma directamente proporcional a la variable dependiente, se realiza el análisis de regresión múltiple, donde se sigue la ecuación de la línea recta, pero con varios coeficientes de regresión, uno para cada una de las variables independientes, por ejemplo, si se quisiera saber la relación de dependencia que hay entre la edad, la estatura, el peso y la temperatura corporal en personas en desarrollo, es decir, en edades entre 0 y 20 años, se podría realizar un análisis de regresión entre cada par de variable, por ejemplo entre edad y estatura, entre edad y peso, entre edad y temperatura corporal, entre estatura y peso, etc, hasta completar todas las combinaciones posibles. Esto sería muy tedioso y largo, pero además se perdería la visión del conjunto, puesto que esas variables actúan en conjunto, no separadas, por lo que en este caso, lo recomendable es realizar un solo análisis de regresión múltiple, incluyendo a todas las variables. Se selecciona cuál es la variable dependiente y cuáles son las independientes, por ejemplo, la variable dependiente será el peso, entonces se hace el análisis del efecto que sobre el peso (y) tendrán la edad (x_1), la estatura (x_2) y la temperatura corporal (x_3). La ecuación a calcular será, en este caso: $y = a + b_1x_1 + b_2x_2 + b_3x_3$.

Hay otros casos donde la o las variables independientes no varía en forma directamente proporciona, sino en otras formas, por ejemplo cuadrática o logarítmica. En estos casos hay que resolver la ecuación de acuerdo con la forma de variación de los datos. En estos casos citados, las ecuaciones de la regresión son: $y = a + b^2$, $y = a + b \log x$, y las líneas serán unas líneas curvas.

Igualmente, de acuerdo con la distribución de los datos, puede ser la variable dependiente la que varíe en forma no directa sino, por ejemplo, logarítmica. En este caso la ecuación podría ser: $\log y = a + bx$. También podría ser: $\log y = a + b \log x$, etc. En estos casos y de acuerdo con esas distribuciones de los datos de ambas variables, las líneas resultantes serán parábolas, hipérbolas, combinaciones de ambas, de signo positivo o negativo. En casos más complejos, resultarán en una curva que pasa de una forma a otra, como en el caso de la Ley de los Rendimientos Decrecientes.

Advertencia: Al trabajar con cifras logarítmicas, debe sumarse la cifra 1 a cada dato para evitar que cuando haya un dato 0, el logaritmo será infinito y no se podrá hacer ningún cálculo sobre ese dato. Al sumar uno a los datos, cuando hay un 0, será: $\log 0 + 1$ y el $\log 1 = 0$.

CAPÍTULO 9.

ANÁLISIS DE CORRELACIÓN.

Análisis de Correlación.

Así como existe el análisis de regresión para determinar el grado de dependencia que existe entre dos variables (dependiente e independiente), algunas veces hay más de una variable independiente, también existe un análisis estadístico que permite determinar el grado de relación que puede existir entre dos variables, pero sin que necesariamente haya una relación de dependencia, es decir, hay una asociación entre las variables pero no necesariamente es una relación de causa-efecto.. En otras palabras, las dos variables pueden ser independientes una de la otra y aún así tener una estrecha relación estadística. Por ejemplo, el aumento de la población de una localidad o de un país puede estar estrechamente relacionado, en realidad correlacionado, con el aumento de la edad de una persona en particular, es decir, a medida que aumenta la edad de l a persona aumenta la población del a localidad o país. Estadísticamente se puede demostrar que ambas variables están estrechamente correlacionadas, pero por supuesto no tienen ningún grado de dependencia una de la otra. No porque a causa del aumento de la edad de la persona aumentará la población de la localidad o país, ni que por causa del aumento de la población ocurrirá un aumento de la edad del a persona. Se podría decir que el análisis de correlación mide el grado de cohesión que existe entre los valores de dos variables. En otras palabras mientras más cerca o cohesionados están los pares de valores de las dos variables más correlacionadas están las variables y mientras más lejos se encuentren los valores de las variables menos correlacionadas estarán las mismas.

Vale la pena destacar que a diferencia con el análisis de regresión, en el análisis de regresión no hay variable dependiente ni variable independiente, pues como hemos dicho antes, no existe relación de dependencia entre ambas. Por lo anterior se ha convenido, para fines del análisis, en que ambas variables son independientes y por lo tanto se les denomina x_1 y x_2 para evitar cualquier confusión en asignar dependencia a alguna de ellas.

Por otra parte, el análisis de correlación es una prueba complementaria, muy importante, del análisis de regresión y por lo general se realizan juntos ambos análisis. Cuando esto ocurre y sabiendo que en el análisis de regresión tenemos una variable independiente (x), y una variable dependiente (y), optamos por llamar a la independiente x_1 y a la dependiente x_2. Sin embargo, esto no tiene la menor importancia y puede invertirse la denominación de las dos variables, pues para los fines del análisis no importa cual sea cual, como veremos más adelante.

Para el análisis de correlación, debe calcularse el coeficiente de correlación r. Su cálculo se realiza como sigue.

Hay que destacar que el valor de r puede tomar cualquier valor desde - 1 hasta +1, pasando por 0. Un valor de 0 nos indica que no hay relación entre las variables (correlación), es decir, son totalmente independientes. Mientras r aumenta se hacen más correlacionadas las dos variables. La nube de dispersión de los datos se hace más estrecha, pasando de ser un circulo uniforme con $r = 0$, hasta hacerse como un pepino o un cigarro habano con r alrededor de + 0.5 o – 0.5, hasta llegar a lo máximo que sería tener todos los datos sobre una línea, con $r = +1$ o $r = -1$.

$$r = \frac{Covarianza\ de\ x_1.x_2 \quad (antes\ xy)}{Media\ geométrica\ de\ las\ Varianzas}$$

$$r = \frac{Covarianza\ de\ x_1.x_2}{\sqrt{Varianza\ de\ x_1\ .\ Varianza\ de\ x_2}}$$

Ejemplo:

En el mismo ejemplo usado arriba para la regresión, tendríamos:

$$r = \frac{38.942}{\sqrt{24.457\ x\ 63.400}} \qquad r = \frac{38.942}{\sqrt{1550.574}} \qquad r = \frac{38.942}{39.377} \qquad r = 0.989$$

Relación de b con r:

$$b = r\frac{s_y}{s_x}$$

s_y = Desviación standard de y s_x = Desviación standard de x

Prueba de significancia de r:

$$t = \frac{r\sqrt{n-2}}{1-r^2}$$ Grados de libertad = n – 2

$$t = \frac{0.989\sqrt{15-2}}{1-0.989^2}$$

$$t = \frac{0.989(3.605)}{1-0.978}$$

$$t = \frac{3.565}{0.022}$$

$$t = 162.045$$

En el ejemplo de arriba, el valor calculado de t es de 162.045 con $p \leq 0.000$, es decir, la prueba de t indica que es muy poca la probabilidad (menos de 1 en 10000) de que los resultados de r se deban al azar.

Otra alternativa para calcular el coeficiente de correlación es:

$$r = \frac{n\sum xy - \sum x \sum y}{(\sqrt{n\sum x^2 - (\sum x)^2})(\sqrt{n\sum y^2 - (\sum y)^2}}$$

Interpretación de estadística descriptiva.

Es importante interpretar adecuadamente los resultados del análisis de estadística descriptiva para evitar errores en dicha interpretación. La media solo tiene sentido si los datos están ubicados en una distribución normal, es decir, están distribuidos uniformemente alrededor de la media. La media o promedio en si misma tiene un valor limitado. Por ejemplo, si una persona tiene un pie en agua hirviendo y el otro pie en agua helada, estadísticamente, en promedio esa persona está muy confortable, la temperatura promedio será muy cómoda. Igualmente, si se toma el promedio de la estatura de los cadetes de la Escuela Militar se obtendrá, por ejemplo, 1.70 metros y si se toma el promedio de estatura de los miembros de un circo donde hay muchos enanos y algunos gigantes, se tendrá también 1.70 metros, por lo que se diría que la estatura de los cadetes y de los miembros del circo son iguales. Por lo tanto, para una buena interpretación hay que conocer y por ende acompañar a la media con medidas de dispersión de los datos como el rango y la desviación standard.

En ciencias de la salud, las estadísticas descriptivas no pueden usarse para definir enfermedad. El promedio no puede tomarse para indicar "normal". La desviación standard no debe usarse como definición del rango de lo "normal". Tomar un punto de corte en una distribución estadística para definir una enfermedad es erróneo. En muchos casos, por ejemplo, en los valores de pruebas de laboratorio, los valores se basan en el 95% de las personas aparentemente sanas, por lo que los que se salen del 95% se consideran fuera del rango normal, pero no es indicación d enfermedad.

En cuanto a la significancia, hay que tener en cuenta que todo lo que es significante o significativo, no es necesariamente importante. En investigación, significante e importante no son sinónimos.

Para permitir una interpretación apropiada hay que mostrar los valores exactos de la probabilidad P e indicar la prueba estadística utilizada. Cuando no se indica la prueba utilizada, se dice que son valores de P "huérfanos".

Es necesario tomar en cuenta que el tamaño de la P no es indicativo de la importancia de los resultados. Los resultados deben ser importantes por sí mismos y por su relación con los objetivos de la investigación.

Las diferencias en una investigación pueden no ser estadísticamente significativas y aun ser muy importantes.
El uso de programas de computación muy complejos y sofisticados, no garantizan la validez de la investigación. Si se insertan en el programa de la computadora, mala información o datos, se obtendrá como salida malos resultados. Se dice “basura que entra, basura que sale”.
Por otra parte, en las estadísticas de relación o asociación, hay que tener en cuenta la temporalidad, es decir, que la causa debe obviamente siempre preceder al efecto o consecuencia. Este principio, a veces no se toma en cuenta cuando se interpretan estudios o investigaciones de corte transversal o estudios de caso-control, en los cuales la causa y el efecto se miden al mismo tiempo. Por ejemplo, si se quiere saber el efecto de fumar sobre el cáncer de pulmones y se hace un estudio de corte transversal durante un día o semana o mes en una ciudad, no se puede establecer que el número de personas con cáncer de pulmón se debía al número de personas que fumaban.

CAPÍTULO 10.

ANÁLISIS DE LA VARIANZA. PRUEBA DE FISHER.

ANÁLISIS DE LA VARIANZA

El análisis de la varianza es quizá la prueba de mayor utilidad en investigación científica. Fue desarrollada por el eminente estadístico inglés Sir Ronald Fisher, de la Estación Experimental Agrícola de Rothamsted, en Saint Alban, Inglaterra. Recientemente esta Estación pasó del gobierno inglés al Colegio Imperial de Ciencia, Tecnología y Medicina de la Universidad de Londres (Este Colegio, a partir de fines del 2007, cuando cumplió 100. años, pasó a ser una universidad por sí mismo con el nombre de Imperial Collage London). Fisher llamó esta prueba, la Prueba de la Razón de las Varianzas (*Variance Ratio*), debido a que analiza la razón (matemática) que hay entre las varianzas de las fuentes de variación con la varianza del Error o Residuo (en inglés *Error* o *Residual*).

Se llama *Tratamiento* a cada uno de los "asuntos" que se han puesto en prueba, por ejemplo, diferentes medicamentos para una enfermedad, o diferentes fertilizantes para un cultivo agrícola, o diferentes insecticidas contra alguna plaga, etc. Se llaman *Repeticiones* a las diferentes veces que se repite un tratamiento, por ejemplo el medicamento que se repite en varios hospitales diferentes, el nutriente que se repite en varias fincas diferentes o el insecticida que se repite en tres barrios diferentes.

Esta prueba compara las medias de diferentes tratamientos mediante el análisis de sus varianzas. Este análisis parte de la hipótesis nula de que las medias de todos los tratamientos son iguales, es decir, que no hay diferencias estadísticamente significativas entre ellas. Para facilitar su desarrollo se crea un CUADRO DE ANÁLISIS DE LA VARIANZA como sigue, en cuyo ejemplo se trata de un experimento donde hubo cuatro tratamientos y tres repeticiones.

Tabla 12. CUADRO DE ANÁLISIS DE LA VARIANZA

Fuente de Variación FV	Grados de Libertad GL	Suma de Cuadrados SC	Varianza V	Razón de las Varianzas F
Tratamientos T = 4	$GL_T = T - 1 = 3$	$SC_T = [(\Sigma\ T_1^2 + T_2^2 + T_3^2 + T_4^2)/R\ (3)] - FC$	$V_T = SC_T/GL_T$	$F_T = V_T / V_E$
Repeticiones R = 3	$GL_R = R - 1 = 2$	$SC_R = [(\Sigma\ R_1^2 + R_2^2 + R_3^2)/\ T\ (4)] - FC$	$V_R = SC_R/GL_R$	$F_R = V_R / V_E$
Error o Residuo E = (To – [T + R]) = (12 – [4 + 3]) = 5	$GL_E = GL_{To} - ([GL_T + GL_R]) = 11 - ([3 + 2])$	$SC_E = SC_{To} - (SC_T + SC_R)$	$V_E = SC_E/GL_E$	
Total To = 12	$GL_{To} = To - 1 = 11$	$SC_{To} = x_1^2 + x_2^2 + \ldots\ \ldots + x_{12}^2 - FC$		

Donde: FV es Fuente de Variación (en inglés *SV* o *Source of Variation*)

GL es Grados de Libertad (en inglés *DF* o *Degrees of Freedom*)

SC es Suma de Cuadrados (en inglés *SS* o *Sum of Squares*) También se le conoce como Cuadrado de la Media, Media al Cuadrado o Media Cuadrática (en inglés *MS* o *Mean Square*). Debe recordarse que no se trata de una simple suma de los valores elevados al cuadrado, sino que es esa suma menos un Factor de Corrección (FC).

FC es el Factor de Corrección. El Factor de Corrección es la suma (sencilla) de cada uno de los valores individuales de los elementos (observaciones, sujetos, individuos, etc.) elevada al cuadrado y luego dividida entre el número total de los elementos: $FC = (\Sigma x_1 + x_2 + \ldots\ \ldots + x_n)^2/n$. En nuestro caso sería: $FC = (\Sigma x_1 + x_2 + \ldots\ \ldots + x_{12})^2/n$, donde n = 12.

Varianza (en inglés *Variance*), su nombre indica de lo que se trata, es la variación estadística que existe en los datos en los cuales se calcula, es decir, en la fuente de variación en la que se calcula. Su cálculo se realiza dividiendo la Suma de Cuadrados (para la Fuente de Variación) entre los Grados de Libertad (para esa Fuente de Variación).

F es la prueba de significancia de las diferencias entre las varianzas. Sir Ronald Fisher, de Inglaterra, quien la creó y desarrolló, la llamó Prueba de la Razón de las Varianzas. Posteriormente, el Profesor George W. Snedecor, también eminente estadístico, profesor en la Universidad de Raleigh en Carolina del Norte,

Estados Unidos, refinó la prueba y la denominó Prueba de Fisher en honor a su creador y le designó la letra F para abreviar.

La prueba comienza por establecer las Fuentes de Variación, pongamos por caso un experimento donde se probaron cuatro condiciones, por ejemplo, rendimiento escolar en cuatro institutos o percepción de trato en cuatro tiendas comerciales, o cuatro métodos de construcción de casas, o productos (cuatro detergentes, o medicamentos, o fertilizantes o alimentos), etc. A los cuales se les denominará Tratamientos. Vale la pena recordar que el término Tratamiento, desde el punto de vista estadístico, no solo incluye los elementos o sujetos sometidos a la intervención sino también los elementos de comparación que no llevan o no son sometidos a la intervención, conocidos también como Controles o Testigos. En nuestro ejemplo, se someterán tres sujetos o individuos a cada tratamiento. En este caso cada grupo sometido a un mismo tratamiento se le denomina Repetición (en inglés *Replication*). No debería usarse el término Replica o Replicación en castellano, pues esto implica una reproducción en menor escala del original, y en nuestro caso cada elemento en cada Repetición es una copia exacta del original. Así tendremos cuatro Tratamientos con tres Repeticiones, es decir, tendremos un Total de 4 x 3 = 12 elementos (sujetos, individuos, etc.). Las Fuentes de Variación serán: Tratamiento, Repeticiones, Error o Residuo y Total. Los Grados de Libertad serán en cada caso n-1, es decir, para Tratamientos será el número de tratamientos menos uno (T-1 ó 4-1 = 3). En el caso de las Repeticiones será el número de Repeticiones menos uno (R-1 ó 3-1 = 2). Para el Total será el número total de elementos (individuos, sujetos, observaciones, etc) menos uno (To-1 ó 12-1 = 11). Para el Error o Residuo será la diferencia de los grados de libertad del Total menos la suma de los grados de libertad de los Tratamientos más los de las Repeticiones $GL_E = GL_{To} – (GL_T + GL_R)$. En nuestro caso será 11-(3+2) = 11-5 = 6. Cuando el diseño experimental tiene esta misma estructura, es decir, solo tratamientos y repeticiones (no hay divisiones o subdivisiones de los tratamientos tal como ocurre en diseño experimental conocido como Parcelas Divididas (en inglés *Split Plots*), se pueden calcular los grados de libertad para el Error multiplicando los grados de libertad del Tratamiento por los de las Repeticiones, por ejemplo, en nuestro caso será 3 x 2 = 6.

Luego se calculan las Sumas de Cuadrado de cada fuente de variación. Para este caso es conveniente indicar que en el análisis de la varianza, el Factor de Corrección es uno solo para todas las fuentes de variación. Para calcular la Suma de Cuadrados en el caso de los Tratamientos, se eleva al cuadrado el total de cada tratamiento, se suman y se divide entre tres (ya que cada tratamiento está compuesto por tres componentes, uno por cada repetición) y se le resta el Factor de Corrección: En nuestro ejemplo: $SC_T = [(\Sigma\ T_1^2 + T_2^2 + T_3^2 + T_4^2)/R\ (3)] – FC$. Para las Repeticiones se eleva al cuadrado el total de cada repetición, se suman y se divide entre cuatro (ya que cada repetición está compuesta por cuatro componentes, uno por cada tratamiento) y se le resta el Factor de Corrección, que debe ser el mismo valor. En nuestro ejemplo: $SC_R = [(\Sigma\ R_1^2 + R_2^2 + R_3^2)/T\ (4)] – FC$. Para el Total, se eleva al cuadrado cada elemento (observación, sujeto, etc.), se suman y se le resta el Factor de Corrección. En nuestro ejemplo: $SC_{To} = x_1^2 + x_2^2 + \ldots\ \ldots + x_{12}^2 - FC$. La Suma de Cuadrados para el Error o Residuo se calcula por diferencia, es decir, a la Suma de Cuadrados para el Total se le resta las Sumas de Cuadrados combinadas de Tratamientos más la de Repeticiones: $SC_E = SC_{To} – (SC_T + SC_R)$.

Luego se calculan las Varianzas de cada fuente de variación, excepto la Varianza del Total porque por ser valores individuales, dentro de cada individuo o unidad no hay (no puede haber) variación. En este caso para calcular la Varianza, se divide la Suma de Cuadrados de cada fuente de variación entre sus respectivos Grados de Libertad. En el caso de la Varianza de los Tratamientos será la Suma de Cuadrados para Tratamientos entre los Grados de Libertad para Tratamientos: $V_T = SC_T/GL_T$. Para las Repeticiones será la Suma de Cuadrados para Repeticiones entre los Grados de Libertad para Repeticiones: $V_R = SC_R/GL_R$. Para el Error o Residuo será la Suma de Cuadrados para el Error entre los Grados de Libertad para el Error: $V_E = SC_E/GL_E$.

Luego se calcula la Razón de las Varianzas, mejor conocida como la F, nombre dado en honor a Sir Ronald Fisher quien la creó. En el caso de la F para los Tratamientos se calcula dividiendo la Varianza de los Tratamientos entre la Varianza del Error o Residuo: $F_T = V_T / V_E$. En el caso de la F para las Repeticiones, se calcula dividiendo la Varianza para las Repeticiones entre la Varianza para el Error o Residuo: $F_R = V_R / V_E$. Para determinar la existencia de diferencias estadísticamente significativas entre los tratamientos o entre las repeticiones, se comparan los valores ahora calculados, que llamaremos F_{Tc} y F_{Rc} con aquellos establecidos en las Tablas de F o de Razón de las Varianzas que se encuentran en la mayoría de los textos de estadística y que mostramos aquí como **Anexo 3**. Se toman los grados de libertad para tratamientos o repeticiones como valores del numerador y los grados de libertad para el error o residuo como valor del denominador. Si el valor calculado es igual o mayor al valor encontrado en las tablas se dice que hay diferencias estadísticamente significativas entre los tratamientos o repeticiones, según sea el caso. Las tablas se

presentan, generalmente, para probabilidades de P = 0.05; 0.01; 0.005 y 0.001. Aunque hay algunas tablas con valores de probabilidad más altos o más bajos.

Pruebas de la mínima diferencia significativa. El análisis de la varianza nos indica si hay o no diferencias estadísticamente significativas entre los diferentes tratamientos, que en algunos casos es lo que busca el investigador, por ejemplo, si se trata de varios productos comerciales, el objetivo puede ser demostrar que son estadísticamente diferentes y que se debe preferir el de mejor efecto, rendimiento, etc. De igual manera, el investigador puede estar interesado en que no haya diferencias estadísticamente significativas entre los tratamientos, en el mismo ejemplo, el objetivo puede ser demostrar que los productos son estadísticamente iguales y que se debe preferir el de menor precio. En todos los casos, siempre se desea que no haya diferencias estadísticamente significativas entre las repeticiones, ya que si las hay, se puede confundir el efecto entre tratamientos, es decir que si hay diferencias entre las repeticiones, las diferencias encontradas entre los tratamientos pueden deberse a las diferencias dentro de una o más de las repeticiones de uno o más tratamientos y no a los tratamientos en sí. En este caso, los resultados de la investigación quedan en duda, no se puede llegar a una conclusión cierta.
Si el análisis de varianza nos indica que hay diferencias entre los tratamientos y no hay entre las repeticiones, nos preguntaremos ¿entre cuáles tratamientos hay las diferencias?, en el ejemplo de los cuatro productos, llámense A, B, C y D, con diferencias estadísticamente significativas, no sabemos si la diferencia es entre A y el resto o A yB son iguales entre sí, pero diferentes del resto, o si A, B y C son iguales entre sí pero diferentes de D, etc. Para dilucidar esta duda se utiliza alguna de las pruebas de la mínima diferencia significativa (*mds*), en inglés *Least Significant Difference(LSD)*. Estas pruebas consisten en determinar cuál es la mínima diferencia que debe existir entre las medias de dos tratamientos para que haya diferencias estadísticamente significativas entre ellas.

Prueba de Tukey.

Mínima diferencia significativa (*mds*) = $Q\, s_{\bar{x}}$

Q: se encuentra su valor en la tabla de *Q* en el **Anexo 3** (tomada de Snedecor 1956) o se puede calcular su valor de acuerdo con: $Q = \dfrac{\bar{x}máx - \bar{x}mín}{s_{\bar{x}}}$

El error standard de la diferencia entre medias es:

$$s_{\bar{x}} = \sqrt{\frac{s_x^{\,2}}{n}}$$

Calculado a partir de la varianza del error:

$s_{\bar{x}}^{\,2}$

n = Número de repeticiones o número de diferencias entre medias o tratamientos.

Tabla 13. Cuadro de Análisis de la Varianza. Ejemplo.

	Tratamiento A	Tratamiento B	Tratamiento C	Tratamiento D	Total Repeticiones	Media Repeticiones	(Repetición)^2
Repetición 1	5	12	30	450	497	124.25	247009
Repetición 2	3	15	45	600	663	165.75	439569
Repetición 3	8	18	60	800	886	221.50	784996
Total Tratamientos	16	45	135	1850	2046		4186116
Media Tratamientos	5.33	15.00	45.00	616.67			
(Tratamiento)^2	256	2025	18225	3422500			

$$n = \frac{a(a-1)}{2}$$

a = Número de tratamientos o medias.

Ejemplo:

Tratamientos: A, B, C, D.
Repeticiones: 1, 2, 3.

Tabla 14. Valores del ejemplo, elevados al cuadrado.

	Valores al cuadrado					
	A	B	C	D	Total	
1	25	144	900	202500	203569	
2	9	225	2025	360000	362259	
3	64	324	6400	640000	646788	
Total	98	693	9325	1202500	1212616	

$FC = (5 +3 + 8 + 12 + 15 + 18 + 30 + 45 + 60 +450 + 600 + 800)^2 / 12$
$FC = 2046^2 / 12$
$FC = 4186116 / 12$
$FC = 348843$

$SC\ Total = (5^2 + 3^2 + 8^2 + 12^2 + 15^2 + 18^2 + 30^2 + 45^2 + 60^2 + 450^2 + 600^2 + 800^2) - 348843$
$SC\ Total = (25 + 9 + 64 + 144 + 225 + 324 + 900 + 2025 + 6400 + 202500 + 360000 + 640000) – 348843$
$SC\ Total = 1212618 – 348843$
$SC\ Total = 863775$

$SC\ Tratamientos = [(16^2 + 45^2 + 135^2 + 1850^2) / 3] – 348843$
$SC\ Tratamientos = [(256 + 2025 + 18225 + 3422500) / 3] – 348843$
$SC\ Tratamientos = (3443006 / 3) – 348843$
$SC\ Tratamientos = 1147668.7 – 348843$
$SC\ Tratamientos = 798825.7$

$SC\ Repeticiones = [(497^2 + 663^2 + 886^2) / 4] - 348843$
$SC\ Repeticiones = [(247009 + 439569 + 784996) /4] – 348843$
$SC\ Repeticiones = (1471574 / 4) – 348843$
$SC\ Repeticiones = 367893.50 – 348843$
$SC\ Repeticiones = 19050.5$

$SC\ Error = SC\ Total – (SC\ Tratamientos + SC\ Repeticiones)$
$SC\ Error = 863775 – (798825.7 + 19050.5)$
$SC\ Error = 863775 – 817876.2$
$SC\ Error = 45898.8$

$Varianza\ Tratamientos = 798825.7 / 3$
$Varianza\ Tratamientos = 266275.23$

$Varianza\ Repeticiones = 19050.5 / 2$
$Varianza\ Repeticiones = 9525.25$

$Varianza\ del\ Error = 45898.8 / 6$
$Varianza\ Error = 7649.8$

$F\ Tratamientos = 266275.23 / 7649.8$
$F\ Tratamientos = 34.8081$
F en la Tabla de F (ver **Anexo 3)**: P = 0 0.05 F = 4.76 P = 0.01 F = 9.78 P = 0.001 F = 23.70
Se concluye que hay diferencias estadísticas muy altamente significativas entre los tratamientos.

$F\ Repeticiones = 9525.25 /7649.8$
$F\ Repeticiones = 1.2952$
F en la Tabla de F (ver **Anexo 3)**: P = 0.05 F = 5.14 P = 0.01 F = 10.92 P = 0.001 F = 27.00
Se concluye que no hay diferencias estadísticamente significativas entre las repeticiones.

La mínima diferencia significativa (mds) es:
Q en la Tabla de *Q* (ver **Anexo 4**): 4.90

$$n = \frac{a(a-1)}{2}$$

n = [4(4-1)] / 2
n = 12 / 2
n = 6

Varianza del error = $s_{\bar{x}}^2$

Varianza del error = 7649.8

$$s_{\bar{x}} = \sqrt{\frac{7649.8}{6}}$$

$$s_{\bar{x}} = \sqrt{1274.97}$$

$$s_{\bar{x}} = 35.71$$

mds = 4.90 x 35.71
mds = 174.98

Medias de los tratamientos:

	616.67	45.00	15.00	5.33
616.67		571.67	601.67	611.34
45.00			30.00	39.67
15.00				9.67

Hay diferencias estadísticamente significativas entre el tratamiento A (media = 616.67) y el resto de los tratamientos, pues su diferencia es superior a la mds (174.98), pero no hay entre ninguno de los otros tratamientos, es decir, sus diferencias no superan la mds (174.98).

Diseños experimentales.
En algunos casos se realizan los experimentos, especialmente, los experimentos de campo, mediante una distribución que se ha denominado "diseño experimental", en el cual los tratamientos se dividen en la forma más conveniente que permita disminuir los "vicios" o "sesgos" voluntarios o involuntarios. Los diseños más comunes son:

Bloques al azar. En este diseño los "bloques" son divisiones donde se ubican los diferentes tratamientos, de manera que cada bloque será una repetición, es decir, cada bloque tendrá tantas "parcelas" o divisiones como tratamientos haya. Un ejemplo donde hay cuatro tratamientos (A, B, C, D) y tres bloques o repeticiones (I, II, III, IV), se le abrevia de 4 x 3, tendrá tres bloques y dentro de cada bloque estará cada uno de los tratamientos colocado mediante sorteo al azar. La desventaja de este diseño es que por efecto mismo del azar pueden quedar los tratamientos de uno o más repeticiones uno al lado del otro.

Tabla 15. Ejemplo de Bloques al Azar.

Bloque I	Bloque II	Bloque III
Tratamiento A	Tratamiento C	Tratamiento D
Tratamiento B	Tratamiento A	Tratamiento C
Tratamiento C	Tratamiento D	Tratamiento A
Tratamiento D	Tratamiento B	Tratamiento D

En este caso, el cuadro de Análisis de la Varianza tendrá como fuentes de variación, los tratamientos, las repeticiones (columnas), el error y el total, tal como se expuso anteriormente.

Cuadrado Latino. En este diseño se tienen tantas repeticiones o bloques como tratamientos haya, es decir, en un experimento con cuatro tratamientos deberá haber cuatro bloques y en uno de cinco, deberá haber

cinco bloques y así sucesivamente. En el Cuadrado Latino no se debe repetir un tratamiento en las repeticiones o bloques (como en el diseño de Bloques al Azar) que forman las columnas, pero tampoco debe repetirse en las hileras. A continuación un ejemplo de un Cuadrado Latino de cinco tratamientos y cinco repeticiones que se abrevia de 5 x 5.
En este caso, el cuadro de Análisis de la Varianza tendrá como fuentes de variación, los tratamientos, las repeticiones (columnas), las hileras, el error y el total.

Tabla 16. Ejemplo de Cuadrado Latino.

	Bloque I	Bloque II	Bloque III	Bloque IV	Bloque V
Hilera 1	Tratamiento A	Tratamiento D	Tratamiento E	Tratamiento B	Tratamiento C
Hilera 2	Tratamiento B	Tratamiento A	Tratamiento C	Tratamiento E	Tratamiento D
Hilera 3	Tratamiento C	Tratamiento E	Tratamiento A	Tratamiento D	Tratamiento B
Hilera 4	Tratamiento D	Tratamiento C	Tratamiento B	Tratamiento A	Tratamiento E
Hilera 5	Tratamiento E	Tratamiento B	Tratamiento D	Tratamiento C	Tratamiento A

Cuadrado Greco-Latino. En este diseño se trata de que diferentes tratamientos con alternativas en cada uno, no se repita dentro del cuadrado, así que habrán combinaciones de letras griegas y latinas (de allí su nombre).
En este caso, el cuadro de Análisis de la Varianza tendrá como fuentes de variación, los tratamientos, las repeticiones (columnas), las hileras, la interrelación de griego con latín, el error y el total.

Tabla 17. Ejemplo de Cuadrado Greco-Latino.

	Bloque I	Bloque II	Bloque III	Bloque IV	Bloque V
Hilera 1	Tratamiento Aα	Tratamiento Eγ	Tratamiento Dε	Tratamiento Bγ	Tratamiento Cε
Hilera 2	Tratamiento Bβ	Tratamiento Dα	Tratamiento Cδ	Tratamiento Eβ	Tratamiento Aβ
Hilera 3	Tratamiento Cγ	Tratamiento Aε	Tratamiento Eα	Tratamiento Dγ	Tratamiento Bε
Hilera 4	Tratamiento Dδ	Tratamiento Bδ	Tratamiento Aγ	Tratamiento Cα	Tratamiento Eδ
Hilera 5	Tratamiento Eε	Tratamiento Cβ	Tratamiento Bα	Tratamiento Aε	Tratamiento Dβ

Parcelas divididas o Split Plots. Este diseño, también llamado **Factorial o multifactorial**, se desarrolla para tener más divisiones dentro de cada tratamiento. Por ejemplo, si queremos conocer el efecto de varios medicamentos sobre la eliminación de una enfermedad o de varios fertilizantes sobre el rendimiento de un cultivo, podremos utilizar este diseño, donde se irán dividiendo las posibilidades de tratamiento. Por ejemplo, si se trata de medicamentos, se podrá tener como Fuentes de Variación: los tratamientos, luego se divide por su aplicación de acuerdo con el sexo: masculino, femenino, luego en cada sexo, de acuerdo con la edad: 0 a 5 años, 6 a 10 años, 11 a 15 años, etc, luego se divide para cada edad de acuerdo con la vía de administración luego para cada vía de administración de acuerdo con la procedencia: Población A, Población B, etc, luego se divide para cada procedencia de acuerdo con el grado de instrucción, luego de acuerdo con los ingresos económicos, luego con la ocupación y así sucesivamente hasta completar cuantas divisiones establezca el investigador. Igual criterio para cualquier tipo de experimento.
Un ejemplo abreviado (no completo, por falta de espacio) de cuadro de Análisis de la Varianza de un experimento con varios cultivos (maíz, caraotas, papas, repollo, etc) y varios fertilizantes (Fertilizante Urea, Nitrophoska, NPS, etc), aplicado en diferentes épocas, con y sin riego, con y sin plaguicidas es el siguiente (en las celdas vacías van los rendimientos en kg/ha de los cultivos).

Tabla 18. Ejemplo de Parcelas divididas o Split Plot.

Cultivo	**Fertilizante**	**Aplicación**	**Riego**	**Plaguicidas**	**Bloque 1**	**Bloque 2**	**Blo-que 3**	**Blo-que 4**
Maíz	Urea	Al sembrar	Con riego	Con plaguicidas				
				Sin plaguicidas				
			Sin riego	Con plaguicidas				
				Sin plaguicidas				
		Al primer aporque	Con riego	Con plaguicidas				
				Sin plaguicidas				
			Sin riego	Con plaguicidas				
				Sin plaguicidas				
		Al segundo aporque	Con riego	Con plaguicidas				
				Sin plaguicidas				
			Sin riego	Con plaguicidas				
				Sin plaguicidas				
	Nitrophoska	Al sembrar	Con riego	Con plaguicidas				
				Sin plaguicidas				
			Sin riego	Con plaguicidas				
				Sin plaguicidas				
		Al primer aporque	Con riego	Con plaguicidas				
				Sin plaguicidas				
			Sin riego	Con plaguicidas				
				Sin plaguicidas				
		Al segundo aporque	Con riego	Con plaguicidas				
				Sin plaguicidas				
			Sin riego	Con plaguicidas				
				Sin plaguicidas				
Caraotas	Urea	Al sembrar	Con riego	Con plaguicidas				
				Sin plaguicidas				
			Sin riego	Con plaguicidas				
				Sin plaguicidas				
La tabla continúa con los diferentes valores hasta completar los diferentes cultivos, fertilizantes, épocas de aplicación, etc.								

CAPÍTULO11.

EPIDEMIOLOGÍA. PREVALENCIA. INCIDENCIA. TASA DE INCIDENCIA. RIESGO. FACTOR DE RIESGO. RIESGO RELATIVO.

La **epidemiología** como disciplina e instrumento fundamental en el análisis de los resultados de cualquier elemento o hecho relacionado con la salud, tiene como principal herramienta de trabajo y por tanto de análisis de la información numérica obtenida en cualquier situación, caso, región, etc, sea información espacial o temporal, los valores de diagnóstico para una patología o situación en particular o aún en general. Se pretende, con los valores conocidos acerca de una patología, predecir su comportamiento futuro o tendencia. Para ello es necesario conocer el número de personas en una población dada en quienes se manifiesta la patología, es decir, la frecuencia en que aparece la patología en el total de la población en cuestión y coincide con la probabilidad de que una persona de la población sufra la patología. Esta cifra o frecuencia se le denomina **prevalencia** y se expresa en términos relativos, por ejemplo, en porcentaje. Por ejemplo, la patología se encuentra en 2500 personas de una población de 100000 habitantes, por lo tanto la prevalencia de la patología en esta población es de 2.5%. Por otra parte, se conoce como **incidencia** al número total de nuevos casos de personas que sufren una patología durante un período de tiempo determinado y en una población determinada. Por ejemplo, el número de nuevos casos de la patología que se presentó en el estado Mérida durante el año 2008 es de 450. En este caso la incidencia de la patología en Mérida en el año 2008 es de 450. La **tasa de incidencia,** también denominado **riesgo** o **riesgo absoluto,** es el número total de nuevos casos de personas que sufren una patología durante un período de tiempo determinado y en una población determinada divido entre el número total de la población en riesgo, es decir, la incidencia entre la población en riesgo. En el ejemplo anterior, la tasa de incidencia o riesgo de contraer la patología en el año 2008, suponiendo que la ciudad de Mérida tuviese 300000 habitantes, es de 450/300000 por año, 1.5/1000 por año o 0.15% por año. Toda población está formada por diferentes grupos, en los cuales el riesgo de contraer una patología no es igual para todos. Si en el ejemplo anterior, la patología fuese dermatosis en las manos causada por hongos, el riesgo en mujeres que trabajan en la fabricación de comida, por ejemplo, cocineras, podría ser 600/300000 al año, mientras que en todas las otras mujeres (no cocineras) podría ser de 200/300000, por lo tanto se puede decir que el riesgo de contraer la dermatosis es el triple en las cocineras que en el resto de las mujeres.

En medicina es común encontrar situaciones donde se presentan dos variables dicotómicas o binarias (ver Capítulo 2) en relación con el diagnóstico de alguna enfermedad o patología. En primer término tenemos la variable enfermedad o patología que llamaremos p, con sus dos posibilidades, patología presente (p +) o patología ausente (p -) y luego tenemos la variable prueba de diagnóstico o examen que llamaremos e, con sus dos posibles resultados, examen positivo (e +) o examen negativo (e -). Cualquier grupo de personas o población (n o N) tendrá las siguientes cuatro alternativas cuando se somete a la prueba para diagnosticar la patología:

1) Tiene la enfermedad y la prueba resultó positiva: Verdaderos positivos.
2) Tiene la enfermedad y la prueba resultó negativa: Falsos negativos.
3) No tiene la enfermedad y la prueba resultó positiva: Falsos positivos.
4) No tiene la enfermedad y la prueba resultó negativa: Verdaderos negativos.

Con estos datos se puede construir una tabla de contingencia de 2 x 2, como sigue:

Tabla 19. Tabla de contingencia de 2 x 2, para cálculos epidemiológicos.

		Patología		
		Presente +	Ausente -	Σ
Examen	Positivo +	Verdaderos positivos + + a 90	Falsos positivos - + b 495	Exámenes Positivos a + b 585
	Negativo -	Falsos negativos + - c 10	Verdaderos negativos - - d 9405	Exámenes Negativos c + d 9415
	Σ	Patología Presente (Enfermos) a + c 100	Patología Ausente (Sanos) b + d 9900	Total a + b +c +d = n 10000

Si se conoce la prevalencia de la patología en una muestra o población dada y se conocen los resultados de las pruebas de diagnóstico o exámenes en los individuos de esa población, se podrán calcular, para esa prueba diagnóstica, la sensibilidad, la especificidad, el valor predictivo positivo y el valor predictivo negativo.

La **prevalencia** de una enfermedad es la frecuencia con la cual aparece la enfermedad en el total de la población (n o N), y coincide con la probabilidad de que una persona de la población sufra la enfermedad. En el caso de la tabla de arriba, la prevalencia o probabilidad de enfermar, es decir, de que una persona de la población (n) padezca la enfermedad es: $\text{Pr}obabilidad\ de\ enfermar = \frac{a+c}{n}$

Sensibilidad. La **sensibilidad** del examen o prueba diagnóstica es la probabilidad de que el resultado del examen o prueba sea positivo (Examen positivo +) en una persona que padece la enfermedad; en otras palabras, es la proporción de personas con la enfermedad que tienen un resultado positivo en la prueba diagnóstica. Es la fracción de los verdaderos positivos. Se le considera como la probabilidad condicionada

$$Sensibilidad = \frac{a}{a+c}$$

La sensibilidad, generalmente, se expresa como porcentaje, ya que representa el porcentaje de resultados positivos (a) en relación con el total de personas enfermas (a+c), en otras palabras, es el porcentaje de verdaderos positivos que se obtendría al aplicar la prueba diagnóstica a quienes están enfermos.

Por otra parte, la probabilidad complementaria de la sensibilidad es:

$$1-\frac{a}{a+c}=\frac{c}{a+c}$$

Esto es el porcentaje de resultados negativos del examen en relación con el total de enfermos (falsos negativos).

De esta manera mientras mayor es el número de verdaderos positivos (a) mayor es la sensibilidad y, viceversa, mientras mayor es la sensibilidad, será menor el número de falsos negativos (c). Por lo tanto, cuanto más sensible es una prueba diagnóstica menor es la probabilidad de obtener falsos negativos, de tal forma que un resultado negativo es muy confiable y permite descartar la presencia de enfermedad, es decir, conocer que la persona no está sufriendo la enfermedad.

Especificidad. La **especificidad** del examen o prueba diagnóstica es la probabilidad de que el resultado del examen o prueba sea negativo en una persona sana, es decir, de quien no padece la enfermedad; en otras palabras es la proporción de personas sin la enfermedad que tienen un resultado negativo en la prueba

diagnóstica. La especificidad representa la fracción de verdaderos negativos. Se le considera la probabilidad condicionada

$$Especificidad = \frac{d}{b+d}$$

La especificidad, generalmente, se expresa como porcentaje, ya que representa el porcentaje de personas con resultados negativos (d) en relación con el total de personas sanas (b+d), en otras palabras, es el porcentaje de verdaderos negativos que se obtendría al aplicar la prueba diagnóstica a personas sanas.
Por otra parte, la probabilidad complementaria de la especificidad es:

$$1-\frac{d}{b+d}=\frac{b}{b+d}$$

Esto es el porcentaje de resultados positivos del examen en relación con el total de personas sanas (falsos positivos). De esta manera mientras más específica es una prueba diagnóstica, menor es la probabilidad de obtener un falso positivo. Por lo tanto un resultado positivo en la prueba es muy confiable y permite con gran certeza saber que el paciente padece la enfermedad, es decir, se confirma el diagnóstico. Esto se exceptúa, es decir, no es válido en las enfermedades de baja prevalencia.

Valor Predictivo de Prueba Positiva o Valor Predictivo Positivo. El **valor predictivo de prueba positiva o valor predictivo positivo (VPP)** de una prueba diagnóstica es la probabilidad de que una persona sufra la enfermedad habiendo resultado positivo en la prueba. Se le considera como la probabilidad condicionada:

$$Valor\ \ predicitivo\ de\ \ prueba\ \ positiva = \frac{a}{a+b}$$

El valor predictivo positivo, generalmente, se expresa como porcentaje, ya que representa el porcentaje de personas realmente enfermas (a= verdaderos positivos) respecto del total de personas que han dado positivo en la prueba (a+b). Su complementario será:

$$1-\frac{a}{a+b}=\frac{b}{a+b}$$

Esto es el porcentaje de falsos positivos respecto del total de positivos. Un valor predictivo positivo alto indica que la probabilidad es muy alta de que la persona esté realmente enferma cuando ha resultado positivo en la prueba diagnóstica. También se puede calcular de la siguiente manera:

$$Valor\ \ predicitivo\ \ positivo = \frac{Sensibilidad\ \ de\ la\ \ prueba\ x\ \ prevalencia\ de\ la\ \ patología}{(Sensibilidad\ \ x\ \ prevalencia)+[(1-Especificidad)(1-\Pr evalencia)]}$$

De esta manera, mientras mayor sea la prevalencia de la patología en la población, mayor será el valor predictivo positivo de la prueba diagnóstica, independientemente de que la sensibilidad y la especificidad se mantengan constantes.

Valor Predictivo de Prueba Negativa o Valor Predictivo Negativo (VPN). El valor predictivo de prueba negativa o valor predictivo negativo de una prueba diagnóstica es la probabilidad de que una persona no sufra la enfermedad habiendo resultado negativo en la prueba. Se le considera como la probabilidad condicionada:

$$Valor\ \ predicitivo\ de\ \ prueba\ negativa = \frac{d}{c+d}$$

El valor predictivo negativo, generalmente, se expresa como porcentaje, ya que representa el porcentaje de personas realmente sanas o no enfermas (d= verdaderos negativos) respecto del total de personas que han dado negativo en la prueba (c+d). Su complementario será:

$$1-\frac{d}{c+d}=\frac{c}{c+d}$$

Esto es el porcentaje de falsos positivos respecto del total de positivos. Un valor predictivo negativo alto indica que la probabilidad es muy alta de que la persona esté realmente sana cuando ha resultado negativo en la prueba diagnóstica. También se puede calcular de la siguiente manera:

Ejemplo 1: Supongamos una población de 10000 personas y una enfermedad con prevalencia de 1%. Esto quiere decir que 100 personas padecen la enfermedad mientras que 9900 no tiene la enfermedad. En esta población el 90% de las pacientes presentan resultado del examen o prueba diagnóstica positivo y 95% de los controles presentan resultado de la prueba negativo.

+ + Verdaderos positivos
- - Verdaderos negativos
+ - Falsos negativos
- + Falsos positivos

Tabla 20. Tabla de contingencia de 2 x 2, para determinación de Sensibilidad y Especificidad. Ejemplo 1.

		Patología		
		Presente +	Ausente -	Σ
Prueba o Examen	Positivo +	+ + 90	- + 495	Examen Positivo 585
	Negativo -	+ - 10	- - 9405	Examen Negativo 9415
	Σ	Patología Presente 100	Patología Ausente 9900	Total 10000

$$Sensibilidad = \frac{\text{Prueba positiva} + \text{Patología presente}}{\text{Total prueba positiva}}$$

$$Sensibilidad(p) = \frac{90}{585}$$
$$= 0.15$$
$$\approx 15\%$$

La prueba tiene Baja Sensibilidad

$$Especificidad\ (n) = \frac{\text{Prueba negativa} + \text{Patología ausente}}{\text{Total prueba negativa}}$$

$$Especificidad\ (n) = \frac{9405}{9445}$$

$$= 0.9989$$

$$\approx 100\%$$

La prueba tiene Alta Especificidad

Ejemplo 2: Supongamos que la prevalencia de patología es de 60%. El resto de los datos sigue igual.

Tabla 21. Tabla de contingencia de 2 x 2, para determinación de Sensibilidad y Especificidad. Ejemplo 2.

		Patología		
		Presente +	Ausente -	Σ
Examen	Positivo +	+ + 5400	- + 200	Examen Positivo 5600
	Negativo -	+ - 600	- - 3800	Examen Negativo 4400
	Σ	Patología Presente 6000	Patología Ausente 4000	Total 10000

$$Sensibilidad = \frac{5400}{5600}$$
$$= 0.96$$
$$= 96\%$$

La prueba tiene Alta Sensibilidad

$$Especificidad = \frac{3800}{4400}$$
$$= 0.86$$
$$= 86\%$$

La prueba tiene Baja Especificidad

Factor de riesgo. Se denomina factor de riesgo a cualquier condición o factor cuya presencia está asociada al incremento del riesgo de contraer la patología. En el ejemplo anterior, el oficio de cocinera lleva el factor de riesgo de contraer la dermatosis en manos. Sin embargo, debe destacarse que la presencia de un factor de riesgo no indica que hay una relación de causa a efecto entre el factor y la patología, es decir, que puede haber presencia del factor de riesgo sin que haya la patología y viceversa, puede haber la patología sin que hubiese el factor de riesgo.

Riesgo relativo. Se denomina riesgo relativo al riesgo de enfermarse las personas de un grupo con el factor de riesgo presente dividido entre el riesgo de enfermarse las personas del grupo sin el factor de riesgo.

$$\text{Riesgo relativo} = \frac{Incidencia - en - \exp uestos}{Incidencia - en - no - \exp uestos}$$

$$\text{Riesgo relativo} = \frac{Ie}{Io} = \frac{a/a+b}{c/c+d}$$

En relación con tratamientos, el riesgo relativo es el cuociente entre el riesgo en un grupo tratado y el riesgo en el grupo control. Es una medida de la eficiencia de un tratamiento. Si es igual a 1, el efecto del tratamiento no es distinto del efecto del control. Si el riesgo relativo es mayor (o menor) de 1, el efecto del tratamiento es mayor (o menor) que el del control.

Debe tenerse en cuenta que mientras más alto es el riesgo relativo, mayor es la probabilidad de que haya una relación de causa a efecto, es decir, que la presencia de la patología está asociada a la presencia del factor de riesgo.

El riesgo relativo se estima de la siguiente manera: Supongamos que en un grupo de trabajadores de una fábrica de cal algunos fuman y algunos sufren de enfermedades respiratorias. Para determinar el riesgo relativo, se toma el número de los que fuman y sufren enfermedades respiratorias, de los que fuman y no

sufren enfermedades respiratorias, de los que no fuman y sufren enfermedades respiratorias y de los que no fuman y no sufren enfermedades respiratorias. Se hace una prueba de independencia (χ^2) y si hay diferencias significativas, se procede a la estimación del riesgo relativo mediante la determinación de la relación de ventaja. En el grupo de trabajadores de la fábrica de cal, 450 fumaban y padecían de enfermedades respiratorias, 200 fumaban pero no sufrían de enfermedades respiratorias, 50 no fumaban y padecían de enfermedades respiratorias y 250 no fumaban y no padecían de enfermedades respiratorias.

Tabla 22. Tabla de Contingencia de 2 x 2, para determinación de Riesgo Relativo y Relación de Ventaja.

		Patología (Enfermedad respiratoria)		
		Presente +	Ausente -	Σ
Factor de riesgo (Fumar)	Fuman +	+ + a 450	- + b 200	Fuman a + b 650
	No fuman -	+ - c 50	- - d 250	No fuman c + d 300
	Σ	Patología Presente (Enfermos) a + c 500	Patología Ausente (Sanos) b + d 450	Total a + b +c +d = n 950

$$\chi^2 = \frac{n\left(|ad-bc|-\frac{1}{2}n\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

$\chi^2 = 225.6032$ $p < 0.001$. Hay diferencias estadísticamente significativas, es decir, la prueba nos indica que hay dependencia de la enfermedad en el hábito de fumar. Entonces se procede a estimar el riesgo relativo mediante la relación de ventaja, lo cual puede hacerse de dos formas diferentes:

1) Relación de ventaja = $\frac{ad}{bc}$

Relación de ventaja = $\frac{450\,x\,250}{200\,x\,50}$

Relación de ventaja = $\frac{112500}{10000}$

Relación de ventaja = 11.25

2) Relación de ventaja = $\frac{a/c}{b/d}$

Relación de ventaja = $\frac{450/50}{200/250}$

Relación de ventaja = $\frac{9}{0.8}$

Relación de ventaja = 11.25

Esto indica que en este grupo, los trabajadores que fuman tienen 11.25 veces más probabilidad de padecer enfermedades respiratorias que los que no fuman.

Reducción Absoluta del Riesgo (RAR). Es igual al riesgo atribuible (RA).

Riesgo Atribuible (RA). Es la proporción de casos expuestos (al factor de riesgo) que no hubieran adquirido la enfermedad si no hubiesen estado expuestos (al factor de riesgo).

REFERENCIAS

Fisher, R. A., Yates, F. 1963. Statistical tables for biological, agricultural and medical research. 6th ed. Oliver and Boyd. Edinburgh and London. 146 p.

Panse, V. G., Sukhatme, P.V. 1959. Métodos estadísticos para investigadores agrícolas. Fondo de Cultura Económica. México D. F. 349 p.

Snedecor, G. W. 1956. Statistical methods. The Iowa State University Press. Ames, Iowa. U. S. A. 534 p.

Snedecor, G. W., Cochran W. G. 1989. Statistical methods. 8th ed. Blackwell. New York.

Sokal, R. R., Rohlf, F. J. 1994. Biometry. 3rd ed. W. H. Freeman. New York. 896 p.

Spiegel, M. R. 1991. Estadística. Serie Shaum. 2ª ed. McGraw-Hill. Madrid. 556 p.

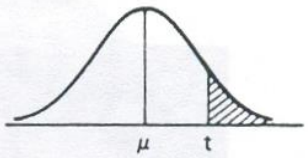
ANEXOS

Anexo 1. Tabla de t de Student.

Anexo 2. Tabla de ji cuadrado.

Anexo 3. Tabla de F (Razón de las Varianzas)

Anexo 1.
Tabla de t de Student



df	$t_{.10}$	$t_{.05}$	$t_{.025}$	$t_{.01}$	$t_{.005}$
1	3.078	6.3138	12.706	31.821	63.657
2	1.886	2.9200	4.3027	6.965	9.9248
3	1.638	2.3534	3.1825	4.541	5.8409
4	1.533	2.1318	2.7764	3.747	4.6041
5	1.476	2.0150	2.5706	3.365	4.0321
6	1.440	1.9432	2.4469	3.143	3.7074
7	1.415	1.8946	2.3646	2.998	3.4995
8	1.397	1.8595	2.3060	2.896	3.3554
9	1.383	1.8331	2.2622	2.821	3.2498
10	1.372	1.8125	2.2281	2.764	3.1693
11	1.363	1.7959	2.2010	2.718	3.1058
12	1.356	1.7823	2.1788	2.681	3.0545
13	1.350	1.7709	2.1604	2.650	3.0123
14	1.345	1.7613	2.1448	2.624	2.9768
15	1.341	1.7530	2.1315	2.602	2.9467
16	1.337	1.7459	2.1199	2.583	2.9208
17	1.333	1.7396	2.1098	2.567	2.8982
18	1.330	1.7341	2.1009	2.552	2.8784
19	1.328	1.7291	2.0930	2.539	2.8609
20	1.325	1.7247	2.0860	2.528	2.8453
21	1.323	1.7207	2.0796	2.518	2.8314
22	1.321	1.7171	2.0739	2.508	2.8188
23	1.319	1.7139	2.0687	2.500	2.8073
24	1.318	1.7109	2.0639	2.492	2.7969
25	1.316	1.7081	2.0595	2.485	2.7874
26	1.315	1.7056	2.0555	2.479	2.7787
27	1.314	1.7033	2.0518	2.473	2.7707
28	1.313	1.7011	2.0484	2.467	2.7633
29	1.311	1.6991	2.0452	2.462	2.7564
30	1.310	1.6973	2.0423	2.457	2.7500
35	1.3062	1.6896	2.0301	2.438	2.7239
40	1.3031	1.6839	2.0211	2.423	2.7045
45	1.3007	1.6794	2.0141	2.412	2.6896
50	1.2987	1.6759	2.0086	2.403	2.6778
60	1.2959	1.6707	2.0003	2.390	2.6603
70	1.2938	1.6669	1.9945	2.381	2.6480
80	1.2922	1.6641	1.9901	2.374	2.6388
90	1.2910	1.6620	1.9867	2.368	2.6316
100	1.2901	1.6602	1.9840	2.364	2.6260
120	1.2887	1.6577	1.9799	2.358	2.6175
140	1.2876	1.6558	1.9771	2.353	2.6114
160	1.2869	1.6545	1.9749	2.350	2.6070
180	1.2863	1.6534	1.9733	2.347	2.6035
200	1.2858	1.6525	1.9719	2.345	2.6006
∞	1.282	1.645	1.96	2.326	2.576

Anexo 2
Tabla de ji (χ^2) cuadrado

Probability.

n	·99	·98	·95	·90	·80	·70	·50	·30	·20	·10	·05	·02	·01	·001
1	$\cdot 0^3157$	$\cdot 0^3628$	·00393	·0158	·0642	·148	·455	1·074	1·642	2·706	3·841	5·412	6·635	10·827
2	·0201	·0404	·103	·211	·446	·713	1·386	2·408	3·219	4·605	5·991	7·824	9·210	13·815
3	·115	·185	·352	·584	1·005	1·424	2·366	3·665	4·642	6·251	7·815	9·837	11·345	16·266
4	·297	·429	·711	1·064	1·649	2·195	3·357	4·878	5·989	7·779	9·488	11·668	13·277	18·467
5	·554	·752	1·145	1·610	2·343	3·000	4·351	6·064	7·289	9·236	11·070	13·388	15·086	20·515
6	·872	1·134	1·635	2·204	3·070	3·828	5·348	7·231	8·558	10·645	12·592	15·033	16·812	22·457
7	1·239	1·564	2·167	2·833	3·822	4·671	6·346	8·383	9·803	12·017	14·067	16·622	18·475	24·322
8	1·646	2·032	2·733	3·490	4·594	5·527	7·344	9·524	11·030	13·362	15·507	18·168	20·090	26·125
9	2·088	2·532	3·325	4·168	5·380	6·393	8·343	10·656	12·242	14·684	16·919	19·679	21·666	27·877
10	2·558	3·059	3·940	4·865	6·179	7·267	9·342	11·781	13·442	15·987	18·307	21·161	23·209	29·588
11	3·053	3·609	4·575	5·578	6·989	8·148	10·341	12·899	14·631	17·275	19·675	22·618	24·725	31·264
12	3·571	4·178	5·226	6·304	7·807	9·034	11·340	14·011	15·812	18·549	21·026	24·054	26·217	32·909
13	4·107	4·765	5·892	7·042	8·634	9·926	12·340	15·119	16·985	19·812	22·362	25·472	27·688	34·528
14	4·660	5·368	6·571	7·790	9·467	10·821	13·339	16·222	18·151	21·064	23·685	26·873	29·141	36·123
15	5·229	5·985	7·261	8·547	10·307	11·721	14·339	17·322	19·311	22·307	24·996	28·259	30·578	37·697
16	5·812	6·614	7·962	9·312	11·152	12·624	15·338	18·418	20·465	23·542	26·296	29·633	32·000	39·252
17	6·408	7·255	8·672	10·085	12·002	13·531	16·338	19·511	21·615	24·769	27·587	30·995	33·409	40·790
18	7·015	7·906	9·390	10·865	12·857	14·440	17·338	20·601	22·760	25·989	28·869	32·346	34·805	42·312
19	7·633	8·567	10·117	11·651	13·716	15·352	18·338	21·689	23·900	27·204	30·144	33·687	36·191	43·820
20	8·260	9·237	10·851	12·443	14·578	16·266	19·337	22·775	25·038	28·412	31·410	35·020	37·566	45·315
21	8·897	9·915	11·591	13·240	15·445	17·182	20·337	23·858	26·171	29·615	32·671	36·343	38·932	46·797
22	9·542	10·600	12·338	14·041	16·314	18·101	21·337	24·939	27·301	30·813	33·924	37·659	40·289	48·268
23	10·196	11·293	13·091	14·848	17·187	19·021	22·337	26·018	28·429	32·007	35·172	38·968	41·638	49·728
24	10·856	11·992	13·848	15·659	18·062	19·943	23·337	27·096	29·553	33·196	36·415	40·270	42·980	51·179
25	11·524	12·697	14·611	16·473	18·940	20·867	24·337	28·172	30·675	34·382	37·652	41·566	44·314	52·620
26	12·198	13·409	15·379	17·292	19·820	21·792	25·336	29·246	31·795	35·563	38·885	42·856	45·642	54·052
27	12·879	14·125	16·151	18·114	20·703	22·719	26·336	30·319	32·912	36·741	40·113	44·140	46·963	55·476
28	13·565	14·847	16·928	18·939	21·588	23·647	27·336	31·391	34·027	37·916	41·337	45·419	48·278	56·893
29	14·256	15·574	17·708	19·768	22·475	24·577	28·336	32·461	35·139	39·087	42·557	46·693	49·588	58·302
30	14·953	16·306	18·493	20·599	23·364	25·508	29·336	33·530	36·250	40·256	43·773	47·962	50·892	59·703
32	16·362	17·783	20·072	22·271	25·148	27·373	31·336	35·665	38·466	42·585	46·194	50·487	53·486	62·487
34	17·789	19·275	21·664	23·952	26·938	29·242	33·336	37·795	40·676	44·903	48·602	52·995	56·061	65·247
36	19·233	20·783	23·269	25·643	28·735	31·115	35·336	39·922	42·879	47·212	50·999	55·489	58·619	67·985
38	20·691	22·304	24·884	27·343	30·537	32·992	37·335	42·045	45·076	49·513	53·384	57·969	61·162	70·703
40	22·164	23·838	26·509	29·051	32·345	34·872	39·335	44·165	47·269	51·805	55·759	60·436	63·691	73·402
42	23·650	25·383	28·144	30·765	34·157	36·755	41·335	46·282	49·456	54·090	58·124	62·892	66·206	76·084
44	25·148	26·939	29·787	32·487	35·974	38·641	43·335	48·396	51·639	56·369	60·481	65·337	68·710	78·750
46	26·657	28·504	31·439	34·215	37·795	40·529	45·335	50·507	53·818	58·641	62·830	67·771	71·201	81·400
48	28·177	30·080	33·098	35·949	39·621	42·420	47·335	52·616	55·993	60·907	65·171	70·197	73·683	84·037
50	29·707	31·664	34·764	37·689	41·449	44·313	49·335	54·723	58·164	63·167	67·505	72·613	76·154	86·661
52	31·246	33·256	36·437	39·433	43·281	46·209	51·335	56·827	60·332	65·422	69·832	75·021	78·616	89·272
54	32·793	34·856	38·116	41·183	45·117	48·106	53·335	58·930	62·496	67·673	72·153	77·422	81·069	91·872
56	34·350	36·464	39·801	42·937	46·955	50·005	55·335	61·031	64·658	69·919	74·468	79·815	83·513	94·461
58	35·913	38·078	41·492	44·696	48·797	51·906	57·335	63·129	66·816	72·160	76·778	82·201	85·950	97·039
60	37·485	39·699	43·188	46·459	50·641	53·809	59·335	65·227	68·972	74·397	79·082	84·580	88·379	99·607
62	39·063	41·327	44·889	48·226	52·487	55·714	61·335	67·322	71·125	76·630	81·381	86·953	90·802	102·166
64	40·649	42·960	46·595	49·996	54·336	57·620	63·335	69·416	73·276	78·860	83·675	89·320	93·217	104·716
66	42·240	44·599	48·305	51·770	56·188	59·527	65·335	71·508	75·424	81·085	85·965	91·681	95·626	107·258
68	43·838	46·244	50·020	53·548	58·042	61·436	67·335	73·600	77·571	83·308	88·250	94·037	98·028	109·791
70	45·442	47·893	51·739	55·329	59·898	63·346	69·334	75·689	79·715	85·527	90·531	96·388	100·425	112·317

For odd values of n between 30 and 70 the mean of the tabular values for $n-1$ and $n+1$ may be taken. For larger values of n, the expression $\sqrt{2\chi^2}-\sqrt{2n-1}$ may be used as a normal deviate with unit variance, remembering that the probability for χ^2 corresponds with that of a single tail of the normal curve. (For fuller formulæ see Introduction.)

Anexo 3.
Tablas de F (Razón de las Varianzas)

5 Per Cent. Points of e^{2z}

n_2 \ n_1	1	2	3	4	5	6	8	12	24	∞
1	161·4	199·5	215·7	224·6	230·2	234·0	238·9	243·9	249·0	254·3
2	18·51	19·00	19·16	19·25	19·30	19·33	19·37	19·41	19·45	19·50
3	10·13	9·55	9·28	9·12	9·01	8·94	8·84	8·74	8·64	8·53
4	7·71	6·94	6·59	6·39	6·26	6·16	6·04	5·91	5·77	5·63
5	6·61	5·79	5·41	5·19	5·05	4·95	4·82	4·68	4·53	4·36
6	5·99	5·14	4·76	4·53	4·39	4·28	4·15	4·00	3·84	3·67
7	5·59	4·74	4·35	4·12	3·97	3·87	3·73	3·57	3·41	3·23
8	5·32	4·46	4·07	3·84	3·69	3·58	3·44	3·28	3·12	2·93
9	5·12	4·26	3·86	3·63	3·48	3·37	3·23	3·07	2·90	2·71
10	4·96	4·10	3·71	3·48	3·33	3·22	3·07	2·91	2·74	2·54
11	4·84	3·98	3·59	3·36	3·20	3·09	2·95	2·79	2·61	2·40
12	4·75	3·88	3·49	3·26	3·11	3·00	2·85	2·69	2·50	2·30
13	4·67	3·80	3·41	3·18	3·02	2·92	2·77	2·60	2·42	2·21
14	4·60	3·74	3·34	3·11	2·96	2·85	2·70	2·53	2·35	2·13
15	4·54	3·68	3·29	3·06	2·90	2·79	2·64	2·48	2·29	2·07
16	4·49	3·63	3·24	3·01	2·85	2·74	2·59	2·42	2·24	2·01
17	4·45	3·59	3·20	2·96	2·81	2·70	2·55	2·38	2·19	1·96
18	4·41	3·55	3·16	2·93	2·77	2·66	2·51	2·34	2·15	1·92
19	4·38	3·52	3·13	2·90	2·74	2·63	2·48	2·31	2·11	1·88
20	4·35	3·49	3·10	2·87	2·71	2·60	2·45	2·28	2·08	1·84
21	4·32	3·47	3·07	2·84	2·68	2·57	2·42	2·25	2·05	1·81
22	4·30	3·44	3·05	2·82	2·66	2·55	2·40	2·23	2·03	1·78
23	4·28	3·42	3·03	2·80	2·64	2·53	2·38	2·20	2·00	1·76
24	4·26	3·40	3·01	2·78	2·62	2·51	2·36	2·18	1·98	1·73
25	4·24	3·38	2·99	2·76	2·60	2·49	2·34	2·16	1·96	1·71
26	4·22	3·37	2·98	2·74	2·59	2·47	2·32	2·15	1·95	1·69
27	4·21	3·35	2·96	2·73	2·57	2·46	2·30	2·13	1·93	1·67
28	4·20	3·34	2·95	2·71	2·56	2·44	2·29	2·12	1·91	1·65
29	4·18	3·33	2·93	2·70	2·54	2·43	2·28	2·10	1·90	1·64
30	4·17	3·32	2·92	2·69	2·53	2·42	2·27	2·09	1·89	1·62
40	4·08	3·23	2·84	2·61	2·45	2·34	2·18	2·00	1·79	1·51
60	4·00	3·15	2·76	2·52	2·37	2·25	2·10	1·92	1·70	1·39
120	3·92	3·07	2·68	2·45	2·29	2·17	2·02	1·83	1·61	1·25
∞	3·84	2·99	2·60	2·37	2·21	2·10	1·94	1·75	1·52	1·00

Lower 5 per cent. points are found by interchange of n_1 and n_2, *i.e.* n_1 must always correspond with the greater mean square.

1 Per Cent. Points of e^{2z}

n_2 \ n_1	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5982	6106	6234	6366
2	98·50	99·00	99·17	99·25	99·30	99·33	99·37	99·42	99·46	99·50
3	34·12	30·82	29·46	28·71	28·24	27·91	27·49	27·05	26·60	26·12
4	21·20	18·00	16·69	15·98	15·52	15·21	14·80	14·37	13·93	13·46
5	16·26	13·27	12·06	11·39	10·97	10·67	10·29	9·89	9·47	9·02
6	13·74	10·92	9·78	9·15	8·75	8·47	8·10	7·72	7·31	6·88
7	12·25	9·55	8·45	7·85	7·46	7·19	6·84	6·47	6·07	5·65
8	11·26	8·65	7·59	7·01	6·63	6·37	6·03	5·67	5·28	4·86
9	10·56	8·02	6·99	6·42	6·06	5·80	5·47	5·11	4·73	4·31
10	10·04	7·56	6·55	5·99	5·64	5·39	5·06	4·71	4·33	3·91
11	9·65	7·20	6·22	5·67	5·32	5·07	4·74	4·40	4·02	3·60
12	9·33	6·93	5·95	5·41	5·06	4·82	4·50	4·16	3·78	3·36
13	9·07	6·70	5·74	5·20	4·86	4·62	4·30	3·96	3·59	3·16
14	8·86	6·51	5·56	5·03	4·69	4·46	4·14	3·80	3·43	3·00
15	8·68	6·36	5·42	4·89	4·56	4·32	4·00	3·67	3·29	2·87
16	8·53	6·23	5·29	4·77	4·44	4·20	3·89	3·55	3·18	2·75
17	8·40	6·11	5·18	4·67	4·34	4·10	3·79	3·45	3·08	2·65
18	8·28	6·01	5·09	4·58	4·25	4·01	3·71	3·37	3·00	2·57
19	8·18	5·93	5·01	4·50	4·17	3·94	3·63	3·30	2·92	2·49
20	8·10	5·85	4·94	4·43	4·10	3·87	3·56	3·23	2·86	2·42
21	8·02	5·78	4·87	4·37	4·04	3·81	3·51	3·17	2·80	2·36
22	7·94	5·72	4·82	4·31	3·99	3·76	3·45	3·12	2·75	2·31
23	7·88	5·66	4·76	4·26	3·94	3·71	3·41	3·07	2·70	2·26
24	7·82	5·61	4·72	4·22	3·90	3·67	3·36	3·03	2·66	2·21
25	7·77	5·57	4·68	4·18	3·86	3·63	3·32	2·99	2·62	2·17
26	7·72	5·53	4·64	4·14	3·82	3·59	3·29	2·96	2·58	2·13
27	7·68	5·49	4·60	4·11	3·78	3·56	3·26	2·93	2·55	2·10
28	7·64	5·45	4·57	4·07	3·75	3·53	3·23	2·90	2·52	2·06
29	7·60	5·42	4·54	4·04	3·73	3·50	3·20	2·87	2·49	2·03
30	7·56	5·39	4·51	4·02	3·70	3·47	3·17	2·84	2·47	2·01
40	7·31	5·18	4·31	3·83	3·51	3·29	2·99	2·66	2·29	1·80
60	7·08	4·98	4·13	3·65	3·34	3·12	2·82	2·50	2·12	1·60
120	6·85	4·79	3·95	3·48	3·17	2·96	2·66	2·34	1·95	1·38
∞	6·64	4·60	3·78	3·32	3·02	2·80	2·51	2·18	1·79	1·00

Lower 1 per cent. points are found by interchange of n_1 and n_2, *i.e.* n_1 must always correspond with the greater mean square.

0·1 Per Cent. Points of e^{2z}

n_2 \ n_1	1	2	3	4	5	6	8	12	24	∞
1	405284	500000	540379	562500	576405	585937	598144	610667	623497	636619
2	998·5	999·0	999·2	999·2	999·3	999·3	999·4	999·4	999·5	999·5
3	167·0	148·5	141·1	137·1	134·6	132·8	130·6	128·3	125·9	123·5
4	74·14	61·25	56·18	53·44	51·71	50·53	49·00	47·41	45·77	44·05
5	47·18	37·12	33·20	31·09	29·75	28·84	27·64	26·42	25·14	23·78
6	35·51	27·00	23·70	21·92	20·81	20·03	19·03	17·99	16·89	15·75
7	29·25	21·69	18·77	17·19	16·21	15·52	14·63	13·71	12·73	11·69
8	25·42	18·49	15·83	14·39	13·49	12·86	12·04	11·19	10·30	9·34
9	22·86	16·39	13·90	12·56	11·71	11·13	10·37	9·57	8·72	7·81
10	21·04	14·91	12·55	11·28	10·48	9·92	9·20	8·45	7·64	6·76
11	19·69	13·81	11·56	10·35	9·58	9·05	8·35	7·63	6·85	6·00
12	18·64	12·97	10·80	9·63	8·89	8·38	7·71	7·00	6·25	5·42
13	17·81	12·31	10·21	9·07	8·35	7·86	7·21	6·52	5·78	4·97
14	17·14	11·78	9·73	8·62	7·92	7·43	6·80	6·13	5·41	4·60
15	16·59	11·34	9·34	8·25	7·57	7·09	6·47	5·81	5·10	4·31
16	16·12	10·97	9·00	7·94	7·27	6·81	6·19	5·55	4·85	4·06
17	15·72	10·66	8·73	7·68	7·02	6·56	5·96	5·32	4·63	3·85
18	15·38	10·39	8·49	7·46	6·81	6·35	5·76	5·13	4·45	3·67
19	15·08	10·16	8·28	7·26	6·62	6·18	5·59	4·97	4·29	3·52
20	14·82	9·95	8·10	7·10	6·46	6·02	5·44	4·82	4·15	3·38
21	14·59	9·77	7·94	6·95	6·32	5·88	5·31	4·70	4·03	3·26
22	14·38	9·61	7·80	6·81	6·19	5·76	5·19	4·58	3·92	3·15
23	14·19	9·47	7·67	6·69	6·08	5·65	5·09	4·48	3·82	3·05
24	14·03	9·34	7·55	6·59	5·98	5·55	4·99	4·39	3·74	2·97
25	13·88	9·22	7·45	6·49	5·88	5·46	4·91	4·31	3·66	2·89
26	13·74	9·12	7·36	6·41	5·80	5·38	4·83	4·24	3·59	2·82
27	13·61	9·02	7·27	6·33	5·73	5·31	4·76	4·17	3·52	2·75
28	13·50	8·93	7·19	6·25	5·66	5·24	4·69	4·11	3·46	2·70
29	13·39	8·85	7·12	6·19	5·59	5·18	4·64	4·05	3·41	2·64
30	13·29	8·77	7·05	6·12	5·53	5·12	4·58	4·00	3·36	2·59
40	12·61	8·25	6·60	5·70	5·13	4·73	4·21	3·64	3·01	2·23
60	11·97	7·76	6·17	5·31	4·76	4·37	3·87	3·31	2·69	1·90
120	11·38	7·32	5·79	4·95	4·42	4·04	3·55	3·02	2·40	1·54
∞	10·83	6·91	5·42	4·62	4·10	3·74	3·27	2·74	2·13	1·00

Lower 0·1 per cent. points are found by interchange of n_1 and n_2, *i.e.* n_1 must always correspond with the greater mean square.

etc.. Es Miembro de cuatro de las seis Comisiones de la UICN, The World Conservation Union. Ha sido organizador, directivo, invitado especial, conferencista o ponente en numerosas reuniones científicas, nacionales o internacionales. Ha sido jurado en concursos internacionales. Es miembro del World Cultural Council. Ha sido invitado a diferentes universidades nacionales y extranjeras a dictar cursos, conferencias o asesorar investigaciones. Su semblanza aparece en el Who is Who in Conservation y en el Who is Who in Medicine. Ha sido consultor para UICN, BID, Universidad Nacional de Piura (Perú). Ha recibido diferentes premios y condecoraciones, incluyendo la Orden "Henri Pittier" 1ª y 2ª Clases, la más alta condecoración por méritos en conservación de la naturaleza en Venezuela. Ha publicado varios libros y artículos en revistas nacionales e internacionales. Ha sido y es Editor y/o Asesor de varias revistas científicas. Su libro "Iniciación Práctica a la Investigación Científica" (actualmente agotado) es el texto en "Metodología de la Investigación Científica" en las Facultades de Medicina y de Odontología de la Universidad de Los Andes y es usado en varios países latinoamericanos; recibió el premio "Libro Dorado" (al libro más vendido de la Universidad de Los Andes), durante los dos años de existencia de ese premio. Es directivo y/o miembro de varias asociaciones científicas nacionales e internacionales. Es Integrante del Programa de Promoción al Investigador (PPI 1507) y ocupó el primer lugar en su Facultad en dos convocatorias (1999 y 2001) del Programa de Estímulo al Investigador (PEI ULA).

ESTADÍSTICA PARA INVESTIGADORES

La estadística es la disciplina científica que se encarga de recolectar, ordenar y analizar datos numéricos para sacar conclusiones que sirvan de apoyo a las actividades humanas,
cualesquiera que sean.

Este texto pretende enseñar en la forma más sencilla, clara, precisa y amena posible, el uso de las técnicas de la estadística más comúnmente usadas por los investigadores, tanto científicos como sociales, desde sus componentes más elementales (media, moda, mediana, etc.) hasta otras más avanzadas (análisis de regresión, de correlación, de varianza, etc.). Se hace énfasis en el cálculo manual (hasta con solo papel y lápiz) de todas las fórmulas con el fin de que cada lector pueda entender de dónde salen las diferentes pruebas, qué significan los valores obtenidos, cómo se deben interpretar y cómo pueden aplicarse en diferentes situaciones. En vista de la importancia que tiene la estadística para las ciencias de la salud, el capítulo final se dedica a diferentes aspectos básicos en epidemiología.

Todos los aspectos tratados están apoyados con ejemplos, paso a paso, y su posterior interpretación. De esta forma el investigador que realice sus cálculos mediante los programas de computación de uso común (SAS, SPSS, etc.) podrá interpretar apropiadamente los resultados obtenidos.

CDCHTA
ULA

ISBN: 978-980-12-5034-0

9 789801 250340