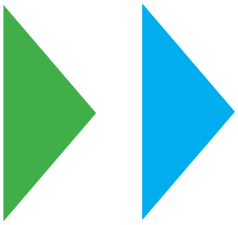
ESTADÍSTICA PARA ADMINISTRADORES

MAWIL
Publicaciones Impresas
y Digitales

www.mawil.us

Estadística

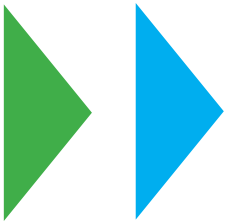


para
Administradores

MAWIL
Publicaciones Impresas
y Digitales

www.mawil.us

Estadística



para Administradores

Romero Izurieta Richard
Magister en Administración de Empresas mención en Negocios Internacionales;
Ingeniero en Computación
Docente de la Universidad de Guayaquil
richard.romeroi@ug.edu.ec

Puente Tituaña Silvia Paulina
Magister en Finanzas; Ingeniera Comercial
Docente de la Universidad de Guayaquil
silvia.puentet@ug.edu.ec

Espinoza Roca Leonardo Roberto
Magister en Docencia y Gerencia en Educación Superior;
Especialista en Gerencia en Educación Superior; Diploma Superior en Docencia y Evaluación en la Educación Superior; Ingeniero Comercial
Docente de la Universidad de Guayaquil
leonardo.espinozar@ug.edu.ec

Gonzalez Gonzalez Rosa de los Ángeles
Magister en Administración de Empresas con mención en Proyectos Estratégicos y Negocios Fiduciarios; Diploma Superior en Gerencia Estratégica de Mercadeo; Ingeniera Comercial y Empresarial Especialización Comercio Exterior y Marketing
Docente de la Universidad de Guayaquil
rosa.gonzalezg@ug.edu.ec

Arce Recalde Jazmin Katherine
Magister en Administración de Empresas;
Licenciada en Sistemas de Información; Analista de Sistemas
Docente de la Universidad de Guayaquil
jazmin.arcer@ug.edu.ec

Solis Marin Betzabeth Johanna
Magister en Negocios Internacionales y Gestión de Comercio Exterior; Diploma Superior en Economía Internacional; Ingeniero en Gestión Empresarial Internacional
Docente de la Universidad de Guayaquil
betzabeth.solism@ug.edu.ec

MAWIL
Publicaciones Impresas y Digitales

DATOS DE CATALOGACIÓN

AUTORES: Romero Izurieta Richard
Puente Tituaña Silvia Paulina
Espinoza Roca Leonardo Roberto
Gonzalez Gonzalez Rosa de los Ángeles
Arce Recalde Jazmin Katherine
Solis Marin Betzabeth Johanna

Título: Estadística para Administradores

Descriptores: Matemáticas; Probabilidades; Conjuntos; Variancia; Hipótesis.

Edición: 1era

ISBN: 978-9942-787-44-6

Editorial: Mawil Publicaciones de Ecuador, 2018

Área: Educación Superior

Formato: 148 x 210 mm.

Páginas: 204

DOI: 10.26820/mawil/estadística-para-administradores

MAWIL
Publicaciones Impresas y Digitales

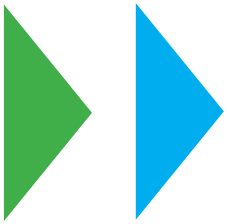
Texto para Docentes y Estudiantes Universitarios

El proyecto didáctico *Estadística para Administradores*, es una obra colectiva creada por sus autores y publicada por *MAWIL*; publicación revisada por el equipo profesional y editorial siguiendo los lineamientos y estructuras establecidos por el departamento de publicaciones de *MAWIL* de New Jersey.

© Reservados todos los derechos. La reproducción parcial o total queda estrictamente prohibida, sin la autorización expresa de los autores, bajo sanciones establecidas en las leyes, por cualquier medio o procedimiento.

*Director General: Mg. Vanessa Quishpe Morocho
*Dirección Central MAWIL: Office 18 Center Avenue Caldwell; New Jersey # 07006
*Gerencia Editorial MAWIL-Ecuador: Lenin Suasnabas
*Editor de Arte y Diseño: Alejandro Plúa

Estadística



para Administradores

REVISORES

Econ. Xavier Francisco Flores Torres Msc.
xavier.florest@ug.edu.ec
Docente de la Universidad de Guayaquil
Facultad de Ciencias Administrativas
Coordinador de Logística Nivelación y Admisión

Ing. Verónica Patricia Silva Ortega MGP.
veronica.silvaor@ug.edu.ec
Docente de la Universidad de Guayaquil
Doctorando Pontificia
Universidad Católica de Santa María de Buenos Aires
en Administración de Empresas

PROLOGO

Las técnicas destinadas a ordenar y disponer de recursos, conforme se involucran grandes cantidades de información sobre los territorios, bienes y personas del dominio de una organización dada, requiere de un rigor científico que dote de solidez a las herramientas clásicas de decisión administrativa, y más aquellas que provienen del sentido común y la intuición.

Las nociones básicas de la administración, inversión y retorno, necesitan de formalidades que solo la estadística le puede proporcionar. De allí que sea necesario seguir varios cursos básicos de estadística que cubran los contenidos necesario para medir y estimar tendencias, situaciones de azar y riesgo, razonamientos que habiliten al administrador para tomar resoluciones con recompensas positivas, razonables y efectivas.

Aunque pueda parecer arduo estudiar estadística, dada la cantidad de conceptos matemáticos que se necesita dominar previamente, no debe perderse de vista que se trata de una disciplina científica que dota de rigor al sentido común en el análisis de datos, la descripción de su información intrínseca y la inferencia de conclusiones plausibles a partir de ellos. Son, pues, ideas que todos pueden entender y poner a su servicio con tan solo un poco de paciencia, disciplina e interés.

Estas notas están dispuestas a facilitar el aprendizaje, haciendo énfasis en los conceptos básicos, su interrelación y ejemplo práctico. El lector dispuesto podrá incursionar en ellas con la certeza de que ampliará sus capacidades intelectuales, sus conocimientos y habilidades en el ejercicio del razonamiento inferencial así como en la exposición de la información contenida por un conjunto de datos.

Ya han llegado los días en que el entendimiento estadístico se ha vuelto tan imprescindible como leer y escribir. Este libro está en función de alfabetizar en el dominio de las herramientas numéricas de la estadística para el uso de la administración.

Los autores

INTRODUCCIÓN

En el Ecuador la Administración Tributaria se vio en la obligación de combaEste texto está compuesto con el contenido para guiar dos o tres cursos de estadística para administradores. Está conformado por doce capítulos interdependientes uno del otro en un orden lineal y ascendente. Comprende tres grupos de estudio: la estadística descriptiva, la estadística inferencial y las aplicaciones, siendo el primero estudiado en los capítulos I y II, el segundo del capítulo V al VIII con un preliminar imprescindible de teoría de probabilidades, en los capítulo III y IV y las aplicaciones que complementan y profundizan a los dos grupos anteriores, en los capítulos del IX al XII.

Cada uno de los capítulos está balanceado para funcionar con dominio tanto de teoría como de práctica, salvo quizás el capítulo IX que es estrictamente teórico, y el capítulo I que también cuenta con numerosas definiciones. La longitud de cada uno es variable, así como su intensidad o dificultad. El corazón del libro está en aquellos que tratan las nociones de la inferencia estadística, y puede, junto con aquellos capítulos de probabilidades, funcionar solo, aunque lo recomendable sea seguir desde el principio. El libro cuenta además con un anexo de tablas de valores numéricos de tres distribuciones célebres.

Los prerrequisitos para aprovechar por completo a este curso son bastante simples: un dominio completo la aritmética y un poco de álgebra, nociones básicas de teoría de conjuntos, conocimiento elemental de los conjuntos numéricos y la recta real. Aunque pueda ser útil conocer un poco de análisis matemático, no es privativo saberlo.

Aunque este libro posea todas las herramientas para avanzar en él sin otra referencia, se recomienda acompañar su estudio y lectura con otros libros del tema, así como problemas resueltos y guías de ejercicio. Este es un libro básico, su destino es ser una introducción a la estadística. Su éxito se medirá en las habilidades que el lector adquiera para desenvolverse por sí mismo ante los problemas de datos que admitan un modelado probabilístico, una inferencia y una decisión en situaciones de incertidumbre, acompañado de la experiencia profesional, la literatura especializada y la tecnología.

CONTENIDO

PRÓLOGO 05

INTRODUCCIÓN 06

CAPÍTULO I:
Qué es la estadística. Los Datos y su Representación Gráfica..... 08

CAPÍTULO II:
La descripción a través de medidas numéricas 27

CAPÍTULO III:
Introducción a la teoría de las probabilidades..... 47

CAPÍTULO IV:
Distribuciones de probabilidad 72

CAPÍTULO V:
Estimación puntual. El muestreo y sus distribuciones..... ...104

CAPÍTULO VI:
Estimación por intervalo. Medidas de confianza 119

CAPÍTULO VII:
Principios de Pruebas de Hipótesis 128

CAPÍTULO VIII:
Inferencias sobre medias y varianzas..... ...146

CAPÍTULO IX:
Introducción al diseño experimental..... ...161

CAPÍTULO X:
El análisis de varianza..... ...171

CAPÍTULO XI:
Algunos fundamentos de regresión..... ...179

CAPÍTULO XII:
El análisis de datos categóricos..... ...189

TRABAJOS CITADOS 203

CAPÍTULO I

Qué es la estadística: Los datos y su representación gráfica

MAWIL
Publicaciones Impresas y Digitales

www.mawil.us

1. Qué es la estadística: Los datos y su representación gráfica

1.1 Definición de la estadística

Hoy en día el mundo, en la totalidad de sus aspectos, se presenta como una inmensa acumulación de datos. La información del estado actual de la economía, la demografía de las naciones, el promedio de consumo de combustibles fósiles per cápita, la eficacia de un novedoso fármaco y el rendimiento escolar en una escuela básica son ejemplos de cómo se presenta el estado actual de las cosas. Así, pues, a los datos numéricos de las frases anteriores se les llama estadísticas. En este sentido, para entender una extensa gama situaciones se debe estudiar a tal cúmulo de magnitudes con un método sistemático y racional.

Ante todo, debemos definir el objeto de estudio. **Los datos son informaciones de hechos recogidas en cifras y códigos, magnitudes estas que son susceptibles a ser analizadas y resumidas para su exposición, interpretación y utilización**. El método con el que se recopilan, ordenan y estudian a dichos datos se conoce como **estadística**.

No hay consenso claro al momento de definirla como ciencia o arte. De hecho, su definición ha evolucionado conforme sus herramientas han ido perfeccionándose. Se refiere que:

El término Estadística, que se utilizó hasta el siglo XVIII como una

> abreviatura de la ciencia descriptiva de los Estados, se identificó cada vez más, en el siglo XIX, con las cifras cuantitivas. (Ross, 2007, pág. 9)

Puede ser muy útil exponer un recorrido por sus múltiples definiciones a lo largo de tres siglos:

> La Estadística tiene el objetivo de realizar una representación fiable de un Estado en una época determinada (Quetelet, 1849). La Estadística es la única herramienta mediante la cual se puede conseguir una apertura en la formidable espesura de dificultades que entorpece el camino de aquellos que estudian la Ciencia del hombre (Galton, 1889). La Estadística puede considerarse (i) como el estudio de las poblaciones, (ii) como el estudio de las variaciones y (iii) como el estudio de los métodos de reducción de datos (Fisher, 1925). La estadística es la disciplina científica relativa a la recopilación, el análisis y la interpretación de datos obtenidos mediante la observación o la experimentación. Tiene una estructura coherente basada en la Teoría de la Probabilidad e incluye muchos procedimientos diferentes que contribuyen a la investigación y el desarrollo en todas las ramas de la ciencia y la tecnología (Pearson, 1936). La estadística es el nombre de la ciencia que trata de llevar a cabo inferencias bajo

> situaciones de incertidumbre; para ello, usa los números para averiguar cuestiones relativas a la naturaleza y la experiencia (Weaver, 1952). (Ross, 2007, pág. 13)

Podemos sintetizar todo lo anterior aseverando que **la estadística se define como el método matemático para describir, clasificar y analizar datos**, con el fin de aprender de ellos, alcanzar una mejor comprensión de un determinado entorno y así estar en condiciones de tomar las mejores decisiones en base a una información depurada y certera.

1.2 Definición y clasificación de los datos

Para comprender mejor el objeto de la estadística, debemos estudiar algunos conceptos, como población, muestra, variable y observación.

Una **población** es un conjunto de elementos de interés. Por ejemplo, los días del año, las personas que asisten a una universidad, el conjunto de los gastos diarios de un estudiante, etc. Una **muestra** es un subgrupo de una población determinada, como por ejemplo, los días laborables del año, las mujeres que asisten a una universidad, el gasto de un estudiante durante un fin de semana.

Una **variable** es una característica de los elementos de una pobla-

ción o de una muestra, así la benevolencia del clima durante todos los días del año, la estatura de los estudiantes universitarios, el gasto total durante un año de un estudiante extranjero. Una **observación** es un conjunto de medidas de una variable determinada.

De acuerdo a lo anterior, se pueden clasificar a los datos de muchas formas. Debemos recordar que en el fondo se tratan de magnitudes, es decir, características que son sometidas a una medición.

Según el origen de los datos, estos se clasifican en:

1. **Observados**: se definen así a los datos que se obtienen de manera espontánea y sin que el observador controle o intervenga en la producción del hecho que será medido.

2. **Experimentales**: estos datos presuponen un ensayo contralado, diseñado racionalmente para la recolección de determinada información.

El origen de los datos también permite clasificarlos en:

1. Datos de fuente.

2. Datos nuevos.

Los datos también se clasifican en cualitativos y cuantitativos:

1. **Cuantitativos**: se asocian a valores numéricos que indiquen cantidad, longitud, área, volumen, peso, entre otros. Es decir, se obtienen usando comparaciones con conjuntos numéricos, sea con una correspondencia uno a uno, por intervalos o por cocientes.

2. **Cualitativos**: se definen por etiquetas o nombres clave que se usan para asociarles un atributo que no son fáciles de cuantificar. A veces, los datos cualitativos suelen denominarse como datos categóricos, se ordenan en escalas nominales, o de acuerdo a un número si se mide, por ejemplo, intensidades.

Esta distinción es fundamental porque es la que prohíbe o posibilita la utilización de la aritmética en su procesamiento. Por lo general, la suma, resta o multiplicación de los datos cualitativos no significan nada.

Los datos cuantitativos también se clasifican en dos tipos, los datos cuantitativos discretos y los datos cuantitativos continuos:

1. **Discretos**: cuando la variable se mide con respecto a un conjunto numérico finito o numerable. Se utilizan, sobre todo, para hablar de cantidades.

2. **Continuos**: cuando la variable se mide con respecto a un conjunto numérico como los números reales, es decir, cuando se trabajan con medi-

das continuas como la longitud, el peso, las tasas de interés.

Un ejemplo de datos cuantitativos discretos es el número de llamadas recibidas en 5 minutos, son medidas de cantidad puntual. Mientras que los datos cuantitativos peso o tiempo son continuos porque entre los posibles valores de los datos no hay separación.

Finalmente, hay una forma de catalogar a los datos que obedece a cómo se obtienen las observaciones de acuerdo a su relación con el tiempo.

1. **Datos de sección transversal**: son los que se obtienen aproximadamente en el mismo momento. Son formas de tomar medidas sin que se involucre la relación temporal

2. **Datos de series de tiempo**: son aquellos adquiridos a lo largo de varios periodos. Se trata de estudiar a la pareja formada por una variable y el tiempo a lo largo de un período determinado.

Hemos visto entonces que los datos son información que se puede catalogar de acuerdo a cómo se recolecte la información de los hechos, a la naturaleza mesurable que poseen, al conjunto numérico que se asocie a su medida y a su relación con el tiempo. Podemos concluir que la información posee muchas facetas que deben ser consideradas antes de su estudio estadístico.

1.3 Tabulación y graficación de datos

Para extraer información sobre un conjunto particular de datos es útil realizar resúmenes gráficos y tabulares. Para ello es primario realizar las siguientes definiciones que forman parte del proceso de graficación e interpretación de un conjunto de datos.

- **Datos sueltos**: son los datos recolectados sin organización numérica. Así, la observación de la estatura de 100 sujetos ordenados alfabéticamente.
- **Ordenación**: una ordenación es un conjunto de datos numéricos en orden creciente o decreciente de la magnitud registrada. Normalmente se realiza por clases que no tienen números en común.
- **Rango**: a la diferencia entre el número mayor y el menor de los datos registrados se le conoce rango de los datos. (Schaum, p.35)

Antes de definir la primera herramienta tabular y gráfica, es necesario mirar un ejemplo sencillo. Una compañía A con 50 empleados tiene la información de cuántos días han dejado de trabajar cada uno de sus empleados por ajenas a sus voluntades individuales en un período de dos semanas. Dicha información se recoge como datos sueltos en la siguiente tabla que, por razones de espacios, omite la mayoría de los datos. Se destaca que el rango de los datos es nueve.

Trabajador	1	2	3	4	5	…	48	49	50
Cantidad de días sin trabajar	2	2	0	0	5	…	7	5	1

Tab. 1.1

El director de recursos humano desea saber cuántos trabajadores han faltado (1) Al menos un día, (2) Entre tres y cinco días inclusive y (3) Más de cinco días.

Si se quiere responder tales preguntas conviene construir una nueva tabla que prescinda de la identidad del trabajador y concentre la información sobre la cantidad de trabajadores que han faltado cero días, un día, dos días, y así hasta nueve días. A continuación una tabla que sintetiza esos datos de frecuencia:

Días de ausentismo laboral	**Frecuencia**
0	12
1	8
2	5
3	4
4	5
5	8
6	0
7	5
8	2
9	1

Tab.1.2

Dicha tabla se construyó sumando cuántos trabajadores no faltaron nunca, cuántos faltaron dos días, y así hasta contar cuántos faltaron nueve días. Nótese que el total de la suma de la fila columna de Frecuencia es 50. Ya, con dicha información tabulada, podemos decir que (1) 38 trabajadores faltaron al menos un día, (2) 4+5+8=17 trabajadores faltaron 3, 4 o 5 días y (3) Si sumamos las frecuencias relativas de 6,7, 8 y 9 días obtendremos la cantidad de trabajadores que faltaron más de 5 días, que es 8.

Al trabajar con grandes cantidades de datos, es útil distribuirlos en clases, grupos o categorías (días de ausentismo en nuestro ejemplo anterior) y determinar el número de individuos que pertenecen a cada clase, es decir, la frecuencia de cada grupo. A la disposición tabular de tal información, las clases con sus correspondientes frecuencias de clase, se le conoce como **distribución de frecuencias** o tabla de frecuencias.

No siempre las clases con puntuales, como en el ejemplo del ausentismo laboral de la compañía A. Véase por ejemplo, la siguiente tabla de frecuencias donde la clases son estaturas de un conjunto de 100 estudiantes:

Clase	Estatura (en centimetros)	Cantidad de estudiantes
1	152-157	5
2	158-165	18
3	166-172	42
4	173-180	27
5	181-187	8
Total		100

Tab. 1.3

Para este caso es necesario definir los límites de las clase como los valores más pequeño (límite inferior) y más grande (límite superior) que se registran en dicho grupo. El intervalo de clase son todos los datos comprendidos entre el límite superior y el inferior.

El tamaño o amplitud de un intervalo de clase es la diferencia entre el límite superior y el límite inferior. La marca de clase o punto medio de clase es la suma de los límites dividida entre dos.

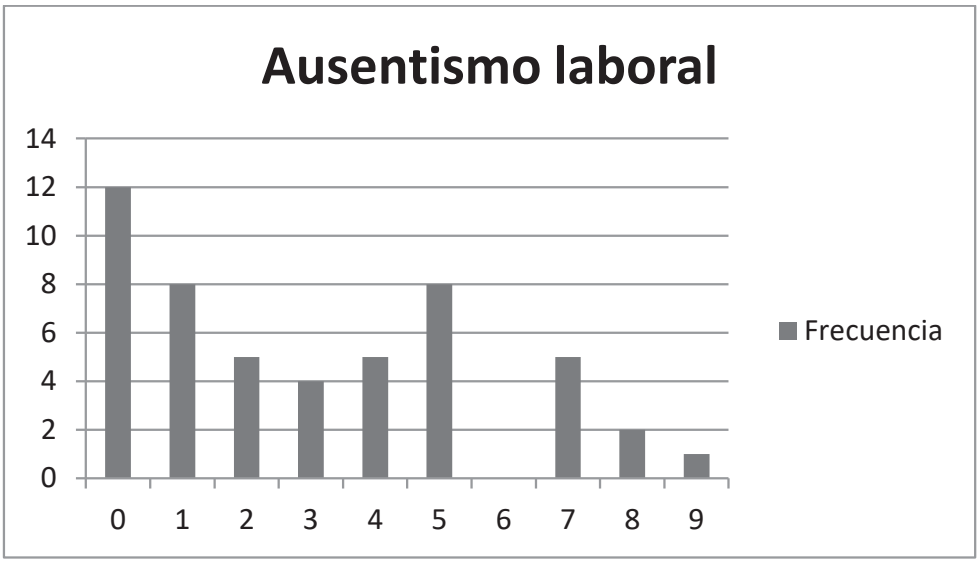
En nuestro ejemplo particular de las estaturas, hay 100 estudiantes divididos entre 5 clases de acuerdo a su estatura, sin que ninguno ocupe más de una clase. El límite inferior de la clase 1 es 152 centímetros, mientras que su límite superior es 157 centímetros. El tamaño de la primera clase es de 5 centímetros y su marca es de 154,5 centímetros.

Dado un conjunto de datos sueltos, hay una serie de pasos para construir su **tabla de frecuencias**:

1. Determinar el número mayor y el menor de los datos para calcular su rango
2. Ordenar los datos
3. Determinar el número de clases disjuntas. Se recomienda que sea superior a 5 e inferior a 20.
4. Determinar los límites de clase, así como su tamaño.
5. Determinar el número de observaciones que corresponden a cada intervalo de clase, es decir, hallar las frecuencias de cada clase.

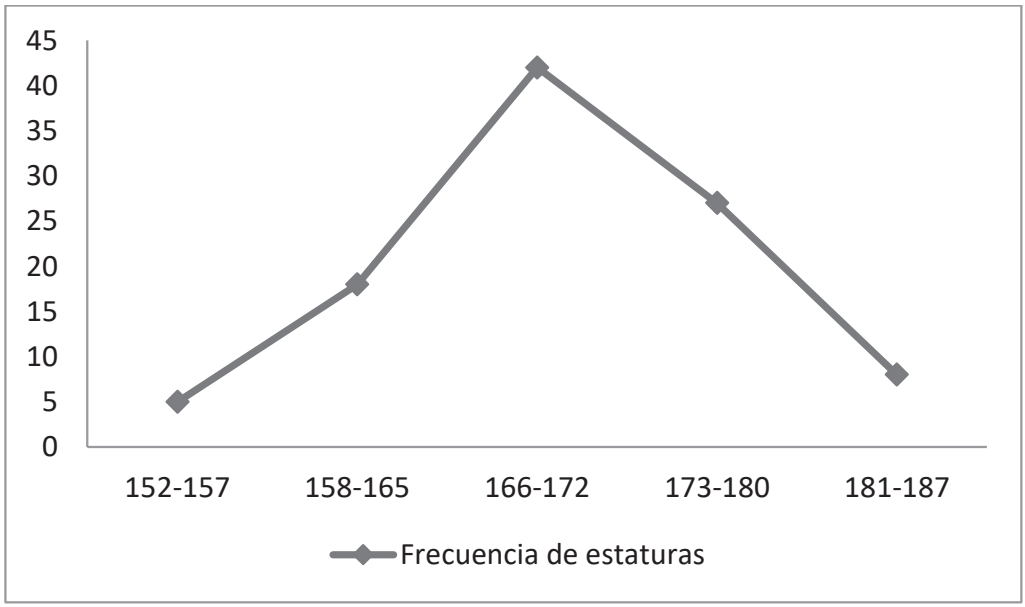
Otra forma de sintetizar más aún la información contenida en una tabla de frecuencias, consiste en dibujar en el plano cartesiano una representación gráfica de dichas tablas, colocando en el eje X a las clases y en el eje Y a las frecuencias. A dicha representación se le conoce como **histograma** y tiene múltiples presentaciones.

Así, el contenido de la tabla 1.2 en el siguiente histograma de barras:



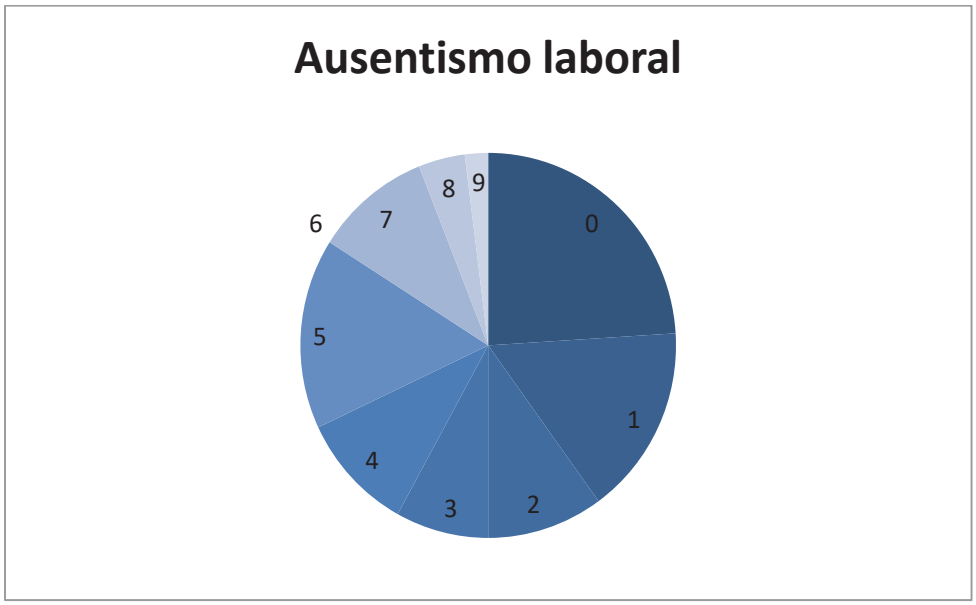
Graf. 1.1

Mientras que el contenido de la tabla de frecuencias 1.3 se muestra en el siguiente histograma poligonal:

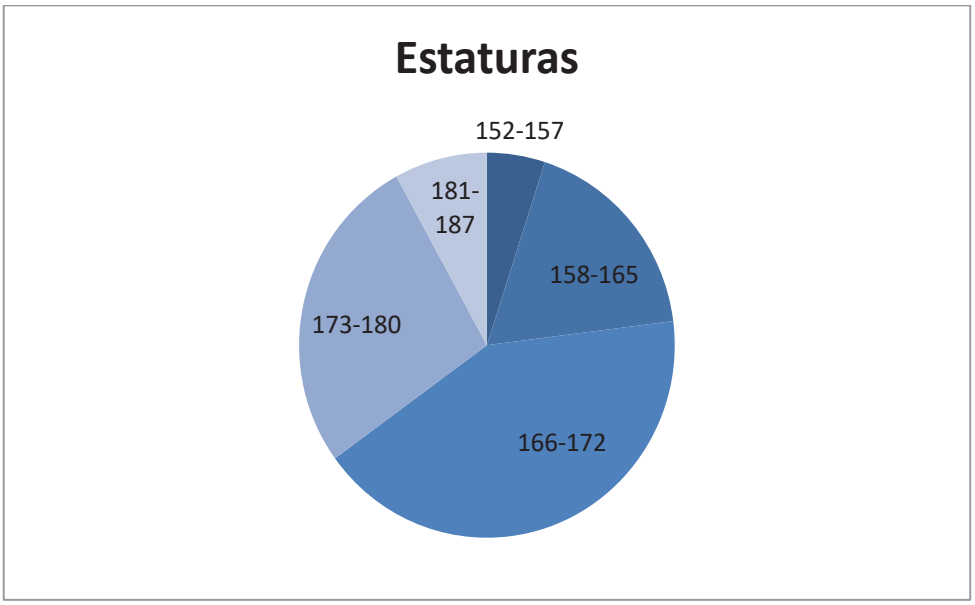


Graf. 1.2

También existen los gráficos de torta, que son círculos divididos de acuerdo a la proporción entre las frecuencias tabuladas. A continuación, la gráfica de torta de las tablas 1.2 y 1.3:



Graf. 1.3



Graf. 1.4

Otra noción importante distinguir es la **frecuencia absoluta** y la **frecuencia relativa**. Cuando contamos la cantidad de observaciones que hay por clase, nos referíamos a una frecuencia absoluta, mientras que la frecuencia relativa es el resultado de dividir la frecuencia absoluta entre la cantidad total de observaciones. Así, la tabla de frecuencias relativas del ausentismo laboral, quedaría como:

Días de ausentismo laboral	Frecuencias relativas
0	0,24
1	0,16
2	0,1
3	0,08
4	0,1
5	0,16
6	0
7	0,1
8	0,04
9	0,02

Tab.1.4

Mientras que la tabla de frecuencias relativas de las estaturas del grupo de 100 estudiantes quedaría como:

Estatura (en centimetros)	Frecuencia relativa	Proporción porcentual
152-157	0,05	5%
158-165	0,18	18%
166-172	0,42	42%
173-180	0,27	27%

181-187	0,08	8%
Total	1	100%

Tab. 1.5

Es de notar que la diferencia es simplemente un cambio en la escala. Ahora el total suma 1 y a partir de ello se puede obtener otra escala porcentual simplemente multiplicando por 100. Es usual construir los histogramas de acuerdo a la tabla de frecuencias relativas, también conocidas como distribuciones de frecuencia.

Un último concepto de utilidad e interés es el de **la frecuencia acumulada** que no es más que la frecuencia total de todos los valores menores que la frontera de clase superior de un intervalo de clase dado. A la tabulación de tales frecuencias se le conoce como tabla de distribución de frecuencias acumuladas y a la gráfica que le corresponde a esta se le denomina **ojiva**. Veamos un ejemplo de dicho concepto para la tabla 1.5

Grupo	**Estatura (en centimetros)**	**Frecuencia absoluta acumulada**	**Frecuencia relativa acumulada**
1	Menor que 152	0	0
2	Menor que 157	5	0.05
3	Menor que 165	23	0.23
4	Menor que 172	65	0.65
5	Menor que 180	92	0.92
6	Menor que 187	100	1

Su gráfica u ojiva es como se presenta a continuación:

1.4 Una clasificación de las distribuciones según la forma de los datos.

Las distribuciones de frecuencia se suelen clasificar de acuerdo a la forma en que se concentren los valores de las frecuencias relativas. Para ello los histogramas son de gran utilidad.

La clasificación comprende, básicamente, dos tipos:

1. Distribuciones simétricas.
2. Distribuciones asimétricas.

Entre las distribuciones asimétricas, se encuentran:

1. Distribuciones sesgadas hacia la derecha.
2. Distribuciones sesgadas hacia la izquierda.
3. Crecientes.
4. Decrecientes.
5. Bimodales.
6. Multimodales. (Schaum, p. 40)

1.5 Comentarios finales

Hemos estudiado la definición de las estadísticas y los datos, así como su clasificación. Presentamos luego algunos métodos tabulares y gráficos utilizados para estudiar el comportamiento de los datos. Los resúmenes gráficos o tabulares de datos son de uso común, tanto en prensa como en investigaciones especializadas. Es muy ventajoso saber cómo se realizan y se interpretan.

Ejercicio

La siguiente tabla muestra una distribución de frecuencias de los salarios semanales de 65 empleados de la empresa S&C. De acuerdo a dichos datos, determine:

a) El límite inferior de la sexta clase.

b) El límite superior de la cuarta clase.

c) El punto medio de la tercera clase.

d) El tamaño del quinto intervalo de clase.

El intervalo de clase con mayor frecuencia se llama intervalo de clase modal. A su frecuencia se le denomina frecuencia de clase modal.

.Determínelas para este caso.

Construya la tabla de frecuencias relativas y grafique su histograma. A partir de todo el estudio anterior, decida sobre:

a) El porcentaje de empleados que ganan menos de $280.00 a la semana.

b) El porcentaje de empleados que reciben más de $260.00 pero menos de $300.00, a la semana.

¿Qué forma tiene el histograma de las frecuencias relativas?

CAPÍTULO II

La descripción a través de medidas numéricas

MAWIL
Publicaciones Impresas
y Digitales

www.mawil.us

2. La descripción a través de medidas numéricas

2.1 ¿Qué es la Estadística Descriptiva?

Las estadísticas tienen distintas formas de abordar analíticamente a los datos. Hemos estudiado la importancia del ordenamiento, la tabulación y la graficación para su resumen, pero existen herramientas numéricas para describir, clasificar y sintetizar a un conjunto de datos. Al uso de estas herramientas se le denomina **Estadística Descriptiva**. Esta rama no pretende ir más allá de los datos, no le interesa la población de donde provienen las muestras y observaciones.

Supongamos que se dispone de un conjunto de datos muestrales, ya sea en bruto o sintetizados en una tabla de frecuencias. Antes que por su síntesis gráfica, no interesan ciertas medidas más condensadas referidas a los datos. Tales medidas merecen una definición, a saber, la de **estadístico**, que es cualquier magnitud numérica cuyo valor se pueda determinar a partir de los datos. Estos instrumentos pueden dan características numéricas a los datos, sea sobre la **localización** de los datos, conocidas como medidas de tendencia central, sea de **dispersión** o qué cantidad de variación hay entre los datos, sea de forma o sea de **asociación**. Si a tales medidas se les obtiene a partir de muestras se les denominan estadísticos muestrales.

2.2 Medidas de tendencia central o estadísticos de localización

Existen tres estadísticos de localización, la **media**, la **mediana** y la **moda**. **La media:** también conocida como promedio o media aritmética, es la medida de localización más importante de una variable. La media proporciona un número que indica la ubicación central de los datos. Si los datos provienen de una muestra, a la media se le denota por $\overline{x}$, mientras que si los datos son los de una población entera, su media se denota por μ, la letra griega mu.

Si los datos están constituidos por n observaciones, presentadas en una tabla ordenada como la siguiente:

i	1	2	3	4	5	…	n-2	n-1	n
X_i	X_1	X_2	X_3	X_4	X_5	…	X_{n-2}	X_{n-1}	X_n

Donde las X subindexadas son las observaciones que componen a los datos, la fórmula para la media muestral es la siguiente

$$\overline{x}=(X1+X2+X3+X4+ \quad \ldots+Xn-1+Xn)/n=\sum x_i / n$$

Es decir, la suma de todas las observaciones dividas entre la cantidad total de tales.

La mediana es otra medida de localización central. Es el valor a la mitad de los datos ordenados de menor a mayor, de manera ascendente. Cuando se tiene un número impar de observaciones, la mediana es el valor de en medio. Pero cuando la cantidad de observaciones es par, no hay tal valor en medio y en ese caso, se sigue una convención y la mediana es definida como el promedio de las dos observaciones de la mitad. Por convención, la definición de mediana se plantea de la siguiente forma:

Una vez ordenados los datos de menor a mayor y en forma ascendente se tiene que:

- Si el número de observaciones es impar, la mediana es el valor de la mitad.
- Si el número de observaciones es par, la mediana es el promedio de las dos observaciones de la mitad.

La mediana es la medida de localización preferida cuando se trata de ingresos anuales y valores de rentas, debido a que la media puede inflarse por unos cuantos ingresos o valores muy altos pero atípicos. En tales casos, la mediana es el estadístico de tendencia central *más utilizado*. Se puede hallar solo para valor cuantitativos.

La Moda es el estadístico de valor central que indica el dato o el intervalo de datos con mayor frecuencia. Dicho estadístico no siempre es único, pues hay casos en donde las distribuciones tienen múltiples modas. Cuando los datos son agrupados por clase, también se habla de clase o intervalo modal.

Las definiciones anteriores apenas si hacen referencia a los cálculos. Por ejemplo, la media se calcula a partir de los datos sueltos. Así mismo, tanto la mediana como la moda se definieron y determinaron de acuerdo a la inspección del histograma de los datos. A continuación estudiaremos cómo calcular dichos valores a partir de su tabulación por frecuencias agrupados en clases.

Para calcular la media se necesita conocer la marca de clase o punto medio de las clases y la frecuencia absoluta. Se procede sumando todos los productos del punto medio con la frecuencia absoluta que le corresponde a la clase, para que se divida al final por el número total de datos. Dicha operación se sintetiza en la siguiente fórmula:

Media=Suma de (i-ésima marca de clase * su frecuencia respectiva) / Total de datos

Para el cálculo de la mediana, se tiene que considerar que ella se encuentra en el intervalo donde la frecuencia acumulada llega hasta la mitad

de la suma de las frecuencias absolutas. Así, la siguiente fórmula da como resultado a la mediana:

$$\textbf{Mediana}=L_i+t_i*(N/2-F_{i-1})/f_i$$

Donde:

L_i es el límite inferior de la clase donde se encuentra la mediana.

N / 2 es la semisuma de las frecuencias absolutas.

F_i-1 es la frecuencia acumulada anterior a la clase mediana.

f_i es la frecuencia absoluta del intervalo mediano.

t_i es la amplitud de los intervalos.

Ahora, para la moda o el intervalo modal, se calcula con la fórmula:

$$\textbf{Moda}=L_i+t_i*(f_i-f_{i-1})/((f_i-f_{i-1})+(f_{i+1}-f_i))$$

Donde:

- **L_i** es el extremo inferior del intervalo modal (intervalo que tiene mayor frecuencia absoluta).
- **f_i** Frecuencia absoluta del intervalo modal.
- **f_i-1** Frecuencia absoluta del intervalo anterior al modal.

- **f_i+1** Frecuencia absoluta del intervalo posterior al modal.
- **t_i** Amplitud de los intervalos.

Ahora veamos un ejemplo con la tabla de estaturas

Estatura (en centimetros)	Marcas de clase	Frecuencia absoluta	Frecuencia acumulada
152-157	154.5	5	5
158-165	161.5	18	23
166-172	**169**	**42**	**65**
173-180	176.5	27	92
181-187	184	8	100

El cálculo de la media, indica que

$$\overline{x} = (154.5*5 + 161.5*18+169*42 + 176.5*27 + 184*8)/100$$

$$\overline{x} = 17015/100$$

$$\overline{x} = 170.15$$

Ahora, calculamos la mediana (Me) según las fórmulas anteriores: primero identificamos la **clase mediana**, para esto tenemos que buscar el intervalo en el que se encuentre la primera frecuencia acumulada mayor o igual a la cantidad total de observaciones entre dos. En este caso N / 2 = 100 / 2 = 50. Así, la clase mediana es aquella que comprende las estatura entre 166 y 172 centímetros, destacada en la tabla.

Identificamos los valores de la fórmula:

- **L_i** es el límite inferior de la clase mediana, y es 166.
- **N / 2** es la mitad de la cantidad de observaciones, y vale 50.
- **F_i-1** es la frecuencia acumulada anterior a la clase mediana, la cual es 23.
- **f_i** es la frecuencia absoluta de la clase mediana, que es 42.
- **t_i** es la amplitud de la clase mediana, que es 172-166=6.

Luego, la mediana resulta ser:

$$Me=166 + 6*(50-23)/42$$

$$Me=166+6*27/42$$

$$Me=166+3.86$$

$$\mathbf{Me=169.86}$$

Para la moda (Mo) lo primero que debemos hacer es identificar el intervalo modal, que justamente coincide con la clase mediana antes estudiada, aquella que agrupa los individuos con estatura entre 166 y 172 centímetros, indicada en la tabla. A continuación identificamos las cantidades involucradas en la fórmula:

- **L_i** vale 166.
- **f_i** vale 42.
- **f_i-1,** la frecuencia absoluta del intervalo anterior al modal, vale 18.
- **f_i+1,** la frecuencia absoluta del intervalo posterior al modal, vale 27.
- **t_i**, es la amplitud de la clase modal, a saber 6.

Luego, la moda es:

$$Mo=166+6*(42-18)/((42-18) + (42-27))$$

$$Mo=166+6*24/(24+15)$$

$$Mo=166+6*24/39$$

$$Mo=166+3.69$$

$$\mathbf{Mo=169.69}$$

2.3 Relación empírica entre los tres estadísticos anteriores

Cuando los datos presentan una simetría normal, los valores de la media, la moda y la mediana coinciden. En caso contrario, se tiene que la mediana está entre la moda y la media. La ubicación de la moda, el pico

de los datos frecuenciales, es la que dictamina sobre el sesgo, u orientación hacia la izquierda o hacia la derecha, de la distribución de frecuencias de los datos.

2.4 Cuantiles

Si un conjunto de datos se ordena de acuerdo a su magnitud, la mediana divide al conjunto en dos partes iguales. Si nos preguntamos cuáles serían los estadísticos que los dividen en más partes, aparecen los estadísticos que se denominan **cuantiles**, los cuales aportan información sobre la ubicación de los datos en el intervalo que va del menor al mayor valor de los datos. Existen **los cuartiles**, si se dividen a los datos en 4 partes, los **deciles**, al dividir entre 10 y los **percentiles**, que resultan de dividir a los datos en 100 partes iguales. Considerando a p como un número entre 0 y 100, en los datos que no tienen muchos valores repetidos, el **percentil-p** divide a los datos en dos partes, no simétricas. Cerca del p% de las observaciones tienen valores menores que el percentil-p y aproximadamente (100-p)% de las observaciones tienen valores mayores que el percentil p.

Es usual trabajar con los percentiles, los que serán definidos a continuación. El **percentil-p** es un valor tal que por lo menos p por ciento de las observaciones son menores o iguales que este valor y por lo menos (100

- p) por ciento de las observaciones son mayores o iguales que este valor.

Así, por ejemplo, las puntuaciones de las evaluaciones en los procesos educativos suelen darse en términos de percentiles. Por ejemplo, suponga que un estudiante obtiene 52 puntos en la evaluación teórica. Por sí mismo, tal información no dice mucho acerca de este estudiante en relación con los demás estudiantes que fueron evaluados. Sin embargo, si esta puntuación corresponde al percentil 70, entonces 70% de los estudiantes obtuvieron una puntuación menor a la de dicho estudiante y 30% de los estudiantes obtuvieron una puntuación mayor, pudiendo tener una mejor apreciación del desempeño del estudiante en cuestión, en relación al universo completo de evaluados. Un algoritmo sencillo para calcular el percentil-p es el siguiente.

Cálculo del Percentil-p

- Paso 0. Se fija el número p entre 0 y 100, que será el percentil deseado.
- Paso 1. Ordenar los n datos de menor a mayor.
- Paso 2. Calcular el índice i de acuerdo a la fórmula i=n*(p/100)
- Paso 3. (a) Si el índice no es un número entero, debe redondearlo

Así, el primer entero mayor que i denota la posición del percentil-p.

Paso 3. (b) Si i es un numero entero, el percentil-p es el promedio de los datos en las posiciones i e i+1.

Los percentiles también se pueden obtener por simple inspección de una ojiva de frecuencias absolutas.

Veamos el siguiente ejemplo. Estudiemos la tabla 1.2, donde ya están ordenados los datos, dada la frecuencia. Su moda es, evidentemente, 0, pues es la cantidad de días de ausentismo de mayor frecuencia. Su media es 3.

Días de ausentismo laboral	**Frecuencia**
0	12
1	8
2	5
3	4
4	5
5	8
6	0
7	5
8	2
9	1

Queremos calcular su percentil-50, es decir, p=50. La cantidad de datos es n=50. Luego, el índice es i=50*(50/100)=25. El dato que está en la posición 25 tiene dos días de ausentismo pues así lo indica la suma de las tres primeras filas de frecuencia,12+8+5=25. Luego el, el percentil-50

es (2+3)/2=2.5, la cual es su mediana.

Es de notarse que los percentiles tiene una relación con los cuartiles. El percentil-25 es el primer cuartil **Q1**, el percentil-50 es el segundo cuartil **Q2** y el percentil-75 es el tercer cuartil **Q3**.

2.5 Medidas de dispersión o estadísticos de variación

La **dispersión** o **variación** de los datos es el nivel en el que se esparcen o distribuyen los datos alrededor de la media, cuán alejados o cercanos pueden estar de la medida de comportamiento central.

El estadístico más sencillo para medir dispersión es **el rango**, definido como la diferencia entre el número mayor y el número menor del conjunto de datos, como ya se definió antes:

R=Max. - Min.

A veces se suele representar como un intervalo. Como por ejemplo, del conjunto de datos {2, 3, 3, 4, 5, 5, 5, 5, 7, 8, 10, 10, 11}, su rango es 11-2=9 o también [2,11].

Aunque el rango sea tan sencillo de calcular para medir variabilidad, muy pocas veces se utiliza como única medida. Sobre todo porque el rango se basa sólo en dos observaciones que son, además, los valores

extremos que podrían ser los mismos para distribuciones de frecuencias totalmente distintas. De allí que se utilice otro estadístico de dispersión denominada rango intercuartilar, que es la diferencia entre el tercer cuartil Q3 y el primer cuartil Q1.

$$\mathbf{RQ=Q3-Q1}$$

En dicho rango se encuentra el 50% de los datos interiores de un conjunto dado.

Con el mismo espíritu se define el **rango percentilar 10-90**, como la diferencia entre el percentil-90 y el percentil-10.

El estadístico de dispersión más importante es **la varianza** y, dado un conjunto de n observaciones {x1, x2, x3, …, xn} se define como:

$$\mathbf{s^2=((x1-\overline{x})^2+(x2-\overline{x})^2 + \ldots + (xn-\overline{x})^2)/n=\sum (xi-\overline{x})^2/ n}$$

Dicho estadístico, mide en promedio cuadrático, cuánto se dispersan los datos alrededor del valor medio $\overline{x}$ de los datos. Veamos, con un ejemplo, cómo se calcula tal estadístico para datos agrupados, ampliando la tabla de frecuencias de las estaturas de ejemplos anteriores. Sabemos de antemano que $\overline{x}$ =170.15. Tomaremos a los xi como las marcas de clase. Como los datos están ya agrupados, debemos multiplicar a cada elemento de la sexta clase por su frecuencia de clase respectiva para obtener cada

sumando de la fórmula de s^2:

Clase	Estatura	Frecuencia (Fi)	xi	$xi-\overline{x}$	$(xi-\overline{x})^2$	$Fi*(xi-\overline{x})^2$
1	152-157	5	154.5	-15,65	244,9225	1224,6125
2	158-165	18	161.5	-8,65	74,8225	1346,805
3	166-172	42	169	-1,15	1,3225	55,545
4	173-180	27	176.5	6,35	40,3225	1088,7075
5	181-187	8	184	13,85	191,8225	1534,58

Al sumar la última columna, se obtiene 5225,25, que al dividir entre la suma de toda la columna de frecuencia, es decir, el total de observaciones n=100, nos arroja el valor de la varianza que es 52,5025

Propiedades de la varianza

1. La varianza será siempre un valor positivo o cero.
2. La varianza será cero cuando los datos sean siempre los mismos, es decir, constantes.
3. Si a todos los valores de los se les suma un número la varianza no se modifica.
4. Si todos los valores de los datos se multiplican por un número la varianza queda multiplicada por el cuadrado de dicho número.
5. Se puede calcular con una fórmula más resumida:

$$s^2=((x1)^2+(x2)^2+\ldots+(xn)^2)/n - \overline{x}^2 = \sum (xi)^2/ n - x^2$$

Con el objeto de interpretar más fácil al dato que arroja la varianza como estadístico, se utiliza **la desviación estándar**, que se mide en las mismas unidades que los datos. De modo que se define la desviación estándar como la raíz cuadrada positiva de la varianza; se denota y defina por:

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

En el ejemplo anterior, la desviación estándar se obtiene de la varianza tomando su raíz cuadrada, es decir $s=\sqrt{52,5025}=7,25$. Lo cual indica que los valores de las estaturas de la tabla 1.3 se desvían alrededor de la media 170.15 cm con más o menos 7.25 cm

Existe una **regla emp**írica para los datos simétricos, que dota de un sentido tangible a la desviación estándar. Cuando nos enfrentamos a datos simétricos, se conoce como regla empírica que:

- Aproximadamente el 68% de los datos cae en el intervalo [$\overline{x} - s$, $\overline{x} + s$]
- Aproximadamente el 95% de los datos cae en el intervalo [$\overline{x} - 2s$, $\overline{x} + 2s$]
- Aproximadamente el 99% de los datos cae en el intervalo [$\overline{x} - 3s$,

$\overline{x}$ +3s]

En algunas ocasiones se utiliza un estadístico descriptivo que indica la magnitud de la desviación estándar en relación con la media. A tal medida se le conoce **como coeficiente de variación** y se representa porcentualmente. Se define como:

$$(s / \overline{x}) * 100\ \%$$

Así, en el ejemplo de las estaturas, la media es de 170.15, la desviación estándar de 7,25 y el coeficiente de variación sería (7,25/170,25)*100=4,26%

2.6 Miscelánea de estadísticos

Existen muchos estadísticos de interés. A continuación haremos comentarios sobre algunos de ellos, con el objeto de introducir referencias a herramientas de interés.

Sobre el sesgo

Anteriormente se hizo un breve comentario sobre la clasificación de las distribuciones de acuerdo a su forma. Así, junto a la graficación de las tablas de frecuencia, existe un estadístico que indica hacia donde se concentran más los datos. Este estadístico se denomina sesgo y se define con la siguiente fórmula:

$$\textbf{sesgo}=n/((n-1)*(n-2))\sum((xi-\bar{x})/s)^3$$

Cuando el valor de dicho estadístico es 0, nos encontramos con una distribución simétrica. Cuando es positivo se dice que los datos siguen una distribución asimétrica con un sesgo a la derecha, siendo el caso que la moda es menor que la media. Mientras que cuando el sesgo es negativo, se dice que la asimetría de los datos tiene un sesgo a la izquierda y ocurre que la moda es mayor que la media.

Sobre la correlación

Hasta ahora se han estudiados herramientas que resumen datos en una sola variable. Pero es frecuente la necesidad de examinar y tomar decisiones a partir de la relación entre dos variables. En ese caso, los datos se presentan como parejas de números aleatorios. Cuando se tiene un conjunto de datos apareados (xi,yi), interesa saber cuantitativamente la relación que tienen entre sí las variables asociadas a la observación. Por ejemplo, si se tiene una tabla como la siguiente:

Estatura X del padre	1.65	1.63	1.67	1.64	1.68	1.62	1.70	1.66	1.68
Estatura Y del hijo	1.68	1.66	1.68	1.65	1.69	1.66	1.68	1.65	1.71

Para medir la relación o asociación entre las estaturas de padre e hijo, se utiliza el estadístico que es denominada coeficiente de correlación

y se define y denota por:

$$\mathbf{r=\sum(xi - \overline{x})(yi - \overline{y})/((n-1)s_x\ s_y)}$$

Donde $\overline{x}$ y s_x son, respectivamente, la media y la desviación estándar de los xi. Del mismo modo, se definen a los estadístico $\overline{y}$, s_y con respecto a los valores yi.

Así, cuando r tiende a ser un valor próximo a 1, se dice que cuando los xi crecen, los yi también lo hacen. Mientras que cuando r tiende a ser -1 significa que los yi tienden a decrecer cuando los xi crecen, a la par que crecerían si los xi decrecen. Cuando r es igual a 0, entonces se dice que no hay correlación lineal alguna.

2.7 Conclusiones

Hasta ahora se han presentado varias herramientas cuantitativas que se denominaron estadísticos cuya función es describir un conjunto dado de datos, es decir, que sirven para resumir la tendencia central, la agrupación, la dispersión y forma de la distribución de frecuencias, así como una regla empírica de gran importancia para ubicar porcentualmente a los datos alrededor de la tendencia media. En contraste con los procedimientos gráficos y tabulares presentados antes, los estadísticos estudiados resumen

los datos con valores numéricos, y esto reporta y significa una ventaja en la precisión de la caracterización de la información que porta un conjunto de datos específico.

Ejercicio

La siguiente tabla registra la distribución de frecuencias del número semanal de minutos que pasan viendo la televisión 400 estudiantes universitarios.

Tiempo viendo TV (en minutos)	Números de estudiantes
300-399	14
400-499	46
500-599	58
600-699	76
700-799	68
800-899	62
900-999	48
1000-1099	22
1100-1199	6

- Encuentre su media, su moda y su mediana. Decida sobre la simetría de los datos.
- Determine su desviación estándar. ¿Puede utilizar la regla empírica?.
- Diga qué porcentaje de estudiantes mira televisión por más de 900 minutos. ¿Y más de 500 minutos pero menos de 1000 minutos?.

CAPÍTULO III

Introducción a la teoría de las probabilidades

MAWIL
Publicaciones Impresas
y Digitales

www.mawil.us

3. Introducción a la teoría de las probabilidades

3.1 Introducción y definiciones básicas. Los eventos

Con el objeto de explorar posibilidades de la estadística más enriquecedoras, más profundas y versátiles, es decir, procurar que la estadística trascienda a la simple descripción de un conjunto de datos, debemos conocer determinadas herramientas matemáticas que procuren discernir sobre la posibilidad de que ocurran ciertos sucesos bajo determinadas circunstancias. Así, encontrar la verosimilitud de que ocurra tal o cual evento, es el objetivo de la teoría de probabilidades (Ross, 2007, pág. 144).

La **teoría de probabilidades** se define como la rama de las matemáticas encargada de medir la posibilidad de un determinado suceso cuando se realiza un experimento.

Un **experimento** debe entenderse como cualquier proceso que produzca una observación o resultado. Al conjunto de todos los posibles resultados se le llama **espacio muestral,** denotado por S. Un **suceso** o **evento** es un conjunto de posibles resultados que se encuentren en el espacio muestral.

Como ejemplo, supóngase el experimento de lanzar un dado de 6 caras bien equilibrado. La observación que define al espacio muestral es registrar el número asociada a la cara que cae hacia arriba. Así, el espacio

muestral sería.

$$S=\{1, 2, 3, 4, 5, 6\}$$

Como posibles eventos podríamos interesarnos en los eventos:

- Que salga 1, $A=\{1\}$
- Que salga un número par, $P=\{2, 4, 6\}$
- Que salga 2 ó 5, $Z=\{2,5\}$

La asociación de un evento a ciertos subconjuntos del espacio muestral S es considerablemente útil y fundamental para distinguir entre dos tipos de sucesos:

- **Eventos elementales**: son aquellos que están conformados por un solo resultado.
- **Eventos compuestos**: son aquellos conformados por más de un resultado.

De tal forma que las operaciones de conjunto harto conocidas, es la manera en cómo se definen eventos compuestos, de acuerdo al enunciado de un resultado de interés. Así, la siguiente tabla indica cómo se construyen tales eventos, qué operaciones de conjuntos las defines y qué significan estas en términos de enunciados sobre observaciones experimentales:

Enunciado	Operación de conjuntos y/o conjunto célebre asociados	Evento
Por lo menos uno de los eventos A o B ocurre	Unión	A U B = A+B
Ambos eventos A y B ocurren	Intersección	$A \cap B = AB$
No ocurre el evento A	Complemento	$A^c = A'$
No ocurre ni A ni B	---	$A^c \cap B^c$
Ocurre a la sumo un evento entre A y B, es decir, no ocurre más de uno de los eventos A y B	---	$(A \cap B)^c$
Ocurre exactamente uno de los sucesos A o B	Diferencia simétrica	$A \Delta B=(A \cap B^c) U(A^c \cap B)$
Si A ocurre entonces también B ocurre	Contención o subconjunto	$A < B$
Evento imposible	Conjunto vacío	Ø
Evento seguro	Espacio muestral	S

Tab.3.1

Fuente: (Apostol, 2001, pág. 580)

Así pues, los eventos elementales combinados unos con otros de acuerdo a toda la gama de operaciones de conjuntos, producen y condicionan eventos compuestos de la más variada índole, recordando que se pueden operar entre ellos más de una vez, de la misma forma que evidenciar relaciones entre sí. Haciendo referencia al ejemplo anterior, los eventos

P="que salga un número par" y I="que salga un número impar", no tienen nada en común, son mutuamente excluyentes o sea disjuntos (no pueden suceder ambos al mismo tiempo), pues

$$P \cap I = \{2, 4, 6\} \cap \{1, 3, 5\} = Ø$$

El evento elemental C="sale 4" implica al evento compuesto P="que salga un número par", en el sentido en que si C sucede también lo hará A, pues

$$\{4\} < \{2, 4, 6\}$$

3.2 Más definiciones básicas. La probabilidad

Asociar una medida de verosimilitud a un evento se define clásicamente como una forma de conteo relativizada. Así, si lanzamos un dado y nos preguntamos sobre la probabilidad de que salga un número par, pues decidimos sobre la cantidad de formas en que puede salir un número impar (casos favorables o evento I) y lo dividimos entre el total de formas posible del resultados (casos posibles o espacio muestral), es decir, contamos la cantidad de elementos que hay en I y lo dividimos por la cantidad de elementos que hay en S. Sintéticamente, se dice que la probabilidad del evento I es:

$$P(I) = \#I / \#S = 3 / 6 = 1 / 2 = 0.5$$

Esta definición clásica presupone que el dado es justo, es decir, que todos los eventos elementales (cada una de las caras) tiene igual probabilidad de salir. Es decir, es un experimento uniforme. En principio, no tiene nada que ver con el resultado del experimento, pero es una aproximación bastante plausible al cálculo de la **frecuencia relativa** tras, digamos, lanzar el dado 100 veces y contar cuántas de esas 100 veces sale un número impar. Es evidente que, por la aleatoriedad, no saldrán exactamente 50 números impares, pero sí un número bastante cercano.

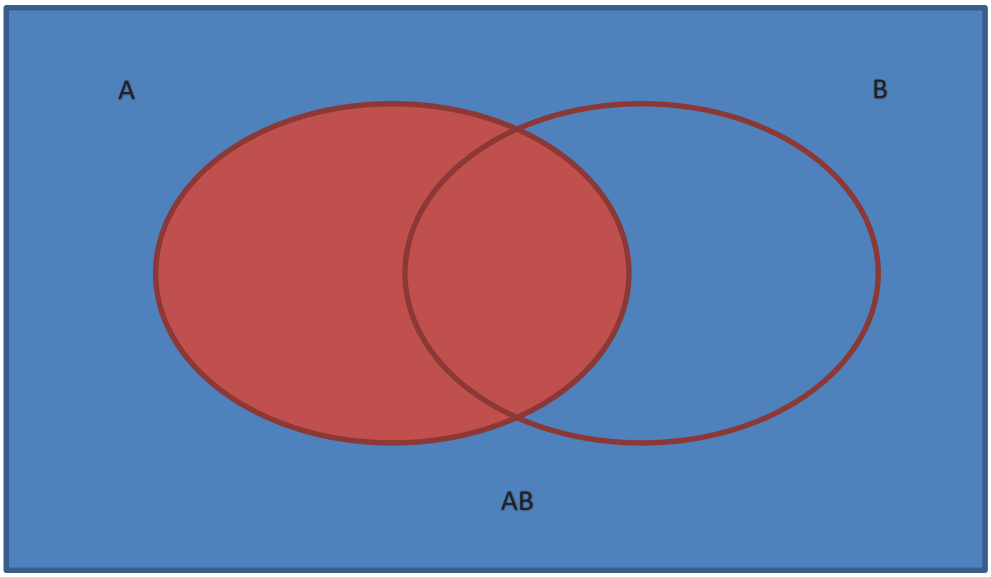
Sin embargo, esta definición clásica de la probabilidad, inspira a un conjunto de propiedades que deseamos que la medida de verosimilitud de los eventos asociados a un experimento tenga, a saber:

1. $P(S)=1$
2. $P(A^c)=1-P(A)$
3. $P(A \cup B)= P(A) + P(B)$ si $A \cap B= \emptyset$
4. $P(A)<P(B)$ si $A<B$

Al conjunto de tales **propiedades de la probabilidad,** se le puede extender al considerar que $S^c= \emptyset$, y por la propiedad 2 se tiene que:

$$P(Ø)=1- P(S)=1 – 1= 0$$

Así mismo, la propiedad 3 se puede ampliar, tras considerar el principio elemental de conteo de la unión de conjuntos conocido por principio de inclusión – exclusión, ilustrado en la tabla 3.2 o ley del doble conteo que establece que:



$$\#(A \cup B) = \#A + \#B - \#(A \cap B)$$

A B

Tab. 3.2

Se concluye que, sin importar qué relación hay entre A y B, se tiene lo que se conoce como ley de adición

$$\mathbf{P(A \cup B)= P(A) + P(B) – P(A \cap B)}$$

Podemos decir que **la probabilidad** es una función de conjuntos que gradúa la posibilidad de ocurrencia de un evento con los números

reales,

1. **De manera monótona**, i.e., eventos más grandes tienen más posibilidad de ocurrir.

2. **Aditiva**, pues la probabilidad de la unión de eventos disjuntos es la suma de las probabilidades de cada evento operado

3. **Definida no negativa**, pues todas las probabilidades son mayores que cero, y al evento imposible se le asigna el número 0

4. **Unitariamente**, pues el evento de ocurrencia segura se le asigna el número 1.

5. **Complementaria**, pues el complemento de un evento tiene probabilidad recíproca, como resta con respecto al uno.

6. Y que cumple con el principio de inclusión-exclusión del conteo básico o **Ley de adición**.

Como ejemplo, los clientes de una consultora suelen poder cancelar los costos de servicio de dos formas, de contado o a crédito. Un 22% de los clientes solo puede pagar de contado, mientras que un 58% de ellos nada más puede hacerlo a crédito. Además, se sabe que un 14% puede pagar de ambas formas. ¿Cuál es la probabilidad de que un cliente pueda pagar con al menos de una de las dos formas?.

La pregunta sobre “al menos” significa, de acuerdo a la tabla 3.1, que nos interesa la probabilidad de un evento definido como la unión de los que pueden pagar de contado con los que pueden pagar a crédito. Definidos los eventos:

A={Los clientes que pueden pagar de contado}

B={ Los clientes que pueden pagar a crédito}

A∩B={Los clientes que pueden pagar de ambas formas}

AUB={ Los clientes que pueden pagar de contado o a créditos, o ambas inclusive}

Se tiene que, por los datos del enunciado del ejemplo,

P(A)=0.22

P(B)=0.58

P(A∩B)=0.14

Ya que si los datos se presentan porcentualmente, las probabilidades se deben dividir entre 100. Así, como nos interesa la P(AUB), utilizamos la Ley de adición:

P(AUB)= P(A) + P(B) – P(A∩B)

P(AUB)= 0.22 + 0.58 - 0.14 = 0.66

Es decir, volviendo a la expresión porcentual, el 66% de los clientes de la consultora, puede pagar o a crédito o de contado.

¿Y si es necesario saber cuántos no pueden pagar de contado? Pues en ese caso el evento de interés es A^c={los clientes que **no** pueden pagar de contado}, así $P(A^c)=1-P(A)=1-0.22=0.78$, es decir un 78% de los clientes no puede pagar de contado.

3.3 Probabilidad condicional, ley de la multiplicación e independencia de eventos.

A menudo, es necesario calcular probabilidades cuando se tiene información parcial relativa al resultado de un experimento. En tales casos, nos enfrentamos a un condicionamiento entre eventos y aparece un nuevo concepto que se denomina **probabilidad condicional**.

Si tenemos dos eventos A y B, la probabilidad de A dado que B haya ocurrido se define y denota por:

$$\mathbf{P(A|B)= P(A\cap B)/P(B)}$$

Siempre que P(B) pueda ocurrir, es decir, $P(B)\neq 0$.

Para ilustrar en qué consiste este nuevo concepto, veamos el ejemplo del dado nuevamente. Supongamos que se lanza un dado equilibrado,

si ya se obtuvo un número impar, ¿cuál es la probabilidad de que ese número sea el 1? Sean pues los eventos:

A={Salió un 1}={1}

B={Salió un número impar}={1, 3, 5}

El sentido común nos indica que tal probabilidad debe ser 1/3 debido a que solo hay tres posibilidades de números impares y una sola para que salga un 1. Pero queremos calcular dicha probabilidad condicional utilizando la fórmula que la define.

Tenemos que $P(A \cap B)=P(A)$, pues al intersecar B con A obtenemos el conjunto más pequeño contenido en ambos, a saber {1}. Calculando con la definición clásica de probabilidades, se tiene que:

$P(A)=1/6$

$P(B)=3/6=1/2$

Así,

$P(A|B)= P(A \cap B)/P(B)= P(A)/P(B)=(1/6)/(1/2)=2/6=1/3$

Como se vio por simple inspección, lo cual corrobora la validez de la definición de la probabilidad condicional.

Es necesario ahora, enunciar una nueva identidad para el cálculo de

las probabilidades, conocidas como **Ley de multiplicación**:

$$P(A \cap B)=P(B)*P(A|B)$$

$$P(A \cap B)=P(A)*P(B|A)$$

Ambas, se derivan de la probabilidad condicional, tras un simple despeje y se utiliza una u otra de acuerdo a los datos de los que se dispongan.

Finalmente, cuando ocurre que la P(A|B)=P(A), es decir, que B no condiciona al evento A, estamos ante un fenómenos que se denomina **independencia**. En este caso, tras sustituir esa igualdad en la definición de probabilidad condicional, obtenemos la condición para definir que dos eventos **A y B son independientes si**

$$P(A \cap B)=P(A)*P(B)$$

Tal igualdad contiene el hecho de que la probabilidad de que ocurra uno de ellos no se ve afectada por la información de que el otro haya ocurrido o no.

Estudiemos ahora un ejemplo más. Supongamos que una caja contiene 3 pelotas blancas y 2 pelotas negras. Se sacan dos pelotas aleatoriamente y **sin reposición**, es decir, que una vez sacada no se vuelve a colocar. Sea el evento E={“la primera pelota es negra”} y F={“la segunda pelota es

negra"}. ¿Cuál es la probabilidad de que ambas pelotas sean negras?

Debemos, pues, calcular P(E∩F). Utilizaremos la ley de multiplicación, con el evento E como condicionante. P(E∩F)=P(E)*P(E|F). Así, P(E)=2/5 mientras que P(E|F)=1/4. Así, P(E∩F)=(2/5)*(1/4)=2/20=1/10.

En este experimento es más o menos evidente que los eventos E y F no son independientes. El solo hecho de que salga o no una pelota negra en la primera extracción condiciona la probabilidad de que salga una pelota negra en la segunda extracción. Sin embargo, si modificamos el enunciado y establecemos que al sacar la primera pelota se registra su color y se vuelve a colocar en la caja, si la extracción se hace **con reposición**, el escenario cambia y P(E|F)=P(E).

Ejercicio. Se lanza un dado dos veces. Calcular la probabilidad de obtener 4, 5 ó 6 en el primer lanzamiento y 1, 2 ó 3 en el segundo. ¿Se trata de eventos independientes?

3.4 Leyes de la probabilidad y el método de la composición de eventos.

En general, hay un algoritmo bastante sencillo para calcular las probabilidades de eventos con un método de composición progresiva de eventos a partir de eventos elementales.

Considerando que las leyes de la probabilidad son:

- Ley de suma: $P(A \cup B)= P(A) + P(B) – P(A \cap B)$
- Ley de complemento: $P(A^c)=1-P(A)$
- Ley de producto: $P(A \cap B)=P(B)*P(A|B)$ y $P(A \cap B)=P(A)*P(B|A)$
- Independencia de eventos: $P(A \cap B)=P(A)*P(B)$

Método de la composición de eventos

Paso 0. Defina el experimento y el evento de interés.

Paso 1. Visualice la naturaleza de los puntos muestrales o eventos elementales. Identifique así algunos de ellos para aclarar las ideas.

Paso 2. Formule una ecuación que expresión el evento de interés, digamos A, como una composición de dos o más eventos mediante las operaciones de conjuntos vistas (unión, intersección y/o complementación. Verifique que el evento A es equivalente a la ecuación que la representa, es decir, que son conjuntos iguales, con los mismos eventos elementales.

Paso 3. Aplique las leyes para el cálculo de las probabilidades antes enunciadas a la ecuación que se determinó en el paso anterior para calcular así P(A). (Mendenhall III, Wackerly, & Scheaffer, 2002, pág. 61)

Ejercicio. En una ciudad democrática, el 40% de los electores son

de derecha y el 60% son de izquierda. Entre los derechistas, 70% está a favor de la renta básica universal, mientras que entre los izquierdistas el 80% la aprueba. Si se elige un elector al azar en la ciudad, ¿cuál es la probabilidad de que esté a favor de la renta básica universal?

3.5 El teorema de Bayes

Hasta ahora, hemos evidenciado que las probabilidades son la fuente para obtener más información sobre la ocurrencia de los eventos. Así, es general que los análisis se inicien con una estimación de probabilidad a priori o probabilidad previa de los eventos de interés. Luego, de fuentes que provengan de muestras, informaciones especiales o pruebas a posteriori, se obtiene más información sobre tales eventos. De modo que con esta nueva información, se revisan y modifican los valores de las probabilidades mediante cálculos hechos posteriormente. El **teorema de Bayes** es una fuente de teoría para calcular dichas probabilidades.

Para abordar dicho teorema debemos explorar un principio previo que se conoce como **Principio de expansión** o **Ley de probabilidad total**. Si miramos con detenimiento la tabla 3.2 hemos de notar que el conjunto A se descompone en dos partes disjuntas, aquella que tiene que común con B, es decir $A \cap B$; y aquella que no tiene nada que ver con B, esto es $A \cap B^c$.

Luego, al reunir ambas partes, es cierto que

$A = (A \cap B) \cup (A \cap B^c)$

De forma tal que si deseamos calcular la probabilidad del evento A, descompuesto así, en dos partes evidentemente disjuntas, tendremos que

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Luego, tras revisar la Ley de multiplicación y aplicarla convenientemente para cada sumando de la expresión inmediatamente anterior, i.e., $P(A \cap B)=P(A|B)P(B)$ y $P(A \cap B^c) = P(A|B^c)\, P(B^c)$, tendremos que:

$$\mathbf{P(A) = P(A|B)P(B) \; + \; P(A|B^c)P(B^c)}$$

La cual se conoce como **Ley de probabilidad total** de un evento descompuesto en 2 partes disjuntas.

Con ella podemos calcular, volviendo al ejemplo de la caja con las 5 pelotas, que la extracción de las pelotas sin reposición, afecta la independencia de los resultados. Veamos pues que con el supuesto de que la caja contiene 3 pelotas blancas y 2 pelotas negras y los eventos E = {"la primera pelota es negra"} y F = {"la segunda pelota es negra"}, calcularemos P(F), descomponiendo el evento según E, ya definido, y E^c = {"la primera pelota es blanca"}. Sabemos que:

$P(F|E)=1/4$

$P(E)=2/5$

$P(F|E^c)=2/4=1/2$

$P(E^c)=3/5$

Así, $P(F) = P(F|E)P(E) + P(F|E^c)P(E^c)=(1/4)*(2/5)+(1/2)*(3/5)=(2/20)+(3/10)=(2+6)/20$. Esto es $P(F)=8/20=2/5$, lo cual es distinto a $P(F|E)$, concluyendo que no es independiente de la ocurrencia de E.

Con la Ley de probabilidad total tenemos que la medida que se le asigne a A es "igual a media ponderada de las probabilidades condicionadas de que ocurra A dado que B haya ocurrido y de que ocurra A dado que B no haya ocurrido" (Ross, p.185), ponderación que está bien hecha, pues $P(B) + P(B^c) = 1$ (Ley del complemento), ya que $B \cup B^c = S$, como descomposición del espacio muestral.

Entonces, ya con esta poderosa herramienta del pensamiento racional moderno, nos preguntamos por el asunto que abrió esta sección, a saber: ¿cómo reevaluar una probabilidad a priori a la luz de una evidencia adicional?.

En principio, supóngase que se está estudiante una cierta proposición hipotética. Sea H el evento que denota a tal proposición y P(H) representará la probabilidad de que sea verdad. A continuación, se consigue

una información adicional asociada al evento llamada E, relacionado con H, dispuesta a refutar a H o a fortalecerla. La reformulación de la hipótesis consiste entonces en evaluar P(H|E). Por la definición de probabilidad condicional se tiene que:

$$P(H|E)=P(H\cap E)/P(E)$$

La cual es, por la ley de multiplicación, aplicada a $P(H\cap E)=P(E|H)$ P(H), igual a:

$$P(H|E)= P(E|H)P(H)/P(E)$$

Si a la probabilidad de E del denominador se le calcula con la Ley de probabilidad total de acuerdo a que H sea una hipótesis verdadera o sea falsa, se obtiene una expresión que se conoce como **teorema de Bayes** y básicamente permite intercambiar en, el cálculo de probabilidades condicionales, el evento condicionado por el evento condicionante, como se sintetiza a continuación:

$$\mathbf{P(B|A) = (P(A|B)P(B)) \div (P(A|B)P(B) + P(A|B^c)P(B^c))}$$

Ejemplo: Un consultora realiza un informe actuarial sobre los riesgos de los empleados de una empresa determinada. En ella se informa que hay dos tipos de trabajadores 1) los descuidados, propensos a tener accidentes y 2) aquellos que son más cautelosos, no propensos a tener

accidentes. Los datos que recolectaron indican que, luego del periodo de formación y prueba de nueve meses de trabajo, una persona descuidada sufrirá uno con probabilidad 0.1, mientras que los cautelosos sufrirán un accidente luego del mismo plazo de tiempo con probabilidad 0.05. Así mismo en el informe se establece que la probabilidad de contratar a un empleado descuidado es de 0.2. Con base en dicho informe, se necesita saber:

a) ¿Cuál es la probabilidad de que un nuevo empleado sufra un accidente luego del periodo de prueba?

b) Si un nuevo empleado tuvo un accidente, luego de los nueve meses, ¿cuál es la probabilidad de que sea del tipo de los descuidados?

Primero, designemos los eventos asociados a la hipótesis según lo referido en el informe. Sean los eventos A={"el empleado nuevo tiene un accidente luego de los nueve meses"}

D={"el empleado nuevo es descuidado"}

D^c={"el empleado nuevo es cauteloso"}

De forma que, según los datos se tiene que:

$P(A|D)=0.1$

$P(A|D^c)=0.05$

$P(D)=0.2$

De acuerdo a la Ley del complemento, se tiene que $P(D^c)=1-P(D)=1-0.2=0.8$, que es la probabilidad de contratar a alguien cauteloso. Entonces, ya estamos en condiciones de calcular la probabilidad de que un nuevo empleado sufra un accidente, utilizando la Ley de probabilidad total, dado que un nuevo empleado o es descuidado o es cauteloso:

$$P(A)=P(A|D)P(D)+P(A|D^c)P(D^c)$$

$$P(A)=(0.1)*(0.2)+(0.05)*(0.8)=0.06$$

Por lo tanto existe un posibilidad del 6% de que un nuevo ingreso en la empresa sufra un accidente luego de los nueve meses de formación y prueba.

Ahora, veamos que conocemos la probabilidad de un descuidado sufra un accidente, pero no sabemos cuál es la probabilidad de que alguien que haya sufrido un accidente sea un descuidado. Para ello utilizaremos la fórmula del teorema de Bayes y la ya conocida probabilidad de que el nuevo tenga un accidente:

$$P(D|A)= P(A|D)P(D)/P(A)$$

$$P(D|A)=(0.1)*(0.2)/(0.06)=1/3$$

Entonces, si un nuevo empleado tiene un accidente, la probabilidad

de que dicho empleado sea descuidado es igual a un 1/3. Hemos visto puesto la potencia del teorema de Bayes en la inferencia a posteriori de ciertas probabilidades.

Este tipo de información, en el caso particular del evento, puede ayudar reforzar las medidas de seguridad, mejorar el proceso de formación o aumentar el período de prueba para determinar si un nuevo ingreso es conveniente o no para el empleo, dada su propensión a los accidentes.

3.6 Elementos de conteo y muestreo aleatorio.

La naturaleza de la mayoría de los experimentos que se desean medir con probabilidades implica la observación de una muestra elegida de una gran cantidad de datos que se identifican como población. Dependiendo de la especificidad de la medida al objeto de estudio, los experimentos se pueden reducir a **un modelo de extracción de una muestra de tamaño n de una población de tamaño N de elementos distinguibles entre sí**, tal y como se extrajeron las dos pelotas de la caja de cinco, en ejemplos anteriores.

Como medir probabilidad es bastante similar a contar, se recurre a hechos combinatorios para ver de cuántas formas pueden suceder eventos determinados de la índole anterior, es decir, **de cuántas formas se pueden**

muestrear tales o cuales individuos de tal o cual población sujeta a determinadas condiciones. No restringiremos a dos condiciones particulares:

1. **Reposición**: si el muestre se realiza con o sin reemplazo del elemento muestreado, permitiendo así o no la repetición del elemento que se tomó una vez.

2. **Orden**: si el muestreo se realiza con o sin orden en la extracción, esto es, distinguiendo si la extracción distingue del primero (segundo, tercero, etc.) en ser tomado o no.

A partir de tal distinción se construye la siguiente tabla de combinatoria establece de cuántas formas se puede realizar **un muestreo aleatorio simple de tamaño n sobre una población de tamaño N**:

MUESTREO	Ordenado	No ordenado
Con reemplazo	N^n	(N+n-1)Cn
Sin reemplazo	NPn	NCn

Tab. 3.3

Donde:

- La potencia N^n significa que se multiplica el factor N, n-veces.
- El factorial significa el producto sucesivo y decreciente N!=N*(N-1)*(N-2)..3*2*1, con la condición de que 0!=1

- La permutación NPn=N!/(N-n)!
- Y el combinatorio NCn=N!/n!(N-n)!

Como ejemplo, calculemos de cuántas formas se pueden escoger 2 nuevos empleados de 4 candidatos para una misma posición laboral. Tenemos un conjunto de aplicantes para un trabajo en el conjunto {C1, C2, C3, C4}. Y se desean escoger tan solo dos, sin ningún orden, es decir, es lo mismo {C1, C4} que {C4, C1}. Estamos pues ante un escenario de muestreo aleatorio simple sin reposición y sin orden. Luego, tras mirar la tabla 3.3, utilizamos la fórma NCn, con N=4 y n=2. Así, 4!=4*3*2*1=24, 2!=2*1 y (4-2)!=2!=2. Luego, 4C2=6. En este caso, para mostrar la validez de la fórmula al lector incrédulo, podemos listar las 6 posibles combinaciones: {C1, C2}, {C1, C3}, {C1, C4}, {C2, C3}, {C2, C4} y {C3, C4}.

El muestro aleatorio perfecto es difícil de conseguir en la práctica. Usualmente se utilizan tablas de números aleatorios y ordenamientos.

3.7 Comentarios finales

La probabilidad es una cuantificación de la posibilidad de que ocurra un suceso dado. Por lo tanto, las probabilidades son una medida del nivel de certidumbre-incertidumbre asociado a cada uno de los eventos de un

experimento aleatorio en cuestión. Los valores de la probabilidad, como función de conjunto, se encuentran en una escala de 0 a 1, donde los valores cercanos al 0 son aquellos con posibilidades de ocurrencia muy bajas. Mientras que los cercanos a 1 indican que es casi seguro que ocurran.

Este no es más que un modelo para la repetición de un experimento, es decir, una teoría para las distribuciones de frecuencia de capítulos anteriores. Conocer esta teoría es, por lo general, el primer paso para construir un mecanismo que permite hacer inferencias.

Se definieron entonces que es un evento, evento elemental o simple, espacio muestral, suceso compuesto, propiedades de la probabilidad, condicionamiento e independencia entre eventos, así como un conjunto de leyes útiles para los cálculos a través del método de los puntos muéstrales.

A este punto es natural formularse la pregunta sobre el valor de la teoría de probabilidades. Se trata, pues, de un fundamento teórico-práctico para asignar una medida de certeza de ocurrencia a eventos aleatorios y, por lo tanto, a distribuciones de variables aleatorias que se estudiarán a continuación. Los eventos de nuestro próximo interés están relacionados con la obtención de información a partir de procesos de muestreo, luego, el modelo de la teoría de probabilidades está en función de una potencialidad de la estadística, que es hacer inferencias sobre una población a partir de

una pequeña parte de sí.

Ejercicio

En una encuesta se les pidió a 1035 adultos y adultas su opinión respecto a los negocios e inversiones. Una de los ítems de dicho instrumentos investigativo era: "¿Cómo califica usted a los pequeños emprendimientos respecto a la calidad de los productos y servicios que ofrecen?". Las respuestas a tal pregunta fueron:

- Excelentes, 18%;
- Bastante bien, 50%;
- Regulares, 26%;
- Malas, 5% y
- No saben / no contestan, 1%.

Diga cuál es la probabilidad de que un encuestado considere a los pequeños emprendimientos bastante buenos o excelentes, cuántos de los encuestados consideraron malos a los pequeños emprendimientos, y cuántos de los encuestados dijo no saber o no contestó.

CAPÍTULO IV
Distribuciones de probabilidad

MAWIL
Publicaciones Impresas
y Digitales

www.mawil.us

4. Distribuciones de probabilidad

Los eventos de mayor interés para la administración y los negocios se asocian a números y reciben la denominación de **eventos numéricos**. La ventaja que ofrece este abordaje es considerable, debido a que homologa experimentos en apariencia distintos, como lanzar una moneda y mirar si sale cara o sello y lanzar un dado y mirar si sale par o impar. De forma que en este capítulo nos ocuparemos del modelado matemático que la teoría de la probabilidad provee para este tipo de eventos.

4.1 Eventos numéricos, variables aleatorias, distribuciones y otros conceptos

Un **evento numérico** es un asociación entre los elementos de un espacio muestral y un conjunto numérico cualquiera. Esta asociación se lleva a cabo a través de una función que se denomina variable aleatoria X, que define unívocamente a la observación, es decir, cuyo dominio es el espacio muestral S y su rango está contenido en los números reales:

$$X: S \rightarrow N$$

Por ejemplo, sea el experimento que consiste en lanzar 3 monedas y registrar (observar) el número de caras que salieron, sin que importe el or-

den. Veamos, a través de un cuadro, cómo se define tal variable aleatoria.

Evento elemental	Variable aleatoria
{S,S,S}=No sale ni una cara	X=0
{C,S,S}=Sale una cara	X=1
{C,C,S}=Salen dos caras	X=2
{C,C,C}=Salen tres caras	X=3

El **rango de una variable aleatoria** es el conjunto de los posibles valores R(X) que puede tomar esta. En el ejemplo anterior, el rango es R(X)={0, 1, 2, 3}.

En general, las variables aleatorias se clasifican en dos tipos:

1) **Discretas**: la variable aleatoria es discretas si su rango R(X) es conjunto numérico discreto, sea finito, como en el ejemplo anterior, o infinito, como los números naturales o enteros.

2) **Continuas**: la variable aleatoria es continua si su rango R(X) es un conjunto numérico continuo, como los números reales.

Si por ahora tratamos exclusivamente con variables aleatorias X discretas y finitas, es decir, que

$$R(X)=\{X1, X2,\ldots, Xk, \ldots , Xn\}$$

Podemos a cada uno de tales valores asignarle una y solo una probabilidad tal que:

$$p1=P(X=X1),\ p2=P(X=X2),\ldots,\ pn=P(X=Xn)$$

De modo tal que $p1+p2+\ldots+pk+\ldots+pn=1$, pues se han recorrido todos los posible valores del espacio muestral al asignar una probabilidad a cada punto del rango de la variable aleatoria. Así, hemos definido una **distribución de probabilidad discreta para la variable aleatoria X**, esto es, una asignación de número pk mayor o igual que cero de forma que la suma de todas las asignaciones no exceda en suma a la unidad, de acuerdo a las propiedades de la función de probabilidad que estudiamos en el capítulo anterior. Usualmente, tales distribuciones, cuando se presentar con una fórmula que depende del valor del rango de la variable aleatoria o, cuando es posible, en una tabla como la siguiente:

X	X1	X2	X3	…	Xk	…	Xn-1	Xn
P(X)	p1	p2	p3	…	pk	…	pn-1	pn

Tab 4.1

Siendo esta representación tabular oportunamente análoga con aquella de las **distribuciones de frecuencias relativas** que se estudiaron en capítulos anteriores. Así, "se pueden considerar a las distribuciones de probabilidad como formas teóricas o ideales del límite de distribuciones de frecuencias relativas cuando el número de observaciones realizadas es muy grande". (Schaum, 130) De allí que sea necesario distinguir, como se

hará próximamente, en distribuciones poblacionales y muestrales, sobre todo cuando se esté ante un experimento que toma parte de un todo que se desea no solamente describir sino investigar a partir de inferencias.

Si se hace una analogía con las tablas de frecuencia relativa acumulada, se puede enunciar el concepto de **distribución de probabilidad acumulada** como aquella función que va sumando, acumulando, todas las probabilidades anteriores.

En resumen, podemos decir que cualquier distribución de probabilidades discreta debe cumplir:

- $0 \leq pk \leq 1$, para todo k.
- $\sum pk=1$, sumando sobre todos los k.

Para ejemplificar todos los nuevos conceptos, supongamos que se lanzan dos dados de colores distintos, justos y de 6 caras. La observación de interés que definirá la variable aleatoria X es la suma de los resultados obtenidos en cada dado. Así, el rango de tal variable aleatoria será

$$R(X)=\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

Ahora, estudiemos con detalle la probabilidad de obtener cada uno de esos resultados, pues, es más o menos evidente que hay resultado más probables que otros. Para ello, utilizaremos las siguientes hipótesis:

- Representaremos con un par ordenado el resultados de ambos dados (D1,D2).
- Vemos que el espacio muestral se cuenta como un muestreo ordenado y con reemplazo de una población de 6 elementos (caras) y una muestra de tamaño 2 (lanzamientos), luego #S=6*6=36.
- Recordemos que la definición de probabilidad clásica es P(A)=#A÷#S.
- Nos ayudaremos con la tabla para determinar #A.

Índice k	**Variable aleatorio X**	**Eventos elementales que componen a la suma**	**Probabilidad P(X)=pk**
1	X=2	Solo es posible cuando salga (1,1)	1/36
2	X=3	(1,2) y (2,1)	2/36
3	X=4	(2,2), (1,3) y (3,1)	3/36
4	X=5	(4,1), (3,2), (2,3) y (1,4)	4/36
5	X=6	(5,1), (4,2), (3,3), (2,4) y (1,55)	5/36
6	X=7	(6,1), (5,2), (4,3), (3,4), (2,5) y (1,6)	6/36
7	X=8	(6,2), (5,3), (4,4), (3,5) y (2,6)	5/36
8	X=9	(6,3), (5,4), (4,5) y (3,6)	4/36

9	X=10	(5,5), (6,4) y (4,6)	3/36
10	X=11	(6,5) y (5,6)	2/36
11	X=12	Solo es posible cuando salga (6,6)	1/36

Nótese que la suma ∑pk=(1+2+3+4+5+6+5+4+3+2+1)/36=36/36=1. Grafiquemos dichos valores para exhibir la simetría con la que se distribuyeron las probabilidades:

Dada la simetría de la distribución, vemos que el valor central, el más probable, es que la suma de los puntos de los dados de 7. Para describir esto estaríamos tentados a recurrir a la media aritmética que se estudió anteriormente, pero es necesario cierta aproximación teórica, es decir, con la teoría de las probabilidades, para ello es necesario introducir un nuevo concepto, que es el de esperanza. Se define y denota a la esperanza de una variable aleatoria X por:

$$E(X)=X1*p1+X2*p2+\ldots+Xn*pn=\sum Xk*pk=\sum X*P(X)=\mu$$

Es de notar que si las probabilidades pk se sustituyen por las frecuencias relativas, se tendría que $E(X)=\overline{x}$. Luego, esta analogía con la media conduce a la siguiente definición:

- Media muestral: $\overline{x}$.
- Media poblacional: $\mu=E(X)$.

Volviendo al ejemplo de los dos dados, calculemos su esperanza para cercionarnos que coincide con lo que gráficamente inferimos.

E(X)=2*(1/36)+3*(2/36)+4*(3/36)+5*(4/36)+6*(5/36)+7*(6/36)+ 8*(5/36)+9*(3/36)+10*(3/36)+ 11*(2/36)+ 12*(1/36)=252/36=7

Además de la esperanza E(X), es necesario definir otros conceptos que caracterizan a una variable aleatoria, a saber, la **varianza** y la **desviación estándar**:

- $Var(X)=E((X-E(X))^2)$.
- $DS(X)=\sqrt{Var(X)}$.

Cuando se trata con variables aleatorias X discretas infinitas y con variables aleatorias continuas, es necesario dominar conocimientos básicos de análisis real matemático, en particular, series e integración pues R(X) deja de ser un conjunto fácilmente manejable. Se verán ejemplos posteriormente, que se enunciarán con un condensado de sus características y experimentos que modelan. Sin embargo es necesario hacer un par de comentarios al respecto.

Cuando X toma un conjunto infinito continuo de valores, el histograma que representa a la distribución se convierte en una curva de trazo continuo, la probabilidades de eventos puntuales son nulas y los eventos

de interés se presentan como intervalos [a,b], es decir, se miran las probabilidades

$$P(a \leq X \leq b) = \int_{[a,b]} f(x)dx$$

Donde f(x) se denomina **función de densidad de probabilidad** y viene a ser el concepto que sustituye o, mejor dicho, que extiende a la distribución de probabilidad de las variables cuando son discretas. En ese caso, no se trata de un conteo sino de una medida de área bajo una curva. Igual que el caso discreto, se puede definir las distribuciones de probabilidad acumulada para variables aleatorias continuas.

4.2 Distribuciones discretas célebres: binomial, geométrica, hipergeométrica y Poisson

La binomial. El muestreo binario

Supongamos que se lleva a cabo un experimento que consiste en realizar N pruebas independientes entre sí cuyos resultado es binario, digamos, "éxito" o "fracaso". Se puede obtener un "éxito" con probabilidad p, mientras que por complemento, el "fracaso" ocurriría con probabilidad 1-p. Sea entonces la variable aleatoria X que mide el número de éxitos que

ocurren en las N pruebas. Se trata de una variable discreta y finita, cuyo rango es {0, 1, 2, 3,…N} y sus probabilidades son conocidas. También está asociada a experimentos de muestreo binario con reposición, para garantizar independencia. Se dice que X sigue una ley de **distribución binomial** con parámetros N y p, con una distribución de probabilidad igual a:

$$pk=P(X=k)=(NCk)*p^{k}*(1-p)^{N-k}$$

con k siendo cualquier número entre 0 y N. Su forma es acampanada, solamente simétrica cuando p=1/2. Su esperanza y varianza son:

$E(X)=Np$

$Var(X)=Np(1-p)$

Veamos ahora un ejemplo. Si el 20% de las inversiones recomendadas por una consultora de riesgo son un fracaso, ¿cuál es la probabilidad de que tras realizar al azar 4 inversiones, independientes entre sí, sugeridas por la consultora:

- Tres (3) sean un éxito,
- Todas sean un éxito, y
- A lo sumo 2 sean un fracaso?

Estamos ante un caso de distribución binomial, donde los ensayos

son hacer inversiones, en este caso N=4 inversiones, con probabilidad de fracaso 1-p=0.2, es decir, p=0.8.

Para responder la primera, debemos calcular

$P(X=3)=(4C3)*(0.8)^3(0.2)^1=4*(0.512)*(0.2)=0.4096$

Del mismo modo que:

$P(X=4)=(4C4)*(0.8)^4(0.2)^0=0.4096$

En cuanto a la última, interpretemos que "a lo sumo 2 fracasos" significa, no más de 2 fracasos, es decir, pueden haber 0 fracasos (todos éxitos, X=4), 1 fracaso (3 éxitos, X=3) y 2 fracasos (2 éxitos, X=2). Calculemos ahora:

$P(X=2)=(4C2)*(0.8)^2(0.2)^2=6*(0.64)*(0.04)=0.1536$

Por lo que

P(A lo sumo 2 sean un fracaso) = P(al menos 2 éxitos) = $P(X\geq2)$ = $P(X=2) + P(X=3) + P(X=4) = 0.4096+0.4096+0.1536=0.9728$

Este tipo de información es necesaria en situaciones de riesgo de pérdidas y asignación de recursos en productos de inversión. ¿Cuál es en este caso el valor esperado de éxitos? $E(X) = 4*(0.8) = 3.2 \sim 3$.

La geométrica. Las rachas de fracasos para un obtener un éxito

Ahora, en el mismo supuesto de hacer pruebas binarias, pero observando la variable Y aleatoria que cuenta el número necesario de pruebas para obtener un éxito con probabilidad p, nos encontramos con una variable aleatoria discreta pero no finita, pues su rango es

$$R(Y)=\{1, 2, 3, 4, 5, 6, 7, \ldots.\}$$

pues es posible (aunque poco probable, como se verá) tener una racha de fracasos arbitrariamente grande. Este es el caso en donde se dice que Y sigue una ley de **distribución geométrica** con parámetro p, cuya distribución de probabilidades viene dada por la fórmula:

$$pk=P(Y=k)=(1-p)^{k-1}p$$

Su forma es exponencial decreciente. Su esperanza y varianza son:

$$E(Y)=1/p$$

$$Var(Y)=(1-p)/p^2$$

Esta distribución tiene una peculiaridad relevante que suele enunciarse como que la variable carece de memoria, en el sentido de que

$$P(Y>m+n \mid Y>m) = P(Y>n)$$

Esto significa que la distribución de condicional del número de en-

sayos no depende de cuantos fracasos se hayan observado antes. El ensayo que se produce no tiene "memoria" de los fracasos que se acumulan en rachas.

La hipergeométrica. El muestreo binomial sin independencia

La **distribución hipergeométrica** está asociada a una variable aleatoria discreta Z que modele muestreos binarios aleatorios y sin reemplazo, es decir, pierde la independencia que se supone en los experimentos binomiales. Supongamos que se tiene una población de N elementos de los cuales, r son del tipo I y N-r son del tipo II. La distribución hipergeométrica mide la probabilidad de obtener Z=k elementos de tipo I en una muestra sin reemplazo de n elementos de la población original, considerando a k≤n. Los parámetros de esta distribución son N, r y n y su distribución de probabilidades viene dada por la siguiente fórmula:

$$pk=P(Z=k)=(rCk*(N-r)C(n-k))\div(NCn)$$

Su forma es asimétrica. Su esperanza es:

$$E(Z)=n*r/N$$

y su varianza es

$$Var(Z)=(N-n)p(1-p)/(N-1) \text{ con } p=r/N$$

Veamos ahora un ejemplo: considerando que en la caja hay un total

de 10 objetos, 3 de los cuales son defectuosos, si de seleccionan 4 objetos al azar, ¿cuál es la probabilidad de que 2 sean defectuosos?.

Identificamos primero los parámetros.

Cantidad total de objetos N = 10, cantidad de objetos tipo I (defectuosos) r = 3, tamaño de la muestra, n = 4.

Nos preguntamos por P(Z=2), esto es

$$P(Z=2)=(3C2*7C2)\div(10C4)=3/10=0.3$$

Relación entre la hipergeométrica y la binomial

Aunque la distribución hipergeométrica modele un experimento radicalmente opuesto al binomial, puede ser reducida a esta. Cuando N es grande con respecto a n, es decir, la población que se muestrea es considerablemente más grande que la muestra, se toma a p=r/N (la probabilidad de escoger un objeto de tipo I y Z se distribuiría binomial con parámetros N, p. Al respecto, debe apuntarse el hecho que fundamenta esta equivalencia: cuando se toman muestras pequeñas de una población muy grande, hacerlo con o sin reemplazo da resultados similares, pues tomado un elemento, dado el tamaño de la población, se hace muy remota la posibilidad de reposición.

La Poisson. Ocurrencia raras y la binomial al límite

La **distribución de Poisson** es una distribución de probabilidad discreta que mide, a partir de una frecuencia de ocurrencia media λ, la probabilidad de que ocurra un determinado número de eventos durante cierto intervalo de un continuo como el tiempo o el espacio en longitud. Aunque es una distribución extravagante y oscura –como se verá en su forma matemática–, se trata de un caso límite de distribución binomial cuando N es muy grande y p tiende a 0, esto es, cuando el "éxito" es raro o inusual. Su rango es

$$R(X)=\{0, 1, 2, 3, 4,\ldots.\}$$

Y su distribución viene dada por

$$pk=P(X=k)=(\lambda^k \exp(-\lambda))\div(k!)$$

donde $\exp(-\lambda)=e^{-\lambda}$ y $e=2.71828\ldots$

Su único parámetro es el número $\lambda > 0$, y tiene la particularidad de que su esperanza y varianza son:

$E(Y) = \lambda$

$Var(Y) = \lambda$

Como si comentó, en la práctica ella puede ser una aproximación

límite a la distribución binomial cuando $N>50$ y $p \rightarrow 0$, tomando a $\lambda=Np$.

La distribución de Poisson se aplica a numerosos fenómenos discretos del mundo, en particular, para aquellos fenómenos que ocurren n-veces durante un periodo definido de tiempo o en un área dada, siempre y cuando la probabilidad de ocurrencia del evento sea baja y uniforme en el tiempo o el espacio. Ejemplos de tales eventos comprenden:

- El número de automóviles que pasan a través de un intersección durante un periodo definido de tiempo.
- El número de errores de redacción que se cometen comete al escribir una página.
- El número de llamadas telefónicas en una central telefónica por minuto.
- La inventiva y producción de un tecnólogo a lo largo de su vida activa, y
- La distribución de la riqueza humana.

Veamos un ejemplo: en un banco el promedio de atención es 4 clientes por hora, encuentre la probabilidad que en 30 minutos se atiendan menos de 3 personas y que en 180 minutos se atiendan 12 clientes.

Primero estandaricemos la medida de tiempo a media hora, de modo que si el promedio de atención es 4 clientes por hora, será 2 clientes por cada media hora. Tal será nuestro parámetro lambda $\lambda=2$. Así, estamos en condiciones de resolver

$$P(X<3)=P(X=0)+P(X=1)+P(X=2)=(\lambda^{o} \exp(-\lambda))\div(0!)+ (\lambda^{1} \exp(-\lambda))\div(1!)+ (\lambda^{2} \exp(-\lambda))\div(2!)$$

Efectuando las operaciones involucradas tenemos que:

$$P(X<3)=\exp(-2)+2* \exp(-2)+4 \exp(-2)/2=0.1353+0.2706+0.2706=0.6765$$

En cuanto a los 180 minutos para atender 12 clientes, podemos reformular el parámetro lambda, dado que 180 minutos son 3 horas, y si se atienden 4 clientes por hora –en media–, el lambda para 180 minutos sería $\lambda = 12$ y

$$P(Y=12)= (\lambda^{12} \exp(-\lambda))\div(12!)= (12^{12} \exp(-12))\div(12!)=0.1143$$

4.3 Distribuciones continuas célebres: uniforme, exponencial y normal.

Para describir a las distribuciones continuas, nos restringiremos a exhibir sus densidades, parámetros, esperanza, varianza y, en general, cómo se utilizan, salvo con la distribución normal, que por ser un caso

particularísimo merece especial atención. Deben recordarse dos asuntos fundamentales en el tratamiento de una variable aleatoria continua que se fundamentan en el hecho de que ahora los eventos numéricos deben darse en intervalos o combinaciones de intervalos.

- Ya no se estudian las probabilidades de que una variable aleatoria tome un determinado valor puntual, se abordan medidas de certidumbre donde las variables aleatorias tomen un valor dentro de un intervalo determinado.

- La probabilidad de que una variable aleatoria continua tome un valor dentro de un determinado intervalo que va de a hasta b, se define como el área bajo la curva de la función de densidad de probabilidad entre a y b. Debido a que un solo punto es un intervalo cuya longitud es nula, esto implica que la probabilidad de que una variable aleatoria continua tome un valor puntual, es cero. Es decir, que cualquier intervalo la probabilidad de que una variable aleatoria continua tome un valor es la misma, ya sea que se estudie el intervarlo (a,b) o el [a,b].

La uniforme

Sea una variable aleatoria T que mide el tiempo de espera en una terminal de autobuses. Supongamos que cada media hora llega un vehículo. Como que la variable aleatoria T toma cualquier valor en el intervalo de [0,30] minutos, T es una variable aleatoria continua y no discreta. Dado que lo aleatorio es la llegada a la terminal, la probabilidad de que el tiempo de espera esté en cualquier intervalo de 15 minutos es el mismo que la probabilidad de que el tiempo de espera esté en cualquier otro intervalo de 15 minutos dentro del intervalo que va de 0 a 30 minutos. Esto quiere decir que no se tiene preferencia por una hora de llegada, no hay horas pico. Como cualquier intervalo de 15 minutos es igual de probable, se dice que la variable aleatoria T tiene una **distribución de probabilidad uniforme**. La función de densidad de probabilidad que define la distribución uniforme del tiempo de espera, es:

$U(t)=1/30$ si $0 \leq t \leq 30$

$U(t)=0$ en caso contrario

De forma tal que:

$$\mathbf{P(t_i \leq T \leq t_f) = (t_f - t_i) / 30}$$

En general, la distribución uniforme tiene dos parámetros: a y b, su

densidad está dada por

$U(t)=1/(b-a)$ si $a \leq t \leq b$

$U(t)=0$ en caso contrario

Su dibujo no es más que un escalón constante, es simétrica y muy fácil de utilizar. Su esperanza y varianza son:

$E(X)=(a+b)/2$

$Var(X)=(b-a)^2/12$

Siempre que los números $a<b$.

La exponencial. Los tiempos de espera

La **distribución exponencial** proviene de considerar como variable aleatoria a la longitud de los intervalos continuos –de tiempo, por ejemplo– que transcurren entre dos sucesos discretos distribuidos según Poisson. Así, por ejemplo, el tiempo transcurrido en una central telefónica antes de recibir una llamada se podría modelar como una distribución exponencial. También el intervalo de tiempo entre eventos telúricos sigue una distribución exponencial.

La distribución exponencial es una distribución de probabilidad con-

tinua con un parámetro λ cuya función de densidad es:

$$f(t)=\lambda*\exp(-\lambda t) \quad \text{si } t > 0$$

$$f(t)=0 \qquad \text{si } t \leq 0$$

Tal y como la distribución de Poisson, con la que está íntimamente relacionada, su único parámetro es el número $\lambda > 0$, y su esperanza y varianza son:

$$E(T) = 1/\lambda$$

$$Var(T) = 1/\lambda^2$$

Su gráfica, para tres parámetros distintos, se muestra a continuación en la fig. 4.2:

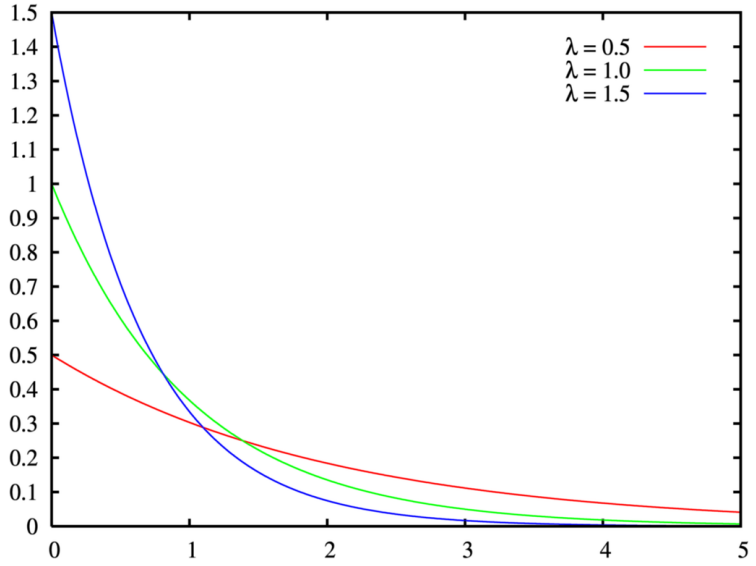


Fig.4.2

Para calcular la probabilidad de que una variable aleatoria T ocurra

en el intervalo [a, b], se calcula resolviendo una integral más o menos sencilla, suponiendo que a>0:

$$P(a \leq T \leq b) = \int_{[a,b]} f(t)dt = \int_{[a,b]} \lambda * \exp(-\lambda t)\, dt = \exp(-\lambda a) - \exp(-\lambda b)$$

Esta distribución es la única continua que carece de memoria, en el sentido mismo sentido que la Poisson, esto es:

$$P(T>m+n \mid T>m) = P(T>n)$$

Cuando T se interpreta como el tiempo de espera para un evento que se produzca con respecto a un tiempo inicial, la pérdida de memoria implica que, si T está condicionada a un fracaso para observar el evento sobre un cierto período de tiempo inicial [0, m], la distribución del tiempo restante de espera es la misma que la distribución incondicional. Por ejemplo, si un evento no se ha producido después de 30 segundos, la probabilidad condicional de que ocurrencia llevará al menos 10 segundos más es igual a la probabilidad incondicional de observar el evento más de 10 segundos después del tiempo inicial.

La normal. Una distribución universal

La distribución más utilizada para describir variables aleatorias continuas X es **la distribución de probabilidad normal**. La distribución nor-

mal tiene un inmenso volumen de aplicaciones prácticas, en las cuales la variable aleatoria puede ser el peso, la estatura de un conjunto de humanos, las puntuaciones de evaluaciones, los resultados de medidas científicas, datos atmosféricos, entre otras cantidades análogas. La distribución normal también tiene una importante aplicación para la estadística que procura inferir, tema del que nos ocuparemos en los próximos capítulos. En tales aplicaciones, la distribución normal muestra qué tan probables son los resultados obtenidos de un muestreo aleatorio y sus estadísticos asociados.

La distribución normal está caracterizada por los **dos valores que definen su esperanza y varianza, los cuales son, además, sus parámetros**:

$$E(X)=\mu$$

$$Var(X)=\sigma^2$$

La curva de su densidad tiene forma acampanada, simétrica con respecto a su esperanza, representada por la letra griega mu. Su variabilidad o anchura está determinada por la desviación estándar, la raíz cuadrada de la varianza, que se representa por la letra griega sigma. En la figura 4.3 se exhiben varias curvas normales para dos valores distintos de su centro μ y cuatro valores para su dispersión.

Dada la simetría que tiene con respecto μ, se tiene que es igualmente probable que X se encuentre por encima o por debajo de μ, esto es:

$$P(X<\mu)=P(\mu<X)=1/2$$

Su densidad tiene una fórmula cerrada, no así su distribución de probabilidad acumulada. La altura de la curva de densidad viene dada por:

$$f(x)=(1/\sqrt{2\pi\sigma})*\exp(-(x-\mu)^2/2\,\sigma^2)$$

Con π=3.1415…

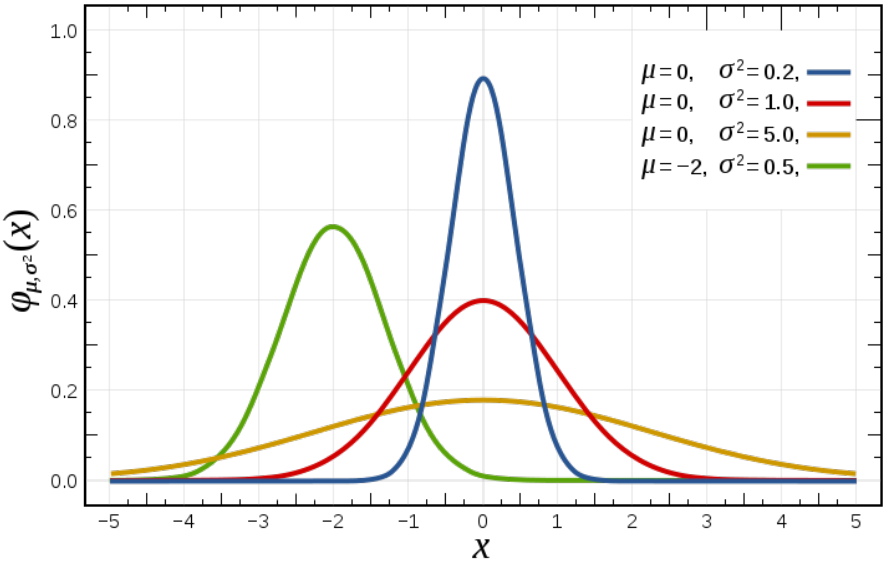


Fig.4.3

Cuando μ=0 y σ=1, la variable aleatoria se dice que la variable aleatoria tiene una **distribución normal estándar** y se reserva la letra **Z** para representarla. Toda variable aleatoria X que se distribuya normal mu, sigma -abreviado **X ~ N(μ,σ)**-, puede ser estandarizada para estar en condi-

ciones de calcularle su probabilidad pues, la variable Z definida por

$$\mathbf{Z=(X-\mu)/\sigma}$$

Tiene la propiedad de que se distribuye N(0,1), es decir, X se estandarizó, por lo que:

$$P(a<X<b)= P(a-\mu<X-\mu<b-\mu)$$

$$P(a<X<b)= P((a-\mu)/\sigma<(X-\mu)/\sigma<(b-\mu)/\sigma)$$

$$P(a<X<b)= P((a-\mu)/\sigma<Z<(b-\mu)/\sigma)$$

$$\mathbf{P(a<X<b) = P(a_z<Z<b_z)}$$

Donde $a_z=(a-\mu)/\sigma$ y $b_z=(b-\mu)/\sigma$, son los valores estandarizados de los números a y b que definen a un evento numérico que se desea medir a través de una distribución normal de X. Así, en general, para calcular las probabilidades de eventos aleatorios normales, basta solo conocer los valores de la distribución normal estándar.

Dada Z una variable aleatoria normal estándar, la probabilidad de que Z esté en el intervalo [a,b] es igual al área bajo la densidad normal estándar entre a y b.

Dada la dificultad de calcular tales áreas analíticamente, se han encontrado **propiedades algebraicas de la normal** deducidas a partir de

propiedades básicas de las áreas, vinculándolas con las leyes de la probabilidad y aprovechando la simetría de la curva normal. Tales propiedades se sintetizan en las siguientes fórmulas:

1. $P(a<Z<b) = P(Z<b) - P(Z<a)$
2. $P(Z < -a) = P(Z>a)$
3. $P(Z>a) = 1 - P(Z<a)$
4. $P(|Z|>a) = 2P(Z>a)$ siempre que $a>0$

Para utilizar tales fórmulas y calcular efectivamente los valores para la **$Z \sim N(0, 1)$**, se han **tabulado una serie de valores que permite calcular dicha probabilidades** y que se presenta en la tabla 4.4. En dicho cuadro se encuentran los valores para determinar

$P(Z<x)$ con $x \geq 0$

Por eso su primer valor es igual a ½=0,5. Su forma de utilizar es bastante simple, pues se trata de un sistema matricial (fila-columna) para distribuir cuantiles reales a 2 decimales.

Así, si queremos hallar el valor de $P(Z<1,54)$, debemos descomponer a

$$1,54 = 1,5 + 0,04.$$

A continuación buscamos la fila bajo la x que esté nominada por 1,5 y luego nos movemos hacia la columna 0.04 donde encontraremos el valor deseado para concluir que:

P(Z<1,54)=0,9382

x	0	0,01	0,02	0,03	**0,04**	0,05	0,06	0,07	0,08	0,09
0	0,5	0,504	0,508	0,512	**0,516**	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	**0,5557**	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,591	**0,5948**	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	**0,6331**	0,6368	0,6406	0,6443	0,648	0,6517
0,4	0,6554	0,6591	0,Ç6628	0,6664	**0,67**	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,695	0,6985	0,7019	**0,7054**	0,7088	0,7123	0,7157	0,719	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	**0,7389**	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,758	0,7611	0,7642	0,7673	**0,7704**	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,791	0,7939	0,7967	**0,7995**	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	**0,8264**	0,8289	0,8315	0,834	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	**0,8508**	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	**0,8729**	0,8749	0,877	0,879	0,881	0,883
1,2	0,8849	0,8869	0,8888	0,8907	**0,8925**	0,8944	0,8962	0,898	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	**0,9099**	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	**0,9251**	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	**0,9332**	**0,9345**	**0,9357**	**0,937**	0,9382	**0,9394**	**0,9406**	**0,9418**	**0,9429**	**0,9441**
1,6	0,9452	0,9463	0,9474	0,9484	**0,9495**	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	**0,9591**	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	**0,9671**	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	**0,9738**	0,9744	0,975	0,9756	0,9761	0,9767

2	0,9772	0,9778	0,9783	0,9788	**0,9793**	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,983	0,9834	**0,9838**	0,9842	0,9846	0,985	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	**0,9875**	0,9878	0,9881	0,9884	0,9887	0,989
2,3	0,9893	0,9896	0,9898	0,9901	**0,9904**	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,992	0,9922	0,9925	**0,9927**	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,994	0,9941	0,9943	**0,9945**	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	**0,9959**	0,996	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	**0,9969**	0,997	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	**0,9977**	0,9978	0,9979	0,9979	0,998	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	**0,9984**	0,9984	0,9985	0,9985	0,9986	0,9986
3	0,9987	0,9987	0,9987	0,9988	**0,9988**	0,9989	0,9989	0,9989	0,999	0,999
3,1	0,999	0,9991	0,9991	0,9991	**0,9992**	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	**0,9994**	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	**0,9996**	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	**0,9997**	0,9997	0,9997	0,9997	0,9997	0,9998

Tab. 4.4: Probabilidades de la normal estándar

Ya estamos en condiciones de ejercitar su uso en sendos problemas. Por ejemplo, el precio promedio de las acciones de una empresa es de $30 y la desviación estándar es $8,20. Suponga que los precios de las acciones están distribuidos normalmente. ¿Cuál es la probabilidad de que el precio de las acciones de una empresa sea por lo menos de $40? ¿De que el precio de las acciones de una empresa no sea mayor a $20? ¿De cuánto deben ser los precios de las acciones de una empresa para que esté entre las 10% mejores?.

Sea X={"precio de las acciones"}. Así, X~ N($30; $8,20). Para responder las preguntas enunciadas, queremos determinar:

a. P(X>$40)

b. P(X<$20)

Y la tercera pregunta, que es la más peculiar, pues se trata de encontrar el cuantil dada el área, que es 0.1 como cola superior, es decir, los precios más altos.

c. $P(X>x_{0.1})=0.1$

Estandaricemos los b=$40 a través del cambio de variable $b_z=(b - \mu) / \sigma$:

$$b_z=(40 - 30) / 8.2 =1.22$$

Así, la probabilidad $P(X>\$40)=P(Z>b_z)=P(Z>1.22)$, que por las propiedades de la normal, es igual $P(X>\$40)=1 - P(Z<1.22)=1–0.8888=0.1112$

Ahora, para a=$20, $a_z=(20 - 30) / 8.2 = -1.22$, por lo que $P(X<\$20) = P(Z < -1.22)$, que debe ser modificada para utilizar la tabla pues ella solo provee probabilidades para los x mayores o iguales a 0.

Entonces $P(Z < -1.22)= P(Z > 1.22)= 1- P(Z < 1.22)=0.1112= P(X<\$20)$.

Finalmente, el valor $x_{0.1}$ tal que $P(X>x_{0.1})=0.1$ es el mismo tal que

$P(X<x_{0.1})=0.9$. Así, miramos el valor más cercano al área 0.9 en la tabla y concluimos que $x_{0.1}=1.28$. Pero como necesitamos el precio, des-estandarizamos con la fórmula

$$c_x = \mu + \sigma * c_z$$

$$\text{Así que } Precio_{10\%} = \mu + \sigma * x_{0.1} = 30+8.2*1.28 = \$40{,}50$$

Para formalizar un objeto propuesto en el ejercico, se definen como **las colas de la normal** a los valores z_p tales que $P(Z>z_p)=p$. También se les conoce como percentiles de las variables aleatorias normales, pues significan que el 100(1-p)% de las veces que se observa una normal estándar el valor es inferior a z_p.

Otra propiedad importante de la normal es que es **la suma de normales es normal**, esto es: si $X \sim N(\mu_x, \sigma_x)$ y $Y \sim N(\mu_x, \sigma_x)$ entonces:

$$X+Y \sim N(\mu_x + \mu_y; \sqrt{\sigma_x^2 + \sigma_y^2})$$

4.4 De lo continuo a lo discreto: el vínculo entre la normal y la binomial.

Recordemos que la distribución binomial discreta aparece en un ex-

perimento que consiste en una serie de N ensayos idénticos e independientes, habiendo para cada prueba tan solo dos resultados posibles, éxito o fracaso. La probabilidad de éxito en una prueba es la misma que en cualquier otra de las pruebas y vale p. La variable aleatoria binomial cuenta el número de éxitos en N pruebas. Cuando el número de ensayos se incrementa considerablemente, la manipulación de la función de probabilidad binomial, sea a mano o sea con una calculadora, reporta un dificulta considerable. En tales casos en que N>>15 y p no tiende a ser 0, la distribución binomial puede aproximarse a una distribución normal. Solo basta considerar $\mu=Np$ y $\sigma=\sqrt{Np(1-p)}$. Es decir, si X se distribuye binomial con parámetros N,p, entonces

$$\mathbf{Z = (X - Np) \div \sqrt{Np(1-p)}}$$

es una variable aleatoria que se distribuye normal estándar. Esto es una consecuencia de un resultado matemático de suma importancia conocido como el teorema central del límite (Apostol, 2001, pág. 689). La relación entre la Poisson y la normal es evidente.

4.5 Comentarios finales

Las variables aleatorias son instrumentos que al principio pueden parecer abstrusos, pero su utilidad en la teoría de probabilidad y en las

estadísticas justifica el empeño en su estudio y tratamiento. Acá se exhibieron tan solo las distribuciones más comunes y sus equivalencias. Sin embargo, conforme se vaya avanzando, aparecerán otros tipos de distribuciones, sobre todo continuas, como la Ji cuadrado, la t de Student y la Fisher. Todas ellas se trabajan con tablas, tal y como la distribución normal. La fuerza de la inferencia estadística reposa en las herramientas y objetos que se han expuesto hasta ahora.

Ejercicio

Sea un variable aleatoria con distribución de probabilidad binomial cuyos parámetros son N=200 y p=0.6. Determine:

1. ¿Cuál es su esperanza y su desviación estándar?
2. ¿Es posible que pueda ser aproximada por la distribución de probabilidad normal?
3. ¿Cuál es la probabilidad de obtener entre 100 y 110 éxitos?
4. ¿Cuál es la probabilidad de sacar 120 o más éxitos?
5. ¿Cuál es la ventaja de usar la normal para aproximar a la binomial?

CAPÍTULO V

Estimación puntual: El muestreo y sus distribuciones

MAWIL
Publicaciones Impresas y Digitales

www.mawil.us

5. Estimación puntual: El muestreo y sus distribuciones

5.1 ¿Qué es la Estadística Inferencial? Estimación e inferencia. Conceptos básicos.

Hasta ahora, cuando hemos explorado el uso de las técnicas asociadas a la estadística, las hemos utilizado para **describir** y clasificar a un conjunto de datos. Contando con las herramientas que la teoría de las probabilidades nos ha brindado, podemos ir más allá, podemos utilizar a las herramientas de la estadística para **inferir** información sobre un conjunto de datos, datos que por lo general de una muestra de individuos de una población. Por eso, en este giro que daremos a la estadística como metodología de tratamiento de la información, el muestreo como variable aleatoria se volverá fundamental y la tarea teórico-práctica será estudiar su distribución, **estimar** sus parámetros. A tal proceder se le denomina **Estadística Inferencial**.

Por estimar entendemos someramente proyectar e inferir. Por supuesto, lo primero es muy amplio y escapa de los objetivos de la estadística. Nos ocuparemos de su segundo sentido, a saber, inferir. Una estimación estadística es un procedimiento mediante el cual construimos qué valor debe tener un parámetro según deducciones que realizamos a partir de los estadísticos, recordando que estos se tratan de que cualquier magni-

tud numérica cuyo valor se pueda determinar a partir de los datos. Entonces, **estimar es sacar conclusiones sobre características poblacionales a partir de resultados muestrales**.

En las ciencias estadísticas, hay dos tipos de estimaciones: la puntual y la que se realiza por intervalo. La segunda es la más común e intuitiva, y será objeto de estudio de capítulos posteriores. En cuanto a **la estimación puntual**, no es menos sencilla, bien al contrario. Comenzaremos con ella pues la estimación por intervalo presupone la estimación puntual.

La estimación puntual se fundamenta en determinar un valor concreto, es decir, un número puntual para el parámetro de una distribución. Tal valor se construye a partir de un estadístico concreto. Como tal **estadístico** sirve para realizar esa estimación, es usual que se le llame **estimador**, y así se le distingue de los estadísticos descriptivos contemplados en capítulos anteriores.

Entendemos por estimador muestral a una regla que indica cómo calcular el valor de una estimación (un parámetro) con base en las mediciones que contiene una muestra. (Mendenhall III, Wackerly, & Scheaffer, 2002, pág. 365). Es como si se tratasen de funciones cuyos argumentos son los valores de una muestra; por ello los estimadores son variables aleatorias, pues las muestras provienen de un proceso aleatorio.

Propiedades de los estimadores

Por lo general, para confiar en el éxito de una estimación, se desea que los estimadores sean

1. **Insesgados**: esto ocurre si el valor esperado del estimador muestral es igual al parámetro poblacional que se estudia.

2. **Eficiente**: que tenga el menor error cuadrático, es decir, que tenga varianza mínima.

3. **Consistente**: es decir, que a medida que la muestra sea más grande, la estimación será mejor. Dicho con formalidad: un estimador muestral es consistente si está más cerca del parámetro poblacional a medida que el tamaño de la muestra aumenta.

4. **Robusto**: si los resultados que proporciona el estimador no se alteran de manera significativa si las hipótesis de partida en los que se basa la estimación no son ciertas o cambian.

5. **Suficiente**: un estimador muestral lo es cuando abrevia toda la información relevante contenida en una muestra, de forma tal que ningún otro estadístico pueda proporcionar información extra sobre el parámetro de la población.

5.2 Teoría elemental del muestreo

Recordemos ahora dos términos fundamentales, a saber, población y muestra. Una población es el conjunto de todos los elementos de interés en una determinada investigación. Una muestra es un subconjunto de una población.

Las características numéricas de una población son variables aleatorias pues no se pueden saber de antemano. Estas tiene distribuciones, conocidas o desconocida, con sus respectivos parámetros, como la media y la desviación estándar.

Así, el objetivo principal de la estadística inferencial es realizar estimaciones y pruebas de hipótesis acerca de los parámetros poblacionales usando aquella información que proporciona una muestra. Es allí donde reside la razón de estudiar qué es, cómo se clasifica y cómo sea realiza el muestro.

En el capítulo pasado hicimos algunos apuntes sobre el muestreo, pero no definimos su teoría. La teoría del muestreo es “el estudio de las relaciones existentes entre una población y las muestras extraídas de ella”. (Schaum, 183). Ella sirve para estudiar si las variaciones entre dos extracciones de elementos de una población dada, son producto del azar o si hay otras razones más profundas que deban ser consideradas significativas.

Para eso existen las pruebas de hipótesis que sirven a la toma de decisiones.

Junto a la la teoría de probabilidades que hemos estudiado hasta ahora, la teoría elemental del muestreo sirve a ser el fundamento racional de la inferencia estadística.

Existen dos tipos de muestreo:

1. Muestreo probabilístico: aleatorio simple, sistemático, por conglomerados y estratificado.
2. Muestreo no probabilístico: a juicio, conveniente y voluntario

Acá el estudio está restringido al muestreo probabilístico. Los criterios para discernir qué subtipo de muestreo probabilístico se debe realizar para obtener muestras representativas de una población es objeto del diseño de experimentos. A continuación describiremos uno de tales tipos de muestreo.

El muestreo aleatorio simple

Para tomar una muestra de una población hay múltiples métodos; uno de los más usados es el muestreo aleatorio simple. Su definición y realización depende del tamaño de la población, es decir, si la población es finita o infinita (considerablemente grande, al menos).

Cuando la población es finita, digamos de tamaño N, el muestreo aleatorio simple de tamaño se realiza considerando que cada elemento de la población tiene la misma probabilidad de ser escogido.

Dicha selección se realiza, por ejemplo, a través del método del bingo, luego de asignar a cada elemento de la población un único número. Otra forma análoga es ordenar a la población y seleccionar utilizando tablas de números aleatorios.

El muestreo aleatorio simple para poblaciones finitas tiene dos formas de realizarse:

1. Con reposición: en este caso un mismo elemento de la población puede tomarse más de una vez.

2. Sin reposición: en este caso, una vez que un elemento de la población es extraído, no puede volverse a tomar.

Como ya hemos comentado al vincular las distribuciones binomial e hipergeométrica, cuando la población es infinita –muy grande–, el muestreo con reposición es equivalente al que se hace sin reposición.

5.3 Distribuciones del muestreo y estimadores de los parámetros poblacionales

Consideremos el siguiente ejemplo. Consideremos la población de números descrita extensivamente en este conjunto: {1, 2, 5, 7, 10}. Su tamaño es N=5. Notemos que su **promedio poblacional** es igual a (1+2+5+7+10)/5=5=μ. Queremos realizar muestras de tamaño n=2, sin orden y sin reposición. De acuerdo a resultados exhibidos en el capítulo anterior, hay 5C2=10 de realizar tal muestreo, listados a continuación:

Muestra #	**Muestra**	**Promedio muestral $\overline{x}$**
1	{1, 2}	**(1+2)/2=1,5**
2	{1, 5}	3
3	{1, 7}	4
4	{1, 10}	5,5
5	{2, 5}	3,5
6	{2, 7}	4,5
7	{2, 10}	6
8	{5, 7}	6
9	{5, 10}	7,5
10	{7, 10}	8,5

Notemos que utilizando al estadístico promedio como función de la muestra, obtuvimos resultados variables, a veces muy lejanos al parámetro poblacional 5, pero la mayoría de las veces no tanto. Estamos pues ante la presencia de una variable aleatoria $\overline{x}$, ya que depende de un muestreo

aleatorio simple. Dicha variable aleatoria posee una distribución muestral. Veamos qué sucede con el promedio de los promedios:

$$(1,5+3+4+5,5+3,5+4,5+6+6+7,5+8,5)/10=5$$

Lo cual es un caso particular de un hecho general que establece que la media de la distribución muestral coincide con la media de la distribución poblacional:

$$E(\overline{x}) = \mu$$

Este hecho, en apariencia trivial, es fundamental para los casos en donde no se conoce a la población y, por lo tanto, se sabe poco de sus parámetros. Formulemos tal hecho con un poco más de formalidad.

Estimador puntual de la media de una población

Sea x1,..., xn una muestra extraída de una población con media desconocida μ y varianza σ^2 (sus parámetros poblacionales). Entonces, definido el estadístico media muestral como $\overline{x}=(\sum xi)/n$ entonces se tiene que:

$E(\overline{x}) = \mu$

$Var(\overline{x}) = \sigma^2/n$ si la población es grande o el muestreo se hizo con reposición

$Var(\overline{x}) = ((N-n)/(N-1))*(\sigma^2/n)$ si la población es pequeña o se hizo el muestro con reposición

Por lo tanto, cuando se desconoce la media poblacional, **la media muestral es un estimador insesgado puntual de la media de la población.**

Ejemplo: para estimar la cantidad media de reclamos por accidentes automovilísticos por fallas mecánicas de un determinado modelo de carro, para estudiar la factibilidad de una descontinuación del modelo, un grupo de protección al usuario, muestreó en los archivos de una gran compañía de seguros y obtuvo las siguientes cantidades que habían sido reclamadas en 8 accidentes: $121.000, $63.000, $11.000, $9.000, $103.000, $66.000, $42.000, $141.000.

El estimador de la media de las cantidades reclamadas por daños causados por los accidentes es:

$$\overline{x} = \$\ 469.500$$

Es decir, que si hay N vehículos en circulación, y el costo de retirar del mercado a todos los vehículos es menor que $N* \overline{x}$, entonces vale la pena, económicamente, hacerlo.

Estimador puntual de la proporción de una población

Un resultado análogo al anterior establece que el estimador puntual de una proporción poblacional de un tipo I (o tiene una característica definida) de individuos:

$$\overline{p}=X/n$$

Donde X es el número total de individuos de la muestra que son del tipo I, y n es el tamaño de la muestra. Así como con la media muestral, la proporción muestral también es un estimador insesgado, es decir, considerando a p como la proporción poblacional:

$$E(\overline{p})=p$$

A su vez, es cierto que la varianza del estimador es $Var(\overline{p})=p(1-p)/n$.

Estimador puntual de la varianza de una población

Supongamos que se ha extraído una muestra x1,...,xn, de tamaño n, de una población con varianza σ^2 desconocida, tiene como estimador insesgado de varianza al estadístico denominada varianza muestral S^2, definido como:

$$S^2=\sum(xi - \overline{x})^2/(n - 1)$$

Veamos un ejemplo. En una muestra de nueve paquetes de fideos producidos por una determinada compañía, se midió el peso total para verificar la calibración de la máquina empaquetadora. Como se resultado se obtuvieron los siguientes datos en gramos:

1011, 1124, 1097, 1108, 1120, 1116, 1113, 1098, 1097

Su media muestral resulta ser:

$$\overline{x}=1098,222$$

Ahora, ayudándonos con una tabla, calcularemos el estadístico S^2:

xi	xi - $\overline{x}$	$(xi - \overline{x})^2$
1011	-87,222	7607,677
1124	25,778	664,505
1097	-1,222	1,493
1108	9,778	95,609
1120	21,778	474,281
1116	17,778	316,057
1113	14,778	218,389
1098	0,222	0,049
1097	-1,222	1,493

Lo que da en suma 9379,553, para concluir que:

$$S^2=1172,444$$

Es decir, que una estimación plausible para desviación estándar es:

$$S=34.24 \text{ gramos}$$

Sin embargo, si quisiéramos saber si eso está dentro del rango tolerable de desviación admisible para la máquina, tendremos que hacer una estimación un poco más profunda, una estimación por intervalos, que será objeto del próximo capítulo.

5.4 Teorema central del límite

No hicimos ninguna mención sobre cómo son las distribuciones muestrales, tan solo se sumarió cómo son sus parámetros y, por lo tanto, cómo estos sirven para estimar las características de las distribuciones de las variables poblacionales. Sin embargo, hay un resultado que merece mención. Ya se comentó sobre él en capítulos anteriores, pero es acá en donde se puede justificar su importancia. Dicho resultado se conoce como Teorema Central del Límite. Y dice así:

Sea x1, x2, x3,...,xn una muestra aleatoria procedente de una población con media μ y varianza σ^2. Si n, el tamaño de la muestra, es suficientemente grande entonces la variable aleatoria que representa la siguiente suma

$$x1+x2+x3+x4+...+xn$$

Se distribuye normal con media μ y varianza σ^2/n. Es decir, sintetizando el

enunciado,

$$\sum \mathbf{xi} \sim \mathbf{N}(\mu, \sigma^2/n) \quad \mathbf{si} \quad \mathbf{n} \rightarrow \infty$$

Al respecto de este Teorema, el polímata inglés apuntó: "Difícilmente conozco algo tan apto para impresionar tanto mi imaginación como el maravilloso orden cósmico que se deriva de la 'ley de frecuencia de error'. Hubiese sido personificada por los griegos, deificada, si la hubiesen conocido. Ella reina con serenidad y con toda modestia, en medio de la confusión más salvaje. Mientras más grande sea la turba, y mayor la aparente anarquía, tanto más perfecto es su dominio. Es la ley suprema de la sinrazón. Cada vez que una muestra grande de elementos caóticos se tengan a la mano y se calculen en un orden de acuerdo su magnitud, una forma insospechada y bellísima de regularidad demuestra que ha estado latente desde hace mucho". (Galton, 1889) El teorema central del límite era conocido en los siglos XVII y XVIII como la ley de frecuencias de errores, por su aplicación a la medición de errores.

5.5 Comentarios finales

Con esta discusión sobre la estadística inferencial, los estimadores y sus propiedades hemos descubierto un nuevo espectro de aplicaciones de las ciencias de los datos y la información. La introducción a la teoría

elemental del muestro nos han puestos en condiciones de comprender un poco más el potencial de la máxima que establece que basta conocer las partes para aproximarse a conocer el todo, orientados por la racionalidad que las matemáticas brindan.

La comprensión de la estimación puntual es un requisito *sine qua non* para continuar explorando otras técnicas de inferencia estadística que son tanto más vastas y certeras en la tarea de obtener conocimiento a partir de medidas restringidas, tomar decisiones en situaciones de incertidumbre, discernir entre verdades probables e improbables y demás objetivos de las ciencias estadísticas.

CAPÍTULO VI

Estimación por intervalo: Medidas de confianza

MAWIL
Publicaciones Impresas
y Digitales

www.mawil.us

6. Estimación por intervalo: Medidas de confianza

6.1 Motivación y definiciones

Cuando se utilizan las estimaciones puntuales no se puede esperar mucho, si tan solo ellas constituyen el objetivo de la investigación, por más que resuelvan problemas de procedimiento, debido a que no se puede espera que el estadístico en cuestión proporcione el valor exacto del parámetro que se desea estimar. Son procedimientos absolutamente necesarios, pues nos aproximan al parámetro poblacional. Pero con ello nos basta pues es muy improbable que el parámetro tome exactamente el valor del estimador.

Como es natural preguntarnos, en qué rango de valores podría caer el parámetro, procuramos estimar por intervalo. Ilustremos tal hecho con un ejemplo: con el objeto de realizar una inversión y medir su riesgo, se desea saber valor de las ventas medias por día que se producen en un establecimiento comercial. Para ello se realiza un muestreo que consiste en escoger al azar las ventas que se realizaron durante 50 días distintos. Tal muestra arroja resultados que indican que las ventas medias fueron por concepto de $600 y pero la desviación muestral es muy alta, pues llega a ser de $300, es decir, es muy probable que un día se vendan $300 y al día siguiente $900. ¿Pero cómo obtener un rango para la media que sea acer-

tado el 95 % de las veces?.

Para ser más concretos, se procura más bien estimar una región entorno al resultado del estimador puntual para el cual tengamos una confianza sólida en el que parámetro esté contenido en dicho intervalo. Así, pues, definimos como **estimador por intervalo** a conjunto de valores acotados con el que se predice que el parámetro poblacional estará contenido en él. Así mismo, el **nivel de confianza** que se da a dicho intervalo es la probabilidad de que el parámetro pertenezca a él. A la herramienta estadística que se utiliza para tal fin se le llama también **intervalos de confianza**.

Para tal estimación se necesitan dos elementos, como en un círculo: un centro y un radio (una distancia al centro). El centro lo aporta un estimador puntual, y el radio debe representar una medida de imprecisión que dependa tanto del muestreo como del nivel de confianza o precisión que se le asigne.

6.2 Intervalo de confianza para la media y la proporción muestral.

Sea x1,..., xn una muestra aleatoria. A continuación presentaremos una tabla donde se sintetizan los intervalos de confianza para la media poblacional a partir de dichos datos:

Distribución de los datos	**Desviación estándar σ**	**Intervalo**	**Nivel de confianza**
Normal o tamaño de la muestra n grande (n>30)	Conocida	$[\overline{x} - z_{\alpha/2} * \sigma/\sqrt{n}, \overline{x} + z_{\alpha/2} * \sigma/\sqrt{n}]$	$100(1-\alpha)\%$
Normal y tamaño de la muestra n grande (n>30)	Desconocido	$[\overline{x} - z_{\alpha/2} * \mathbf{S}/\sqrt{\mathbf{n}}, \overline{x} + z_{\alpha/2} * \mathbf{S}/\sqrt{\mathbf{n}}]$	$100(1-\alpha)\%$

Donde $z_{\alpha/2}$ es tal que $P(Z> z_{\alpha/2}) = \alpha / 2$

Así, para el ejemplo de las ventas medias por día, nuestros datos indican que

$\overline{x}=600$

$S=300$

$1-\alpha = 0.95$ y $n=50$

Por lo que, $\alpha = 0.05$ Hay que determinar entonces al cuantil $z_{0.025}$, esto es:

$$P(Z> z_{\alpha/2})=0.025=1-P(Z< z_{0.025})$$

$$P(Z< z_{0.025})=1-0.025=0.975$$

El cual es, tras revisar la tabla de valores normales adjunta en el apéndice, $z_{0.025}$=1.96. Luego, el intervalo de confianza tiene como límite inferior a

$$Li= \overline{x} - z_{0.025} * S/\sqrt{n}$$

$$Li=600 – 1.96*300/\sqrt{50}=600 – 1.96*300/5.48=600-107.30=\$492.7$$

Mientras que su límite superior es:

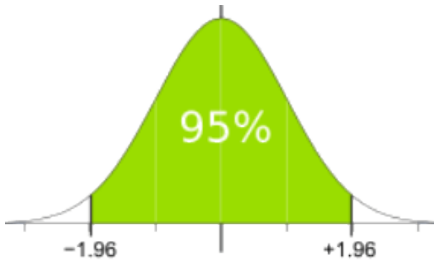
$$Ls= Li= \overline{x} + z_{0.025} * S/\sqrt{n}$$

$$Li=600 + 1.96*300/\sqrt{50}=600 + 1.96*300/5.48=600+107.30=\$707.3$$

Por lo que concluimos que nuestro **intervalo de confianza** para un nivel de confianza del 95% es

[492.7, 707.3]

Es decir, el valor medio de las ventas por día, el 95% de la veces es mayor que 492.7 dólares y menor a 707.3 dólares.



En cuanto a la proporción muestral, tenemos un resultado similar

que se sintetiza en la siguiente fórmula probabilística.

$$\mathbf{P(\ \overline{p} - z_{\alpha/2} * \sqrt{\overline{p}(1-\overline{p})/n} \leq p \leq \overline{p} + z_{\alpha/2} * \sqrt{\overline{p}(1-\overline{p})/n}\ ) = 1 - \alpha}$$

6.3 Intervalo de confianza para desviaciones estándar

Los límites de confianza para la estimación por intervalo de la desviación estándar σ de una población normalmente distribuida, se estiman a partir de una muestra con desviación estándar muestral igual S, y vienen dados por la siguiente fórmula:

$$\mathbf{[S - z_{\alpha/2} * \sigma / \sqrt{2n},\ S + z_{\alpha/2} * \sigma / \sqrt{2n}]}$$

Si la desviación poblacional σ es desconocida, basta con sustituir por S.

Ejemplo: Se calculó que la desviación estándar del tiempo de atención media de un empleado en una muestra de 200 individuos es de 10 minutos. Determinemos el intervalo de confianza para la desviación estándar con un nivel de confianza igual al 99%. En este caso los datos son:

S=10

N=200

1 - α=0.99

Por lo, al mirar la tabla de la normal, buscando $z_{0.005}$, el cual es:

$z_{0.005}=2.58$

Así, el intervalo de confianza es

$$[Li, Ls]=[10 – 2.58 * 10 / \sqrt{400}, 10 + 2.58 * 10 / \sqrt{400}]$$

$$[Li, Ls]=[10 – 1.29, 10 + 1.29]$$

$$[Li, Ls]=[8.71, 11.29]$$

Existen intervalos de confianza para estimar S^2, sin que importe el tamaño de la muestra, pero para ello es necesario utilizar los cuantiles de la distribución Ji cuadrado con n-1 grados de libertad, técnica que es utilizada por las estadísticas avanzadas. En capítulos posteriores se estudiará qué hacer cuando la muestra no es mayor a 30.

6.4 Sobre la selección del tamaño de la muestra

Las nociones adquiridas en la estimación puntual y por intervalos puede, a este punto, ponerse al servicio de la teoría del muestreo y el diseño experimental. El diseño de un experimento consiste en sustancia en elaborar un plan para conseguir cierto volumen de información. Como cualquier objeto de valor, la información se consigue a un determinado

costo, según la forma en cómo se obtenga y los medios que se tengan a disposición. Algunas medidas pueden contener mucho, poco o nada, de información sobre el parámetro de interés para la investigación.

El objetivo de una investigación seria y de altura es balancear los costos de adquisición de data, junto a la disposición de los datos a proveer de información justa y oportuna. Una de las preguntas principales sobre el procedimiento de muestreo que compete al diseño experimental es sobre la cantidad de mediciones que deben incluirse en la muestra. En cuanto a los fines de la estimación, basta saber el nivel de significancia que sus estadísticos deberían tener, el grado de exactitud que el experimentador desea tener en sus inferencias, lo cual puede establecerse de antemano especificando un límite para el error de estimación.

Por ejemplo, para el caso de la estimación de la media de una población, tal procedimiento se realiza fijando un nivel de significancia del 100(1- α)%, un error de estimación tolerable de valor k y despejando el **n** de la relación

$$\mathbf{z_{\alpha/2} * S/\sqrt{n} = k.}$$

Nótese que tal proceso demanda estimar el valor de la desviación estándar pues la variabilidad del estimador $\overline{x}$ depende de la variabilidad de la población de donde se tomará la muestra. Esto puede considerarse como

una de las aplicaciones del Teorema Central de Límite.

6.5 Conclusiones

En este capítulo profundizamos las nociones de estimación estadística, formalizando ya no solo la estimación puntual sino la construcción de intervalos de confianza con niveles de confianza preestablecidos para la estimación de medias, proporciones y desviaciones poblacionales. Vimos además como esta técnica está al servicio del diseño de experimentos al proveer de herramientas para determinar el tamaño de un muestreo dado a priori un error tolerable para estimar un determinado parámetro de la población.

Ejercicio

En una encuesta de mercadeo un individuo puede responder de formas, forma 1 y forma 2, a una determinada pregunta. Si el encuestador desea estimar la probabilidad p de que una persona responda de la forma 1, ¿cuántas personas debe encuestar? Supongamos que el encuestador quedará contento si el error de estimación es menor de 0.04 con una probabilidad igual a 0.90. Debería esperarse que la proporción esté cercana al 60%.

CAPÍTULO VII

Principios de Pruebas de Hipótesis

MAWIL
Publicaciones Impresas
y Digitales

www.mawil.us

7. Principios de Pruebas de Hipótesis

7.1 Decisiones estadísticas. Un resumen histórico

La estadística y la teoría de las probabilidades son disciplinan modernas, existen desde 1654, cuando un problema de apuestas en juegos de azar motivó un intercambio epistolar entre B. Pascal y P. Fermat, que sentó las bases del estudio racional de la contingencia (Apostol, 2001, pág. 571). Así mismo, problemas de salud pública y vacunación, dieron origen a la estadística como herramienta de los nacientes estudios epidemiológicos cuando se plantearon los problemas sobre la efectividad del tratamiento por inoculación del mismo virus en infantes. Desde entonces y hasta nuestros días, un multitud de aplicaciones, que provienen de la genética, la psicología, la ingeniería, la economía y los negocios, han estimulado el crecimiento conjunto de los gemelos probabilidad y estadística, siendo esta última una disciplina que se encarga de deducir información sobre una población a partir de los datos de una muestra seleccionada al azar de dicha población, (Paulos, 1990, pág. 77) y, sobre todo, de proveer herramientas para **tomar decisiones** en base a determinadas aseveraciones más o menos verosímiles, sobre poblaciones estadísticas.

Una de las herramientas más fuertes de la estadística es **la prueba o contraste de hipótesis**. Como técnica, se viene gestando desde el siglo

XVIII, en los trabajos sobre las proporciones de los sexos en los nacimientos humanos, y se perfeccionó a lo largo del siglo XX con los trabajos de K. Pearson, R. A. Fisher y J. Neyman. Para utilizarla, básicamente, se formula "una suposición (que a menudo (...) se llama hipótesis nula), se diseña un experimento, se realiza, y luego se calcula si los resultados del experimento son suficientemente probables, en el supuesto de que la hipótesis sea cierta." (Paulos, 1990, pág. 72). Con tal información, se plantean **las condiciones de una decisión en donde se acepta o se desecha dicha hipótesis nula**.

En tales test estadísticos, el investigador y quien deba tomar la decisión, se enfrentan a cometer dos **tipos de errores**: el error de tipo I que consiste en rechazar la hipótesis mientras que esta es verdadera y el error de tipo II que consiste en aceptar la hipótesis cuando esta es falsa. La distinción entre ambos errores podría parecer cosmética si se piensa en la equivalencia gruesa que hay entre rechazar la verdad y aceptar una mentira, pero estadísticamente y, más aún, en las consecuencias de la decisión que sigue a una prueba de hipótesis, hay una diferencia sustancial. En este capítulo abordaremos con rigor los principios básicos para la teoría de la decisión estadística, conocidos como pruebas de hipótesis. Veremos a su vez cómo las técnicas de estimación puntual e intervalos de confianza sirven a tales fines.

7.2 Pruebas de hipótesis. Definiciones básicas

Las **decisiones estadísticas** se tratan de determinaciones definitivas que se tienen que tomar en torno a las características de una población con base en la información que provee una muestra. Como por ejemplo, si un nuevo fármaco es efectivo en el tratamiento de una patología, si una máquina embotelladora de bebidas gaseosas perdió la calibración, si la implementación de un nuevo software mejora la administración de una empresa, entre otras.

Las decisiones estadísticas se presentan como resoluciones sobre la verdad o falsedad de una hipótesis sobre distribuciones poblacionales y sus parámetros, estas se cuantifican con el planteamiento de una inecuación. Se denominan **hipótesis estadísticas**. Hay dos tipos de hipótesis:

1. **Hipótesis nula**: son formulaciones sobre un parámetro poblacional, usualmente enunciadas para ser rechazadas o nulificadas. Se denota por H_0.

2. **Hipótesis alternativas**: se llaman así a cualquier proposición que difiera de una hipótesis nula dada. Se denotan por H_1.

Como ejemplos podemos considerar un experimento donde se desea descartar dados para casinos, se formula la hipótesis nula que dice que “la

probabilidad de que salga par es igual a p = ½", mientras que la hipótesis alternativa puede ser "la probabilidad de que salga par es igual a p > ½". También puede considerarse que "la probabilidad de que salga par es igual a p = 1/3" es otra hipótesis alternativa.

Cuando **se supone que una hipótesis H_0 dada es verdadera**, pero se encuentra que los datos observados en una muestra aleatoria son notablemente distintos de los resultados que se esperaban de acuerdo a la hipótesis, se dice que las diferencias observadas son significativas y que **se rechaza la hipótesis**. O, mejor dicho, que no se puede apoyar a la hipótesis con confianza de acuerdo a la evidencia obtenida en el muestreo. Si sucede lo contrario –i.e., que la diferencia no es notable-, se dice que **se acepta la hipótesis**. Los procedimientos o reglas que permiten llegar a una u otra conclusión sobre H_0 se denominan pruebas de significancia, reglas de decisión o simplemente **pruebas de hipótesis**.

Tales procedimientos se realizan, básicamente, utilizando la información contenida en una muestra, sintetizada en un estimador dado (que se denominará **estadístico de prueba**) y verificando luego si el valor de éste está contenido o no en una región numérica que contiene los valores por lo que se rechazaría la hipótesis nula –el complemento del intervalo de confianza que define a la hipótesis– que se llama **región de rechazo**, y se

abrevia por RR.

Así pues, **los elementos de una prueba de hipótesis** son:

1. La hipótesis nula H_0.
2. La hipótesis alternativa H_1.
3. La muestra y el estadístico de prueba.
4. Y la región de rechazo (RR).

Cabe destacarse que rechazar la hipótesis nula no es equivalente a aceptar la hipótesis alternativa, pues hay muchas hipótesis alternativas que no todas pueden ser ciertas.

7.3 Tipos de errores y nivel de significancia

Como estamos ante situaciones de incertidumbre, es posible que sucedan yerros, malas decisiones o errores de juicio. Si se rechaza una hipótesis nula cuando esta es cierta, se dice que se trata de un **error de tipo I**. Cuando, por el contrario, se acepta la hipótesis nula debiéndosele rechazar, se dice que se comete un **error de tipo II**.

Para que las pruebas de hipótesis sean exitosas, tiene que realizarse de tal forma que se minimicen ambos errores. Tarea tal que no es tan sencilla, pues por lo general son errores complementarios en el sentido que

mientras se disminuye la probabilidad de cometer error de tipo I entonces se aumenta la probabilidad de cometer el error tipo II. Así, se tiene que escoger cuál es el error más grave y comprometerse a por reducirlo, ya la solamente se pueden reducir ambos errores aumentando el tamaño de la muestra, opción que no siempre es viable, o económica o ni siquiera posible.

Es necesario, pues, definir el nivel de significancia de las pruebas de acuerdo a las riesgos de cometer tales errores:

1. La probabilidad P(cometer error tipo I) = P(rechazar H_0 | H_0 es cierta) = α se denomina nivel de significancia de la prueba o simplemente nivel de la prueba. Y es, por lo general, **la probabilidad máxima que se estaría dispuesto a aceptar**.

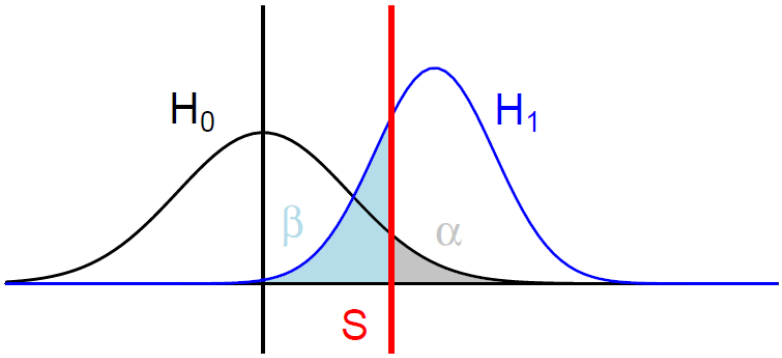
2. La probabilidad P(cometer error tipo II) = P(aceptar H_0 | H_0 es falsa) = β

En la tabla 6.1 se resumen los tipos de errores (así como sus probabilidades de ocurrencia), pues es necesario hacer énfasis en su naturaleza, ya que no son errores equivalentes.

	Se acepta H_0	Se rechaza H_0
H_0 verdadera	Buena decisión 1- α	**Riesgo α**
H_0 falsa	**Riesgo β**	Buena decisión 1- β

Tab. 6.1

En el dibujo 6.2, se supone que la hipótesis alternativa H_1 es la negación completa de la H_0, para ilustrar gráficamente la relación recíproca que existen entre los riesgos de cometer el error de tipo I (área gris a la derecha del segmento rojo S) y el error de tipo 2 (área azul, a la izquierda del valor S). En la práctica se utilizan valores prefijados antes del muestreo de forma tal que los resultados no influyan en los niveles de significancia tolerable. Estos valores son usualmente α=0.05 o **α=0.01**.



Dib. 6.2

Estudiemos el enunciado de un problema básico para ilustrar las situaciones que caracterizan a una prueba de hipótesis: un consorcio comercial está interesado en lanzar una nueva mercancía al público. Luego de hacer una campaña publicitaria, se toma la muestra de 1000 habitantes, de los cuales 25 estimaron útil a la nueva mercancía. Con un nivel de significación del 1%, ¿apoya el estudio la siguiente hipótesis: más del 3% de la población no considera útil a la nueva mercancía?.

Sobre dicho problema podemos identificar que se trata de probar hipótesis sobre proporciones muestrales con n=1000, un nivel de significancia del 0.01 y una hipótesis nulas que se tendría que refutar. Queda por determinar cuál es el estadístico de prueba y la región de rechazo, asunto que estudiaremos a continuación.

7.4 Las pruebas Z para una población normal o para muestras grandes.

Supóngase que se desea probar un conjunto de hipótesis sobre determinado parámetro poblacional θ a partir de una muestra aleatoria x1,..., xn, con n>30. En particular, deseamos probar hipótesis sobre la media $\theta = \mu$ de una población con varianza conocida σ^2, así como también sobre la proporción poblacional $\theta = p$. Ya sabemos que tales prueba requiere un estadístico Z y una región de rechazo RR. Además, tales pruebas se catalogan en dos variedades dependiendo de la hipótesis alternativa que se proponga:

- Las pruebas unilaterales.
- Las pruebas bilaterales.

En la tabla 6.3 se sintetizan en qué consisten operativamente, dichas

pruebas, mientras que el dibujo 6.4 se muestran gráficamente.

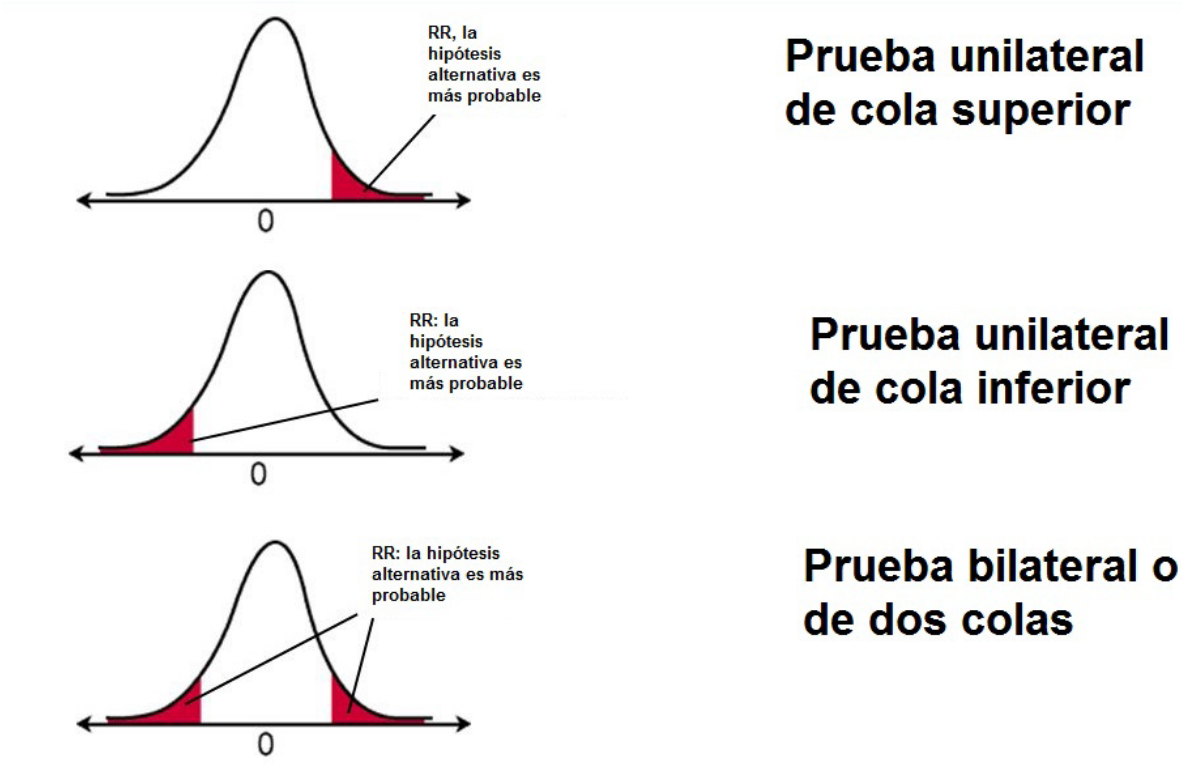
Parámetro θ	**Tipo de prueba**	**Hipótesis nula H_0**	**Hipótesis alternativa H_1**	**Estadístico de prueba Z**	**RR con significancia α**
Media poblacional μ	Bilateral	$\mu = \mu_0$	Dos colas: $\mu \neq \mu_0$	$\sqrt{n}(\bar{x} - \mu_0)/\sigma$	$\{ \lvert Z \rvert > z_{\alpha/2} \}$
Media poblacional μ	Unilateral	$\mu = \mu_0$ ó $\mu \leq \mu_0$	Cola superior: $\mu > \mu_0$	$\sqrt{n}(\bar{x} - \mu_0)/\sigma$	$\{ Z > z_{\alpha} \}$
Media poblacional μ	Unilateral	$\mu = \mu_0$	Cola inferior: $\mu < \mu_0$	$\sqrt{n}(\bar{x} - \mu_0)/\sigma$	$\{ Z < -z_{\alpha} \}$
Proporción poblacional p	Bilateral	$p = p_0$	De dos colas: $p \neq p_0$	$(X-n*p_0)/\sqrt{(n*p_0*q_0)}$	$\{ \lvert Z \rvert > z_{\alpha/2} \}$
Proporción poblacional p	Unilateral	$p = p_0$	Cola superior: $p > p_0$	$(X-n*p_0)/\sqrt{(n*p_0*q_0)}$	$\{ Z > z_{\alpha} \}$

Tab. 6.3

Considerando que:

- z_α es un número real tal que $P(Z > z_\alpha) = \alpha$.
- X mide la cantidad de elementos con determinada propiedad de estudio en proporciones y $q_0 = 1 - p_0$.
- En virtud de la equivalencia entre la binomial y la norma, la prueba de cola inferior para la proporción es análoga a la de la media po-

blacional.



Graf. 6.4

Abordemos, pues, el ejemplo sobre la opinión pública en torno a una nueva mercancía. En ese caso se tiene que:

El tamaño de la muestra es n=1000.

La proporción muestral es $\overline{p} = X/N = 0.025$, es decir el 2.5% de los encuestados opina que la nueva mercancía es útil, luego de haberse realizado la campaña publicitaria.

El nivel de significancia es $\alpha=0.01$.

Y estamos en presencia de una hipótesis nula que pretendemos refutar H_0: $p_0=0.03$.

Se trata de una prueba unilateral porque la hipótesis alternativa es H_1: $p>0.03$.

Realicemos entonces el test, recordando que el estadístico de prueba es $Z=(X-np_0)/\sqrt{(np_0q_0)}$. Esquemáticamente, se procede de acuerdo a los siguientes pasos:

1. Hipótesis nula: $p_0=0.03$
2. Hipótesis alternativa: $p>0.03$
3. Nivel de significancia: $\alpha=0.01$, por lo que $z_\alpha=2.33$
4. El estadístico de prueba $Z=(25-30) / \sqrt{(1000*(0.030)*(0.97))}=-5/5.39=-0.93$
5. La $RR=\{ Z > z_\alpha \}=(2.33, \infty)$

Por lo que se acepta la hipótesis nula, ya que el valor de Z no cae en la región crítica de rechazo, esto es: $-0.93<2.33$. Con tales datos, es plausible concluir que la proporción poblacional de los que consideran al nuevo producto útil es del 3%.

En este caso basta la prueba unilateral, porque no preguntábamos si la proporción era superior al valor, cosa que resultó refutada. En los casos en donde no importe evaluar que sea distinto a un valor, conviene realizar la prueba de dos colas. En el caso de este problema, consideremos $\alpha=0.01/2=0.005$, por lo que $z_{\alpha/2} = 2.57$, y la región de rechazo sería $RR = \{ |Z| > z_{\alpha/2} \} = (-\infty,-2.57) \cup (2.57, \infty)$. Caso en el cual también aceptaríamos la hipótesis nula de que $p_0=0.03$, pues el valor de $|Z| = 0.93$ tampoco está en RR, con una posibilidad del 1% de cometer un error del tipo I. Ahora, nos preguntamos, ¿qué nivel de significancia se hubiese tolerado antes de cambiar de opinión?

7.5 El p-valor

Ante cierta arbitrariedad en la fijación del nivel de la prueba, es conveniente mirar un nuevo objeto. Si en el ejemplo anterior, tras realizar la prueba bilateral, calculamos la probabilidad $P(|x|> Z) = 2*P(x > 0.93)=0.3524$, podemos concluir que el nivel de significancia límite antes de rechazar la hipótesis nula, sería del 35%. Esta medida es de suma importancia, pues, determina qué niveles de confianza son suficientes para rechazar la hipótesis nula. Definamos pues de qué se trata tal número.

Se denomina **p-valor** o valor p al número asociado al estadístico de

prueba que indica el menor nivel de significancia por el cual se debería rechazar la hipótesis nula de acuerdos a los datos observados. En esencia, se trata de un fuerte indicio probabilístico de que la hipótesis nula no es cierta, pues:

- Cuanto más pequeño sea el p-valor, más convincente será la evidencia que se inclina a rechazar la hipótesis nula.
- Contiene más información que los enunciados de las hipótesis.
- Si el valor deseado de significancia α es mayor o igual al p-valor, entonces se debe rechazar la hipótesis nula.
- Si, por el contrario, α es menor que el p-valor, se concluye que no se tiene suficiente confianza para rechazar la hipótesis nula.

Los p-valores de las pruebas Z se calculan de acuerdo al tipo de prueba que se realice:

1. Para prueba unilaterales, p-valor = $P(x \geq Z)$.
2. Para prueba bilaterales, p-valor = $2*P(x \geq | Z |)$.

Notemos que el p-valor para la prueba unilateral del problema de la encuesta sobre la utilidad del nuevo artículo es p-valor=0.1762, es decir, el límite del nivel de significancia es un 18%.

7.6 Pruebas Z para las medias de dos poblaciones normales

Dadas $\overline{x}_1$ y $\overline{x}_2$ dos medias muestrales obtenidas en muestreos de tamaño n_1 y n_2, respectivamente, lo suficientemente grandes, de dos poblaciones distintas que tienen medias y desviaciones estándar μ_1, σ_1 y μ_2, σ_2, respectivamente; es usual que se desee contrastar hipótesis sobre la diferencias de las medias, es decir, que

$$H_0: \mu_1=\mu_2 \text{ ó } \mu_1-\mu_2=0$$

Como se discutió sobre la teoría de la distribución normal, se sabe que la diferencia de medias muestrales está distribuida normalmente y, suponiendo a H_0 así, se tiene que

$$\mu_{\overline{x}1 - \overline{x}2} = 0$$

$$\sigma_{x1 - \overline{x}2} = \sqrt{(\sigma_1^2 /n_1 + \sigma_2^2/n_2)}$$

Por lo que en la tabla 6.5 se sintetizan las pruebas de hipótesis sobre diferencia de medias para poblaciones distintas de varianzas conocidas, que se realizan también con la distribución normal, por lo que son una extensión natural de las pruebas Z que hasta ahora hemos estudiado.

Parámetro θ	**Tipo de prueba**	**H_0**	**H_1**	**Estadístico de prueba Z**	**RR de nivel α**
Diferencia de medias de poblaciones distintas	Bilateral	$\mu_1=\mu_2$	$\mu_1 \neq \mu_2$	$(\overline{x}_1 - \overline{x}_2)$ ÷ $(\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)})$	$\{\|Z\| > z_{\alpha/2}\}$
Diferencia de medias de poblaciones distintas	Unilateral	$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$(\overline{x}_1 - \overline{x}_2)$ ÷ $(\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)})$	$\{Z > z_\alpha\}$

Tab. 6.5

7.7 Comentarios finales

Las pruebas o contrastes de hipótesis son una de las más potentes herramientas de la estadística. Resume un método para resolver la incógnita sobre qué tanto puede inferirse a partir de un conjunto de datos y una aseveración hipotética. Es el arma por excelencia de los estudios de control de calidad, calibración y medidas, pues dota de un esquema de trabajo para determinar si las fluctuaciones de un muestreo se deben al azar o a cambios reales en los procesos que son objeto de estudio. Así mismo, con ellas se pueden estudiar desempeños comparativos entre conjuntos de datos que provengan de poblaciones distintas, y que se estimen diferentes en

media. También se puede utilizar las pruebas de hipótesis para determinar el tamaño de una muestra cuando se desea diseñar un experimento para concluir una hipótesis dada con un determinado nivel de confianza.

En la exposición se hizo énfasis en los tipos de errores que se pueden cometer en dichas pruebas, el nivel de confianza que se tiene en ellos y sobre los índices de medición de tales riesgos que son intrínsecos a los datos utilizados. Todo esto debido a la necesidad de brindar dominio racional a las concesiones y consideraciones que se tienen que hacer al tomar decisiones ante la incertidumbre y el azar.

Se hicieron suposiciones bastante gruesas: el conocimiento a priori de la varianza poblacional y el tamaño de la muestra. También se obviaron las pruebas de hipótesis sobre las desviaciones estándar. Sin embargo, hay pruebas para cuando estos supuestos de normalidad fallan, así como para test donde esté involucrado el estadístico S2, pero requieren conocer dos nuevas distribuciones asociadas a la estimación con muestras pequeñas. Asunto tal que será objeto de estudio en el próximo capítulo.

Ejercicio

Se llevó a cabo un estudio psicológico para comparar la velocidad de reacción de hombres y mujeres ante un determinado estímulo. Se rea-

lizó un muestreo aleatorio independiente de 50 hombres y 50 mujeres. Se obtuvo par de medidas que asocian a los hombres un valor medio de 3.6 y a las mujeres 3.8. Se desea comprobar si estos datos proporcionan suficiente evidencia para inducir la conclusión de que existe diferencia entre las medias reales de velocidad de reacción de hombres y mujeres. Se recomienda utilizar un nivel de significancia $\alpha=0.05$, así como determinar el p-valor y comparar.

CAPÍTULO VIII

Inferencias sobre medias y varianzas

MAWIL
Publicaciones Impresas
y Digitales

www.mawil.us

8. Inferencias sobre medias y varianzas

8.1 Introducción

En todas las técnicas de inferencias hasta ahora estudiadas, sean de estimación o pruebas de hipótesis, hemos supuesto o normalidad de la población o un tamaño N de la muestra superior o igual a 30. En dichos casos las distribuciones muestrales se comportan normalmente. Pero cuando N<30, estamos en presencia de las denominadas muestras pequeñas, en donde la aproximación normal no es buena y empeora conforme N se hace más pequeño.

Entonces se hace necesario estudiar otros tipos de distribuciones probabilísticas que se ajusten mejor a las distribuciones de las muestras pequeñas, aunque sus resultados sean igualmente válidos, e incluso más precisos, para cualquier tipo de muestreo y estimación.

Con ese fin estudiaremos algunas nuevas distribuciones de probabilidad conocidas como la t de Student, la distribución ji cuadrado y la distribución F de Fisher.

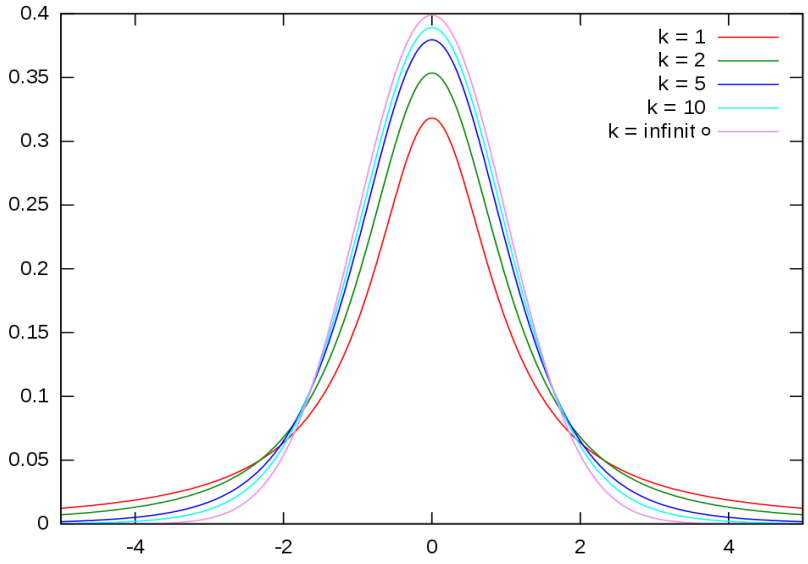
8.2 Las distribuciones t de Student, ji cuadrado y Fisher

En estadística, **la distribución t de Student** surge del problema de estimar la media de una población normalmente distribuida cuando el ta-

maño de la muestra es pequeño. También aparece al realizar la prueba de hipótesis de las diferencias entre dos varianzas muestrales, así como para la construcción del intervalo de confianza para la diferencia entre medias y desviaciones de dos poblaciones cuando se desconoce la varianza de la población y, por lo tanto, debe ser estimada a partir del estadístico S^2 junto a los datos de una muestra.

Es una distribución cuyo rango son todos los números reales. Es **simétrica y acampanada**, aunque ligeramente distinta de la normal pues posee colas más pesadas. Tiene un solo parámetro **n** entero, que caracteriza a su media (aunque esté siempre centrada en el origen) y a su varianza, denominado **grados de libertad**. Conforme los grados de libertad se hacen muy grandes, la distribución t tiende a parecerse a la normal estándar. La forma de su densidad se ilustra en el dibujo 8.1, para cinco parámetros distintos.

Fue descubierta por William Gosset a principios del siglo XX, y publicada bajo un pseudónimo, debido a que provenía de estudios estadísticos en una celosa industria de cerveza en Irlanda.



Dib. 8.1

Debido a fines prácticos, sus valores se presentaron tabulados de acuerdo a su grado de libertad y a sus cuantiles, es decir, en la tabla 8.2 se organizaran los datos tal que por cada fila se registran los grados de libertad (g.l.) de la t, y por columna se indicarán los puntos porcentuales tp de las colas derechas, es decir,

$$\mathbf{tp = P(T > p)}$$

Considerando que T es una variable aleatoria continua que se distribuye t de Student y p es cualquier número real que esté entre 0 y 1, aunque en nuestro caso particular, nos restrinjamos a un conjunto finito de tales valores indicados en la tabla de marras.

g.l. \ tp	0,4	0,3	0,2	0,1	0,05	0,025	0,01	0,005
1	0,3249	0,7265	1,3764	3,0777	6,3138	12,7062	31,8205	63,6567
2	0,2887	0,6172	1,0607	1,8856	2,92	4,3027	6,9646	9,9248
3	0,2767	0,5844	0,9785	1,6377	2,3534	3,1824	4,5407	5,8409
4	0,2707	0,5686	0,941	1,5332	2,1318	2,7764	3,7469	4,6041
5	0,2672	0,5594	0,9195	1,4759	2,015	2,5706	3,3649	4,0321
6	0,2648	0,5534	0,9057	1,4398	1,9432	2,4469	3,1427	3,7074
7	0,2632	0,5491	0,896	1,4149	1,8946	2,3646	2,998	3,4995
8	0,2619	0,5459	0,8889	1,3968	1,8595	2,306	2,8965	3,3554
9	0,261	0,5435	0,8834	1,383	1,8331	2,2622	2,8214	3,2498
10	0,2602	0,5415	0,8791	1,3722	1,8125	2,2281	2,7638	3,1693
11	0,2596	0,5399	0,8755	1,3634	1,7959	2,201	2,7181	3,1058
12	0,259	0,5386	0,8726	1,3562	1,7823	2,1788	2,681	3,0545
13	0,2586	0,5375	0,8702	1,3502	1,7709	2,1604	2,6503	3,0123
14	0,2582	0,5366	0,8681	1,345	1,7613	2,1448	2,6245	2,9768
15	0,2579	0,5357	0,8662	1,3406	1,7531	2,1314	2,6025	2,9467
16	0,2576	0,535	0,8647	1,3368	1,7459	2,1199	2,5835	2,9208
17	0,2573	0,5344	0,8633	1,3334	1,7396	2,1098	2,5669	2,8982
18	0,2571	0,5338	0,862	1,3304	1,7341	2,1009	2,5524	2,8784
19	0,2569	0,5333	0,861	1,3277	1,7291	2,093	2,5395	2,8609
20	0,2567	0,5329	0,86	1,3253	1,7247	2,086	2,528	2,8453
21	0,2566	0,5325	0,8591	1,3232	1,7207	2,0796	2,5176	2,8314
22	0,2564	0,5321	0,8583	1,3212	1,7171	2,0739	2,5083	2,8188
23	0,2563	0,5317	0,8575	1,3195	1,7139	2,0687	2,4999	2,8073
24	0,2562	0,5314	0,8569	1,3178	1,7109	2,0639	2,4922	2,7969
25	0,2561	0,5312	0,8562	1,3163	1,7081	2,0595	2,4851	2,7874
26	0,256	0,5309	0,8557	1,315	1,7056	2,0555	2,4786	2,7787
27	0,2559	0,5306	0,8551	1,3137	1,7033	2,0518	2,4727	2,7707
28	0,2558	0,5304	0,8546	1,3125	1,7011	2,0484	2,4671	2,7633
29	0,2557	0,5302	0,8542	1,3114	1,6991	2,0452	2,462	2,7564
30	0,2556	0,53	0,8538	1,3104	1,6973	2,0423	2,4573	2,75
40	0,255	0,5286	0,8507	1,3031	1,6839	2,0211	2,4233	2,7045
50	0,2547	0,5278	0,8489	1,2987	1,6759	2,0086	2,4033	2,6778
60	0,2545	0,5272	0,8477	1,2958	1,6706	2,0003	2,3901	2,6603
80	0,2542	0,5265	0,8461	1,2922	1,6641	1,9901	2,3739	2,6387
90	0,2541	0,5263	0,8456	1,291	1,662	1,9867	2,3685	2,6316
100	0,254	0,5261	0,8452	1,2901	1,6602	1,984	2,3642	2,6259
110	0,254	0,5259	0,8449	1,2893	1,6588	1,9818	2,3607	2,6213
120	0,2539	0,5258	0,8446	1,2886	1,6577	1,9799	2,3578	2,6174

Tab. 8.2 _ Cuantiles de la t de Student

Así, pues, si tenemos una muestra x1,..., xn; con n<30, el estadístico

muestral definido por:

$$\mathbf{T = \sqrt{n}(\overline{x} - \mu)/S}$$

es un estadístico análogo al Z, pero con la sola diferencia que se distribuye t de Student con n - 1 grados de libertad. Su utilidad es estimable sobremanera y en las siguientes secciones se exhibirá su presencia en la estadística inferencial.

La distribución ji cuadrado, $\chi^{2,k}$, también denominada Chi-Cuadrado o de Pearson, es una distribución de probabilidad continua con **un parámetro $k \geq 2$** que representa los grados de libertad de la variable aleatoria que modela. Esta relacionado con la distribución normal, con la t de Student y con la distribución de Fisher.

La distribución Ji cuadrado tiene **muchas aplicaciones en la estadística inferencial**. Como prueba de independencia, bondad de ajuste y, sobre todo, en **la estimación de varianzas**, asunto en el que reside nuestro interés. También tiene relación en el problema de estimar la media de una población normal, así como con la regresión lineal y el análisis de varianza. Su rango son todos los números reales no negativos, es asimétrica con un ligero sesgo hacia la izquierda. En el dibujo 8.3 se ilustra la forma de su densidad para varios grados de libertad.

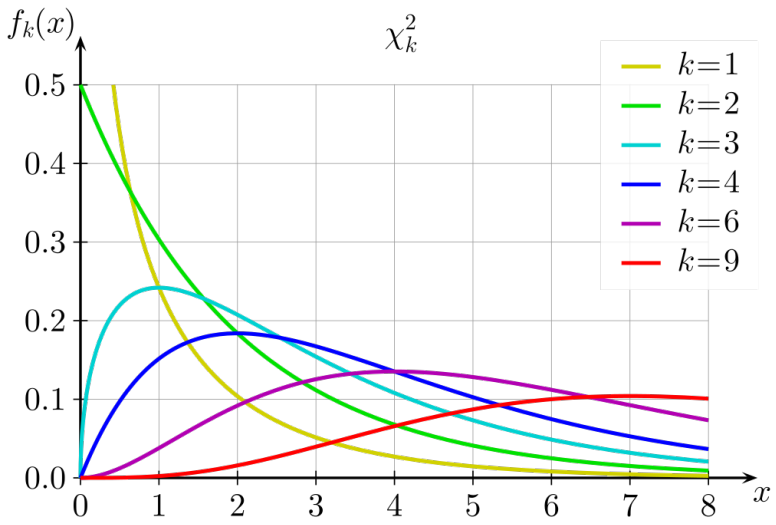
El vínculo con la estadística viene dado por el siguiente hecho: si se

realiza un muestreo de x1,..., xn de una población normal con desviación estándar σ el estadístico muestral definido por:

$$\mathbf{J = (n-1)S^2 / \sigma^2 = ((x1 - \overline{x})^2 + \ldots + (xn - \overline{x})^2) / \sigma^2}$$

Tal se distribuye ji cuadrado con n-1 grados de libertad.

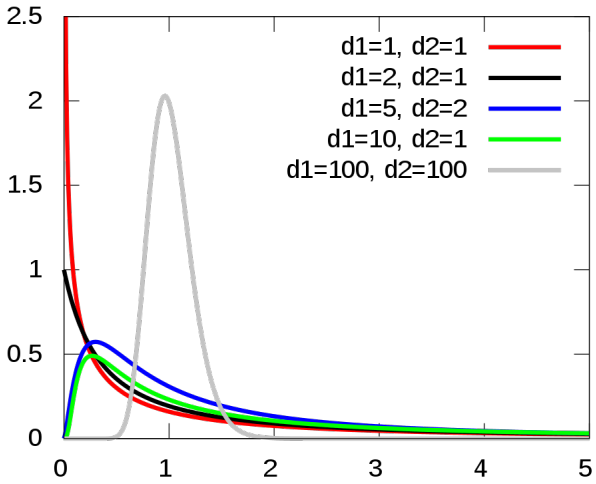
Tal y como se hizo con la t de Student, sus valores se presentara tabulados de acuerdo a su grado de libertad y a sus cuantiles. En las tablas 8.5 y 8.6 se organizaran los datos de tal forma que por cada fila se registran los grados de libertad (g.l.) de la t, y por columna se indicarán los puntos porcentuales Jp de las colas derechas, i.e., **Jp = P(J > p)**



Dib. 8.3

La distribución F de Fisher es una distribución de probabilidad continua. Su rango son los números reales no negativos, posee dos parámetros o grados de libertad mayores a cero. Aparece al considerar el cociente de

dos variables distribuidas Ji cuadrado, por lo que está asociada a los estadísticos que trabajan sobre los cocientes de las varianzas muestrales, que es objeto central en el análisis de varianza. También se utiliza en la Prueba F que se encarga de probar hipótesis sobre igualdad de varianzas. En el dibujo 8.4 se ejemplifican gráficamente algunas de sus densidades. Su tratamiento escapa de los objetivos de este libro, pues es una distribución para técnicas avanzadas.



Dib. 8.4

g.l.	0,995	0,99	0,975	0,95	0,9	0,8	0,7	0,6	0,5
1	0	0,0002	0,001	0,0039	0,0158	0,0642	0,1485	0,275	0,4549
2	0,01	0,0201	0,0506	0,1026	0,2107	0,4463	0,7133	1,0217	1,3863
3	0,0717	0,1148	0,2158	0,3518	0,5844	1,0052	1,4237	1,8692	2,366
4	0,207	0,2971	0,4844	0,7107	1,0636	1,6488	2,1947	2,7528	3,3567
5	0,4117	0,5543	0,8312	1,1455	1,6103	2,3425	2,9999	3,6555	4,3515
6	0,6757	0,8721	1,2373	1,6354	2,2041	3,0701	3,8276	4,5702	5,3481
7	0,9893	1,239	1,6899	2,1673	2,8331	3,8223	4,6713	5,4932	6,3458
8	1,3444	1,6465	2,1797	2,7326	3,4895	4,5936	5,5274	6,4226	7,3441
9	1,7349	2,0879	2,7004	3,3251	4,1682	5,3801	6,3933	7,357	8,3428
10	2,1559	2,5582	3,247	3,9403	4,8652	6,1791	7,2672	8,2955	9,3418
11	2,6032	3,0535	3,8157	4,5748	5,5778	6,9887	8,1479	9,2373	10,341
12	3,0738	3,5706	4,4038	5,226	6,3038	7,8073	9,0343	10,182	11,3403
13	3,565	4,1069	5,0088	5,8919	7,0415	8,6339	9,9257	11,1291	12,3398
14	4,0747	4,6604	5,6287	6,5706	7,7895	9,4673	10,8215	12,0785	13,3393
15	4,6009	5,2293	6,2621	7,2609	8,5468	10,307	11,7212	13,0297	14,3389
16	5,1422	5,8122	6,9077	7,9616	9,3122	11,1521	12,6243	13,9827	15,3385
17	5,6972	6,4078	7,5642	8,6718	10,0852	12,0023	13,5307	14,9373	16,3382
18	6,2648	7,0149	8,2307	9,3905	10,8649	12,857	14,4399	15,8932	17,3379
19	6,844	7,6327	8,9065	10,117	11,6509	13,7158	15,3517	16,8504	18,3377
20	7,4338	8,2604	9,5908	10,8508	12,4426	14,5784	16,2659	17,8088	19,3374
21	8,0337	8,8972	10,2829	11,5913	13,2396	15,4446	17,1823	18,7683	20,3372
22	8,6427	9,5425	10,9823	12,338	14,0415	16,314	18,1007	19,7288	21,337
23	9,2604	10,1957	11,6886	13,0905	14,848	17,1865	19,0211	20,6902	22,3369
24	9,8862	10,8564	12,4012	13,8484	15,6587	18,0618	19,9432	21,6525	23,3367
25	10,5197	11,524	13,1197	14,6114	16,4734	18,9398	20,867	22,6156	24,3366
26	11,1602	12,1981	13,8439	15,3792	17,2919	19,8202	21,7924	23,5794	25,3365
27	11,8076	12,8785	14,5734	16,1514	18,1139	20,703	22,7192	24,544	26,3363
28	12,4613	13,5647	15,3079	16,9279	18,9392	21,588	23,6475	25,5093	27,3362
29	13,1211	14,2565	16,0471	17,7084	19,7677	22,4751	24,577	26,4751	28,3361
30	13,7867	14,9535	16,7908	18,4927	20,5992	23,3641	25,5078	27,4416	29,336
40	20,7065	22,1643	24,433	26,5093	29,0505	32,345	34,8719	37,134	39,3353
50	27,9907	29,7067	32,3574	34,7643	37,6886	41,4492	44,3133	46,8638	49,3349
60	35,5345	37,4849	40,4817	43,188	46,4589	50,6406	53,8091	56,62	59,3347
80	51,1719	53,5401	57,1532	60,3915	64,2778	69,2069	72,9153	76,1879	79,3343
90	59,1963	61,7541	65,6466	69,126	73,2911	78,5584	82,5111	85,9925	89,3342
100	67,3276	70,0649	74,2219	77,9295	82,3581	87,9453	92,1289	95,8078	99,3341
110	75,55	78,4583	82,8671	86,7916	91,471	97,3624	101,7656	105,6323	109,3341
120	83,8516	86,9233	91,5726	95,7046	100,6236	106,8056	111,4186	115,4645	119,334

Tab. 8.5 Cuantiles de la ji cuadrado (parte I)

	Jp	0,4	0,3	0,2	0,1	0,05	0,025	0,01	0,005	0,0025
g.l.										
1		0,7083	1,0742	1,6424	2,7055	3,8415	5,0239	6,6349	7,8794	9,1406
2		1,8326	2,4079	3,2189	4,6052	5,9915	7,3778	9,2103	10,5966	11,9829
3		2,9462	3,6649	4,6416	6,2514	7,8147	9,3484	11,3449	12,8382	14,3203
4		4,0446	4,8784	5,9886	7,7794	9,4877	11,1433	13,2767	14,8603	16,4239
5		5,1319	6,0644	7,2893	9,2364	11,0705	12,8325	15,0863	16,7496	18,3856
6		6,2108	7,2311	8,5581	10,6446	12,5916	14,4494	16,8119	18,5476	20,2494
7		7,2832	8,3834	9,8032	12,017	14,0671	16,0128	18,4753	20,2777	22,0404
8		8,3505	9,5245	11,0301	13,3616	15,5073	17,5345	20,0902	21,955	23,7745
9		9,4136	10,6564	12,2421	14,6837	16,919	19,0228	21,666	23,5894	25,4625
10		10,4732	11,7807	13,442	15,9872	18,307	20,4832	23,2093	25,1882	27,1122
11		11,5298	12,8987	14,6314	17,275	19,6751	21,92	24,725	26,7568	28,7293
12		12,5838	14,0111	15,812	18,5493	21,0261	23,3367	26,217	28,2995	30,3185
13		13,6356	15,1187	16,9848	19,8119	22,362	24,7356	27,6882	29,8195	31,8831
14		14,6853	16,2221	18,1508	21,0641	23,6848	26,1189	29,1412	31,3193	33,426
15		15,7332	17,3217	19,3107	22,3071	24,9958	27,4884	30,5779	32,8013	34,9496
16		16,7795	18,4179	20,4651	23,5418	26,2962	28,8454	31,9999	34,2672	36,4557
17		17,8244	19,511	21,6146	24,769	27,5871	30,191	33,4087	35,7185	37,9461
18		18,8679	20,6014	22,7595	25,9894	28,8693	31,5264	34,8053	37,1565	39,4221
19		19,9102	21,6891	23,9004	27,2036	30,1435	32,8523	36,1909	38,5823	40,885
20		20,9514	22,7745	25,0375	28,412	31,4104	34,1696	37,5662	39,9968	42,3357
21		21,9915	23,8578	26,1711	29,6151	32,6706	35,4789	38,9322	41,4011	43,7751
22		23,0307	24,939	27,3015	30,8133	33,9244	36,7807	40,2894	42,7957	45,2041
23		24,0689	26,0184	28,4288	32,0069	35,1725	38,0756	41,6384	44,1813	46,6235
24		25,1063	27,096	29,5533	33,1962	36,415	39,3641	42,9798	45,5585	48,0337
25		26,143	28,1719	30,6752	34,3816	37,6525	40,6465	44,3141	46,9279	49,4354
26		27,1789	29,2463	31,7946	35,5632	38,8851	41,9232	45,6417	48,2899	50,8291
27		28,2141	30,3193	32,9117	36,7412	40,1133	43,1945	46,9629	49,6449	52,2153
28		29,2486	31,3909	34,0266	37,9159	41,3371	44,4608	48,2782	50,9934	53,5943
29		30,2825	32,4612	35,1394	39,0875	42,557	45,7223	49,5879	52,3356	54,9666
30		31,3159	33,5302	36,2502	40,256	43,773	46,9792	50,8922	53,672	56,3325
40		41,6222	44,1649	47,2685	51,8051	55,7585	59,3417	63,6907	66,766	69,6991
50		51,8916	54,7228	58,1638	63,1671	67,5048	71,4202	76,1539	79,49	82,664
60		62,1348	65,2265	68,9721	74,397	79,0819	83,2977	88,3794	91,9517	95,344
80		82,5663	86,1197	90,4053	96,5782	101,8795	106,6286	112,3288	116,3211	120,1017
90		92,7614	96,5238	101,0537	107,565	113,1453	118,1359	124,1163	128,2989	132,2556
100		102,9459	106,9058	111,6667	118,498	124,3421	129,5612	135,8067	140,1695	144,2928
110		113,1214	117,269	122,2495	129,3851	135,4802	140,9166	147,4143	151,9485	156,2304
120		123,289	127,6159	132,8063	140,2326	146,5674	152,2114	158,9502	163,6482	168,0817

Tab. 8.6 Cuantiles de la ji cuadrado (parte II)

8.2 Intervalos de confianza para muestras pequeñas

Profundizando lo estudiado en el capítulo VI, cuando tenemos una muestra x1,..., xn; con n<30, para una población de datos con varianza desconocida, el intervalo de confianza que mejor acota la media μ, con un nivel de confianza del 100 (1 – α)% es

$$[\ \overline{x} - t_{n-1,\,\alpha/2} * S/\sqrt{n},\ \overline{x} + t_{n-1,\,\alpha/2} * S/\sqrt{n}\ ]$$

Donde n es el tamaño de la muestra,

1-α es el nivel de confianza,

S es el estimador de desviación estándar muestral, definido como la raíz cuadrada de $S^2=\sum(xi - \overline{x})^2/(n - 1)$ y

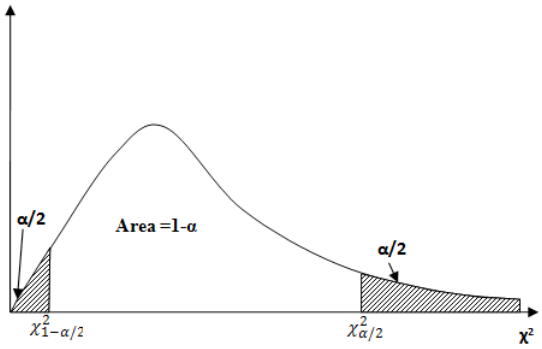
$t_{n-1,\,\alpha/2} = P(T_{n-1} > \alpha/2)$, puntualizando que n-1 son los grados de libertad de la t de Student.

Ahora, para estimar la varianza poblacional σ^2, el índice de variabilidad de una población, con una muestra x1,..., xn, utilizamos el estadístico S^2 que recién hemos citado, para establecer que intervalo de confianza para la varianza poblacional, con un nivel de confianza del 100 (1 – α)% es

$$[\ (n-1)\ S^2/J^2_{\alpha/2}\ ,\ (n-1)\ S^2/\ J^2_{1-\alpha/2}\ ]$$

Donde $J^2_{1-\alpha/2} = P(J< \alpha/2)$ y $J^2_{\alpha/2} = P(J> \alpha/2)$ con J distribuyéndose ji cuadra-

do con g.l. n-1, como se indica en el dibujo 8.7



Dib. 8.7

8.3 Las pruebas T de hipótesis para muestras pequeñas o con varianza desconocida

Supóngase como se hizo en el capítulo anterior, que se desea probar un conjunto de hipótesis sobre determinado parámetro poblacional μ ó σ^2 a partir de una muestra aleatoria x1,..., xn, esta vez con n<30. Consideraremos siempre que la varianza σ^2, no es conocida, pues el otro caso implica sustituir el estadístico S por el valor de σ. Tales pruebas se catalogarán, como antes, en dos variedades y sus RR serán determinadas por la t de Student o la ji cuadrado. En la tabla 8.8 se mostrarán bien discriminadas tales pruebas.

Parámetro	**Tipo**	**H_0**	**H_1**	**Estadístico**	**RR con significancia α**
μ	Bilateral	$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \sqrt{n}(\bar{x} - \mu_0)/S$	$\{ \lvert T \rvert > t_{n-1,\, \alpha/2} \}$
μ	Unilateral	$\mu = \mu_0$ ó $\mu \leq \mu_0$	$\mu > \mu_0$	$T = \sqrt{n}(\bar{x} - \mu_0)/S$	$\{ T > t_{n-1,\, \alpha} \}$
μ	Unilateral	$\mu = \mu_0$	$\mu < \mu_0$	$T = \sqrt{n}(\bar{x} - \mu_0)/S$	$\{ T < - t_{n-1,\, \alpha} \}$
σ^2	Bilateral	$\sigma_0^2 = \sigma^2$	$\sigma_0^2 \neq \sigma^2$	$J = (n-1)S^2 / \sigma_0^2$	$\{ J < J^2_{1-\alpha/2} \}$ ó $\{ J > J^2_{\alpha/2} \}$
σ^2	Bilateral	$\sigma_0^2 = \sigma^2$	$\sigma^2 > \sigma_0^2$	$J = (n-1)S^2 / \sigma_0^2$	$\{ J > J^2_{\alpha} \}$
σ^2	Bilateral	$\sigma_0^2 = \sigma^2$	$\sigma^2 < \sigma_0^2$	$J = (n-1)S^2 / \sigma_0^2$	$\{ J < J^2_{1-\alpha} \}$

Tab.8.8

8.4 Conclusiones

La ampliación del espectro de distribuciones teóricas, tras estudiar la t de Student y la ji cuadrado, contribuyen en gran medida a la estadística inferencial de bajo costo, muestras pequeñas y supuestos débiles sobre la población. Es, quizás, el cénit de la teoría del muestreo y estimación, pues la mayoría de los casos de inferencia en la vida real suceden bajo tales supuestos. Su exposición fue retardada pues demandaba de maduración en los conceptos asociados a las distribuciones teóricas, a los estimadores puntuales y por intervalos, así como de las nociones básicas de las pruebas de hipótesis para poblaciones normales. Se omitieron ejemplos pues casi todos son análogos a los tratados en los capítulos anteriores, así como del manejo de las tablas, que son más sencillas que la de la normal.

Ejercicio

Un experto en eficiencia afirma que introduciendo un nuevo tipo de máquina en un proceso de producción industrial puede disminuir significativamente el tiempo demandado para producir. Dada la magnitud de la inversión necesaria para la adquisición de la maquinaria, así como los costos de mantenimiento de las nuevas máquinas, la gerencia considera que solo reduciendo el tiempo de producción en un 8%, no se podría afron-

tar el gasto de la introducción de la nueva tecnología. Seis experimentos demuestran que el tiempo de producción se disminuyó un 8.4%, con una desviación estándar de 0.32%. Pruebe la hipótesis de que si la nueva maquinaria llegue a implementarse con niveles de significancia iguales a $\alpha = 0.01$ y $\alpha = 0.05$.

CAPÍTULO IX

Introducción al diseño experimental

MAWIL
Publicaciones Impresas
y Digitales

www.mawil.us

9. Introducción al diseño experimental

9.1 Introducción

En capítulos anteriores hemos dotado de rigor a la teoría del muestreo, a través de la extracción de la información contenida en una muestra en función de la estimación de un parámetro poblacional. Vimos, por ejemplo, **cómo la precisión de tal estimación, depende del ancho del intervalo de confianza**, y este por su parte, depende de dos factores: (1) **del tamaño de la muestra** y (2) **la variabilidad de la población**. Esto es, existen dos influencias en la cantidad de información que contiene una muestra, condiciones que dictan la precisión de la inferencia y la significancia de las pruebas de hipótesis que a partir de ella se puedan realizar. Tal conclusión podría parecer trivial, pero es verdadera y guarda en su seno un esquema de análisis que será descrito en este capítulo, con el fin de diseñar experimentos muestrales para mejorar la exactitud de las inferencias estadísticas.

9.2 Nociones básicas de diseño experimental. Variabilidad, elementos y estrategias

El **diseño experimental es una rama de la teoría del muestreo**. Su objeto de estudio son modelos estadísticos elementales cuyo fin es deter-

minar si un conjunto dado de factores influyen en las variables de interés y, de existir tal influencia, mensurarla y catalogarla de acuerdo a su intensidad. Por ejemplo:

1. la comparación de la eficacia de métodos de estudio,
2. la efectividad de una receta de abonos,
3. los efectos de una vacuna o fármaco,
4. la influencia de una política de tarifas escaladas en el uso del servicio de internet o,
5. el rendimiento de una nueva tecnología en un proceso laboral.

La metodología utilizada para diseñar tales modelos se fundamente en la experimentación. Su principal presupuesto dicta que si se repite un experimento, en condiciones similares e independientes, los resultados presentan una variabilidad determinable. Si el experimento se efectúa en un controlado, el error experimental será pequeño y habrá poca variabilidad en los resultados. Si por el contrario, se experimenta en circunstancias industriales, la variabilidad será, por lo general, alta en los más de las unidades experimentales.

El fin del diseño experimental es **estudiar si cuando se combinan determinados factores se producen cambios significativos en un deter-**

minado proceso. Para ellos, se debe definen las noción de un tratamiento que puede aplicarse o no. Así, cuando la variabilidad experimental es grande, se detectará la influencia del uso de un tratamiento cuando este procure grandes cambios relativos al error intrínseco a la observación.

Cuando se realiza un experimento, es de interés determinar:

- la definición de las variables de interés.
- la parametrización, es decir, fijar las principales causas de cambios en las variables.
- la delimitación de las condiciones con las que se consiguen valores extremos en las variables.
- la modulación: comparar las respuestas para diferentes niveles en los parámetros controlables, y
- el modelado: establecer un modelo estadístico que permita predecir comportamientos futuros.

Para el éxito de un experimento es fundamental el análisis de los resultados que se obtienen. Cabe destacarse que pocas veces es posible utilizar estos métodos con datos históricos o que provengan de experimentos registrados sin el detalle de su diseño original, pues los datos pueden ser inconsistentes, se pueden encontrar variables con una alta correlación

y la cantidad de variables controlables no es fácilmente determinable, o su rango de acción es muy limitado.

Para diseñar experimentos es necesario distinguir los tipos de variabilidad de acuerdo a sus orígenes, como se expondrá en la tabla 9.1.

Tipo de variabilidad	**Origen sistemático**	**Origen planificado**	**Origen aleatorio**
Controlable	Sí	Sí	No
No controlable	Sí	No	No siempre
Típica	No	No	Sí
Exploratoria	No	Sí	A veces

Tab. 9.1

La variabilidad controlable viene originada por la posible dispersión de los resultados debido a diferencias introducidas sistemáticamente entre las distintas condiciones experimentales dispuestas expresamente en el modelo experimentador. Se trata del tipo de variabilidad que se intenta estudiar en el diseño. Es deseable que haya tal variabilidad y que sea identificada y cuantificada por el modelo.

La variabilidad no controlable produce una variación sistemática en los resultados, pero es debida a causas desconocidas o, en todo caso, no planificadas. Es decir, los resultados están siendo sesgados de forma ordenada y perceptible por causas desconocidas. La presencia de esta va-

riabilidad supone un error en el modelo, variaciones ambientales no previstas o fallos en la implementación o ajuste del modelo al experimento, es causa usual de conclusiones erróneas. Tal variabilidad se puede evitar mediante un par de técnicas que serán descritas luego. Esta variabilidad no es deseable.

La variabilidad típica es básicamente originada por ruido aleatorio. Estos incluyen a los errores de medida. Es impredecible, inherente a todo proceso experimental y, por lo tanto, inevitable. Por esta variabilidad se explica que tras tomar medidas repetidas de un mismo objeto en una misma circunstancia, ocurra que, en muchos casos, la segunda medida no sea igual a la primera y que, por ello, no se pueda predecir sin error el valor de una tercera medida. Aunque esta variabilidad sea inevitable, es tolerable, pues se puede estimar su comportamiento para poder hacer predicciones.

La variabilidad exploratoria ocurre cuando se desea jugar creativamente con un experimento para medir su robustez o para procurar resultados que se escapen de la sistematización del experimento. Se planifican pero, en principio, no se sabe de dónde provienen. Es útil y deseable solo en la medida en que se considere necesario un abordaje ladino o con técnicas de pensamiento lateral.

Para diseñar un experimento y modelarlo de acuerdo a lo antes ex-

puesto, es necesario contar con un conjunto de definiciones importantes.

Unidades experimentales: son los objetos sobre los cuales se llevan a cabo las medidas del experimento.

Factores: Son el conjunto de variables controladas por el experimentador. Al grado o nivel de intensidad de cada factor como subcategoría se le denomina **nivel del factor**.

Téngase en cuenta que en la manipulación matemática de los modelos de diseño de experimento, los factores cuantitativos son tratados como cualitativos y sus niveles son elegidos se codifican. Por lo general, un factor no tiene más de cuatro niveles.

Tratamiento: es una combinación específica de niveles de factores, es una ecualización de factores.

Observación: es una medida en las condiciones determinadas por alguno de los tratamientos.

Diseño completamente aleatorizado: es una división aleatoria de n unidades experimentales relativamente homogéneas en k grupos, en donde cada grupo recibe un único tratamiento distinto a los otros.

Experimento factorial: es el diseño de experimentos que se formula en términos de todos los conceptos anteriores, y donde existen además

observaciones de todos los posibles tratamientos.

Repetición: las repeticiones son el número de veces que un tratamiento se aplica al mismo número de unidades experimentales.

Bloque: es un conjunto de unidades experimentales de los más homogéneo posible que reciben el mismo tratamiento

Un diseño de bloques aleatorizado consta de b bloques con k unidades experimentales cada uno. Los tratamientos son asignados de forma aleatoria todas la unidades con la restricción de que cada tratamiento aparezca una vez en cada bloque.

En general, al planificar un experimento hay tres principios básicos para seguir: la **aleatorización**, el **bloqueo** y la **factorización** del diseño. Aleatorizar y bloquear son estrategias para tratar a las unidades experimentales sin preocuparse por cual tratamientos considerar. Mientras que la factorización del diseño define una estrategia para elegir los tratamientos sin considerar como asignarlos a las unidades experimentales.

Aleatorizar significa no prestarle atención a los factores no controlables en el diseño. Bloquear es partir a las unidades experimentales en bloques de modo que las observaciones realizadas en cada bloque se realicen bajo condiciones experimentales lo más parecidas posibles y factorizar consiste en cruzar los niveles de todos los factores en todas las combina-

ciones posibles.

Consideremos el siguiente ejemplo: se desean investigar las diferencias en la producción de dos máquinas, cada una de tales debe ser manejada por un operador. La variable respuesta es el rendimiento diario de una máquina, el factor es el tipo de máquina que tiene dos niveles de intensidad, más otro factor de ruido que es el operador designado. En el diseño del experimento, para realizar el estudio se pueden utilizar dos estrategias para controlar el factor del operador que maneja la máquina. O aleatorizar, se seleccionando al azar dos grupos de operarios y asignado al azar cada grupo de operadores a cada una de las dos máquinas, para luego evaluar la producción de las mismas. O bloquear, introduciendo el factor bloque operador y eligiendo un único grupo de operadores para que todos ellos utilicen las dos máquinas.

9.3 Comentarios finales

El objetivo de este capítulo fue exponer brevemente algunos aspectos básicos del diseño experimental, sus conceptos metodológicos y estrategias de diseño. Definimos la variabilidad de un experimento de acuerdo al origen de las modificaciones en las respuestas. Indicamos la existencia de diseños de bloques y completamente aleatorizados. El diseño experimen-

tal es un área muy fructífera de la estadística y posee modelos sencillos y útiles. Sin embargo, su potencial inferencial está en el análisis de varianza, que es la técnica donde se exhibe una manera de medir la cantidad de información que posee un diseño experimental dado, así como determinar el origen de los sesgos experimentales y demás fuentes de error.

Ejercicio

Un publicista cuenta con dos conceptos para campañas publicitarias y tres medios para dirigir dichas campañas: prensa, televisión y redes sociales. ¿Cuántos tratamientos (combinación factor-nivel) hay en este experimento? Describa cuidadosamente dos de estos tratamientos. Cada experimento emplea una combinación concepto-medio: ¿cómo aleatorizaría el orden de los experimentos?.

CAPÍTULO X
El análisis de varianza

MAWIL
Publicaciones Impresas
y Digitales

www.mawil.us

10. El análisis de varianza

10.1 Introducción

La industrialización de la producción de los medios de vida planteó sendos retos a las ciencias estadísticas. El control de calidad, la calibración de la maquinaria, la proyección de la eficiencia, el rendimiento de los procesos, fueron algunos de los problemas que se redimensionaron al intensificarse el trabajo en serie y la extensión de las facultades del hombre económico con herramientas especializadas. Determinar, por ejemplo, detectar errores en el ensamblaje de un gran volumen de electrodomésticos, así como **determinar dónde estuvo el fallo entre los múltiples factores que están involucrado en dicho proceso**, es una situación donde la estadística debía incursionar, dada las dimensiones del problema y su origen. Es así como surgió la necesidad diseñar experimentos muestrales y **estudiar el comportamiento medio y su varianza**.

Existe un método para abordar tales problemas estadísticos donde están involucrados varios parámetros al mismo tiempo, por ejemplo, el estado de un conjunto de medias de varias poblaciones, muestras, tratamientos o bloques que, a diferencia de los métodos ya estudiados en la estimación por intervalo, no posean la misma varianza. A dicho método se le conoce como **Análisis de Varianza**, **ANDEVA** o ANOVA, por sus siglas

en inglés, y se le debe al científico R. Fisher. En este capítulo estudiaremos los estadísticos asociados a dicho análisis para modelos que tengan un factor de variabilidad. No nos debe sorprender el vínculo conceptual con los diseños experimentales, como ya se había adelantado en discusiones anteriores.

10.2 Análisis de varianza para modelos unifactoriales

Se denomina **experimento de un factor** a un conjunto de observaciones distribuidas en un arreglo matricial, donde un total de k filas representan grupos (o tratamientos) y las n columnas representan el número de medidas observadas por cada grupo. Usualmente se asume se dice que el experimento de un factor tiene n réplicas por cada k-ésimo tratamiento. En la tabla 10.1 se ilustra un experimento de este tipo, con sus respectivos promedios por tratamiento.

Tratamiento 1	x11, x12, x13,..., x1n	$\overline{x}1\bullet$
Tratamiento 2	x21, x22, x23,... x2n	$\overline{x}2\bullet$
...	...	$\overline{x}3\bullet$
Tratamiento k	xk1, xk2, xk3,... xkn	$\overline{x}k\bullet$

Tab. 10.1

Es necesario precisar una notación extra para los promedios por fila o tratamiento:

$$\overline{xj}\bullet = (xj1 + xj2 + xj3 + ... + xjn)/n$$

Y, el promedio total de todas las observaciones involucradas sería,

$$\overline{x} = (x1\bullet + x2\bullet + x3\bullet + ... + xk\bullet)/k = (\sum \sum xij) / kn$$

Un modelo matemático para un experimento de un factor es el siguiente. Consideremos que las observaciones de la fila i se desvían de la media poblacional μ por **un solo factor αi** (asociado al tratamiento) más un error aleatorio, es decir, $xi1 = \mu + \alpha i + ei1$, $xi2 = \mu + \alpha i + ei2$,..., $xin = \mu + \alpha i + ein$, o más sucintamente:

$$xij = \mu + \alpha i + eij$$

Con la consideración que dicta que $\alpha 1 + \alpha 2 + \alpha 3 + ... + \alpha k = 0$, pues $\overline{x} = \mu$ dado que los errores eij se distribuyen normales como media 0 y varianza $\sigma 2$, por los que se concluiría que las observaciones también se distribuyen normales con media μ y varianza $\sigma 2$. De hecho, la hipótesis nula para este modelo es que no haya diferencia entre los tratamientos, esto es: H_0: $\alpha 1 = \alpha 2 = \alpha 3 = ... = \alpha k = 0$.

En cuanto a la variación total de todo el experimento, tenemos que es igual a la doble sumatoria:

$$V = \sum \sum (xij - \overline{x})^2$$

Se conoce una descomposición de tal variación dada por la siguiente

ecuación:

$$V = Vd + Ve$$

Donde Vd es la variación dentro de cada tratamiento, es decir:

$$Vd = \sum \sum (xij - \overline{xi}\bullet)^2$$

Y Ve es la variación entre los tratamientos, puesto que se

$$Ve = \sum \sum (\overline{xi}\bullet - \overline{x})^2 = k \sum (\overline{xi}\bullet - \overline{x})^2$$

Es usual que se escriban también como

V = SS total

Vd = SS int = SSE

Ve = SS trat = SST

Es sabido que las esperanzas de cada componente de la variación total vienen dadas por:

$E(Vd)=k(n-1)\sigma^2$

Mientras que sí H_0 es cierta,

$E(Ve)=(k - 1)\sigma^2$

$E(V)=(kn - 1)\sigma^2$

Por lo que se tiene que los estimadores para la varianza dentro de los tratamientos es

$$Sd^2 = Vd/k(n-1)$$

Para la varianza entre los tratamientos es:

$$Se^2 = Ve/(k-1)$$

Y para la varianza total:

$$S^2 = V / (kn-1)$$

Suponiendo igualmente que H_0 es cierta, ambos estadísticos tienen distribución ji cuadrado, el Sd^2 con k(n-1) grados de libertad, Sd^2 con k-1 grados de libertad y S^2 con kn -1 grados de libertad. A partir de todo lo anterior, se tiene sintéticamente la siguiente tabla ANDEVA

Variación	**Grados de libertad (g.l.)**	**Cuadrado medio**
Dentro de tratamientos Vd	k(n-1)	Sd^2
Entre tratamientos Ve	k-1	Se^2
Total V = Vd + Ve	kn – 1	S^2

Tab. 10.2

Construyamos como ejemplo una tabla ANDEVA para el siguiente experimento de un solo factor, entendiendo a los resultados como medidas de desempeño productivo de un equipo de trabajo con k=3 diferentes métodos laborales y n=4 réplicas.

Método A	48, 49, 50, 49
Método B	47, 49, 48, 48
Método C	49, 51, 50, 50

Los promedios por método sería $\overline{x}1\bullet = 49$, $\overline{x}2\bullet = 48$ y $\overline{x}3\bullet = 50$. Así, la media total sería igual a $\overline{x} = 49$.

Ahora, en cuanto a las variaciones, la total sería V=14. Aquella entre tratamientos Ve=8 y la variación dentro de tratamientos es Vd=V-Ve=6. Por lo que su tabla ANDEVA no es más que:

Variación	g.l.	Cuadrado medio
Vd=6	9	$Sd^2=2/3$
Ve=8	2	$Se^2=4$
V=14	11	$S^2=14/11$

10.3 Comentarios finales.

Con la tabla ANDEVA se pueden realizar pruebas de hipótesis e inferencias sobre las diferencias entre los tratamientos, modelando así un diseño experimental de un factor. Los experimentos de dos factores no son más difíciles, básicamente modelan datos organizados como la siguiente tabla:

	Bloque 1	Bloque 2	Bloque 3	...	Bloque n
Tratamiento 1	X11	X12	X13	...	X1n
Tratamiento 2	X21	X22	X23	...	X2n
...	...	...	...	...	...
Tratamiento k	Xk1	Xk2	Xk3	...	Xkn

Mientras que obedecen a un modelo matemático parecido al unifactorial:

$$\mathbf{Xij = \mu + \alpha i + \beta j + Eij}$$

Con las mismas condiciones $\sum \alpha i = 0$ y $\sum \beta j = 0$ y los errores Eij normalmente distribuido con media 0 y varianza σ^2. Sin embargo la construcción de su tabla ANDEVA excede los límites de una aproximación estadística básica, así como profundiza en la prueba Fisher que compara los cocientes de las varianzas entre y dentro de los tratamientos.

CAPÍTULO XI

Algunos fundamentos de regresión

MAWIL
Publicaciones Impresas
y Digitales

www.mawil.us

11. Algunos fundamentos de regresión

11.1 Introducción. La regresión lineal

Hemos acumulado hasta ahora múltiples situaciones que admiten el modelado de datos pareados. Por ejemplo, en el diseño de experimentos podría vincularse la pareja tratamiento-observación como modelo de un estímulo al que le corresponde una única respuesta (E, R). Así, podría vincularse infinidad de datos pareados: inversión versus retorno, precios versus tiempo, ingresos versus gastos, etcétera. En general, se puede abordar para una relación entre dos variables vinculadas por la relación estocástica

$$y = f(x,e)$$

Donde X es la variable independiente o estimulo, Y es la variable dependiente o la respuesta y e sería un error aleatorio. A tal relación se le denomina **modelo de regresión** Dependiendo de la naturaleza de la función F, se obtendrían distintos tipos de modelos de regresión. Nosotros nos enfocaremos en el caso en que

$$y = \alpha + \beta x + e$$

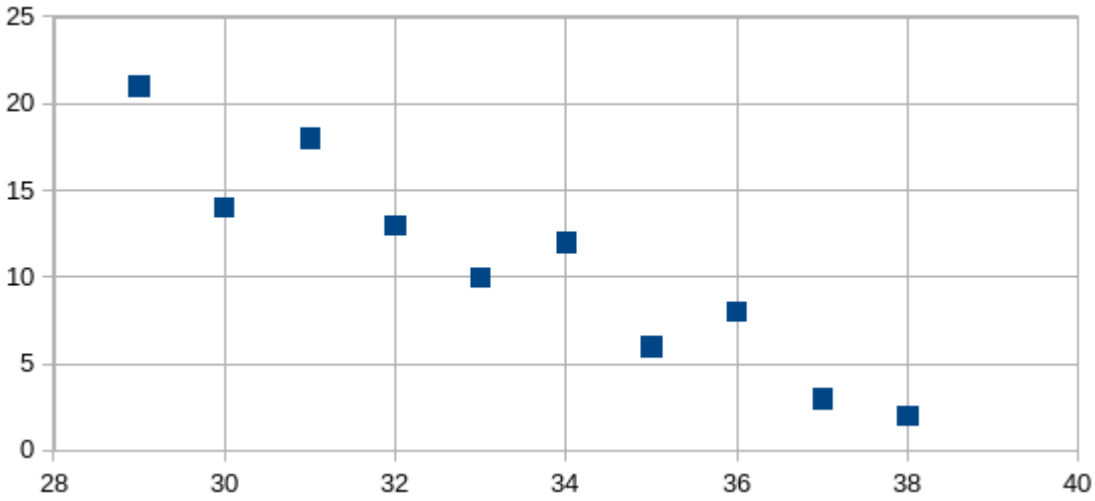
Donde α sería el parámetro de localización, β el parámetro de inclinación y **e** un error aleatorio con media cero, de tal forma que si tomamos la esperanza de la variable aleatoria Y, se obtendría una relación lineal

entre x y y, a saber:

$$\mathbf{E(y) = \alpha + \beta x}$$

El valor de **los parámetros** α y β sería determinado por la información de un conjunto de n datos obtenidos experimentalmente:

(xi, yi)



Dib. 11.1

En el dibujo 11.1 se muestra una graficación del conjunto de datos contenido en la tabla 11.2 del índice de precios ordenados de un conjunto de mercancías versus las ventas por cada precio.

Precios	28	29	30	31	32	33	34	35	36	37	38
Ventas	24	21	14	18	13	10	12	6	8	3	2

Tab. 11.2

A dicha gráfica se le denomina **diagrama de dispersión**. En el eje horizontal se representaron los precios mientras que en el vertical se graficaron las ventas. En ella se evidencia una nube de puntos no colineales con una tendencia al decrecimiento. La forma de dicho decrecimiento es el objeto de la regresión, esto es, se desea estimar a partir de a aleatoriedad de tal fenómeno comercial, una medida de su ubicación y tendencia suponiendo, por la forma de la dispersión, que hay una relación lineal alterada por un **ruido normal**.

Supongamos que dado un conjunto de puntos datos dados por n puntos (xi,yi), al sustituirlos en el modelo, se tendría

$$yi = \alpha + \beta xi + ei$$

Un enfoque matemático para abordar la estimación de los parámetros de regresión consiste en **minimizar la suma de todos los errores al cuadrado**

$$\sum ei^2 = \sum (yi - \alpha - \beta * xi)^2$$

La razón de que sea al cuadrado es que los errores pueden ser positivos y negativos y, por lo tanto, podrían compensarse, mientras que elevados al cuadrado todos serían positivos.

Los **estimadores para dichos parámetros de regresión** serían

$$\beta e = (\sum(xi - \overline{x})*(yi - \overline{y})/(\sum(xi - \overline{x})^2)$$

$$\alpha e = \overline{y} - \beta e * \overline{x}$$

De forma tal que la recta $y = \alpha e + \beta e * x$ se denomina recta de **regresión estimada por mínimos cuadrados**.

Si convenimos en la notación:

$Sxy=\sum(xi- \overline{x})*(yi - \overline{y})$

$Sxx=\sum(xi- \overline{x})^2$

$Syy=\sum(yi - \overline{y})^2$

Se tendría que los estimadores son:

$\beta e = Sxy/Sxx$

$\alpha e = \overline{y} - \beta e * \overline{x}$

Por ejemplo, consideremos la siguiente tabla que vincula las variables x y y:

x	y
70	155
63	150
72	180
60	135
66	156
70	168
74	178
65	160
62	132
67	145
65	139
68	152

Realizando un simple cálculo, los promedios serían $\overline{x}$=66.8 y $\overline{y}$=154.2. Ampliemos entonces dicha tabla para realizar los cálculos que se necesitan para determinar Sxy, Sxx y Syy.

x	y	$x - \overline{x}$	$y - \overline{y}$	$(x - \overline{x})^2$	$(y - \overline{y})^2$	$(x - \overline{x})*(y - \overline{y})$
70	155	3,2	0,8	10,24	0,64	2,56
63	150	-3,8	-4,2	14,44	17,64	15,96
72	180	5,2	25,8	27,04	665,64	134,16
60	135	-6,8	-19,2	46,24	368,64	130,56
66	156	-0,8	1,8	0,64	3,24	-1,44
70	168	3,2	13,8	10,24	190,44	44,16

74	178	7,2	23,8	51,84	566,44	171,36
65	160	-1,8	5,8	3,24	33,64	-10,44
62	132	-4,8	-22,2	23,04	492,84	106,56
67	145	0,2	-9,2	0,04	84,64	-1,84
65	139	-1,8	-15,2	3,24	231,04	27,36
68	152	1,2	-2,2	1,44	4,84	-2,64

Así, sumando por columnas, se tiene que:

Sxy = 616.32

Sxx = 161.98

Syy = 2659.68

Por lo que

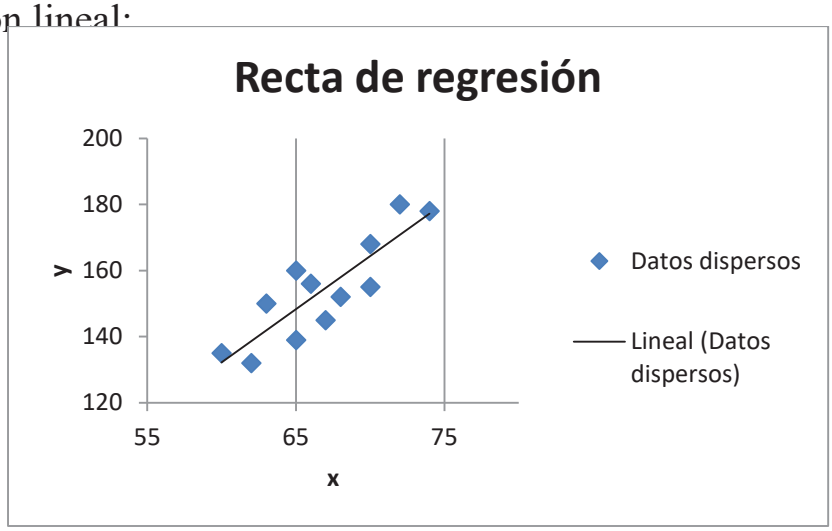
$\beta e = Sxy/Sxx = 3.22$

$\alpha e = \overline{y} - \beta e * \overline{x} = 154.2 - 3.22*66.8 = -60.9$

Para concluir, finalmente, que su recta de regresión es:

$y = 3.22*X - 60.9$

A continuación se puede ver gráficamente a los datos y el resultado de la regresión lineal:



11.2 El error como variable aleatoria. Estimación de la varianza y prueba de hipótesis sobre el βe = 0

Se construyó un modelo estadístico para explicar el vínculo entre dos variables a través de la ecuación lineal con ruido **y = α + βx + e**, con alfa y beta son parámetros a estimar a partir de un conjunto de datos. Ahora es necesario hacer suposiciones extra: como asumir que el error e está distribuido normal con media 0 y varianza σ^2, es decir, que

$$E(y) = \alpha + \beta x \; ; \; Var(y) = \sigma^2$$

También hay que asumir que las observaciones (xi, yi) son independientes por lo que la variable aleatoria

$$(y - E(y))/ \sigma$$

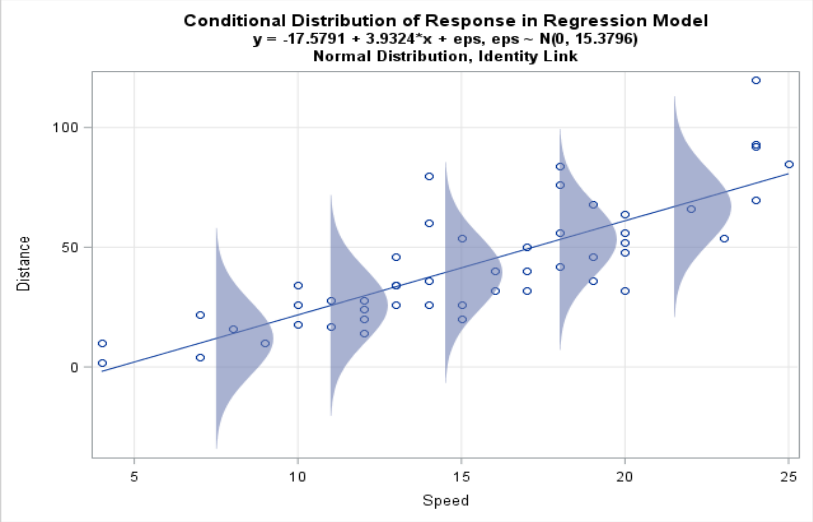
Representado en el dibujo 11.3, se distribuye normal y, de acuerdo a las propiedades de la distribución ji cuadrado

$$\mathbf{(\sum (y - \alpha e - \beta e * x)^2) / \sigma^2}$$

La suma de los cuadrados de los residuos se distribuye ji cuadrado con n-2 grados de libertad pues se estimaron a los dos parámetros αe y βe con un conjunto de n puntos. Por lo que el estadístico definido por:

$$SSR = (\sum (yi - \alpha e - \beta e * xi)^2)$$

Sirve para estimar la varianza considerando la igualdad **SSR / (n-2) = σ^2**



Dib. 11.3

Para fines de cómputo, es útil tener en cuenta que SSR = (Sxx*Syy - Sxy^2)/Sxx. Si abordamos el ejemplo anterior para estimar la varianza de los 12 datos, tendríamos que:

SSR = (Sxx*Syy - Sxy^2)/Sxx = (161.98*2659.68 - 616.32*616.32) / 161.98

SSR = 677.9

Así, un estimado de la varianza, del modelo de regresión lineal es σ^2 = SSR/n-2 = 677.9/ 10 = 6.779, por lo que σ = 2.6 es la desviación estándar.

Un uso importante del modelo de regresión lineal $y = \alpha + \beta*x + e$ consiste en estimar si la variable y depende de x o no. Es decir, queremos refutar la hipótesis $\mathbf{H_0: \beta = 0}$. Proponemos además la hipótesis alternativa $\mathbf{H_1: \beta \neq 0}$.

Para ese fin, se propone el estadístico de prueba: $\mathbf{TS = \beta e*\sqrt{[(n-2)Sxx/SSR}}$

Más la región crítica $\mathbf{RR : |TS| \geq t_{n-2,\ \delta/2}}$

Donde δ es el nivel de significancia de la prueba.

Y el p valor es igual a $\mathbf{p\text{-}valor = P(|T_{n-2}| > \delta)}$

11.3 Comentarios finales

La regresión es una técnica profunda y generosa. Permite no solamente describir la relación entre datos pareados y series temporales, sino también predecir comportamientos en sistema de estímulo-respuesta. Además cuenta con herramientas que miden la susceptibilidad del modelo a adaptarse a los datos, como lo es el coeficiente de correlación

$$r = Sxy/\sqrt{Sxx*Sxy}$$

que, siendo un valor que está entre -1 y 1, indica el signo el sentido de la

relación. Si r = 1, existe una correlación positiva lineal perfecta. El coeficiente indica una dependencia total entre en relación directa. Si $0 < r < 1$, existe una correlación positiva, si $-1 < r < 0$, existe una correlación negativa. Si r = 0, no existe relación lineal, y es necesario ensayar con otros modelos de regresión. Y si r = -1, existe una correlación negativa lineal perfecta e inversa. La teoría de regresión es, junto con el teorema central del límite, la madre de las estadísticas y su uso es generalizado en las ciencias administrativas.

Ejercicio.

Un principio de la producción industrial establece que si se incrementa el número de unidades producidas, se puede habitualmente disminuir su costo por unidad. Los siguiente datos relacionan los costos unitarios de producción con el número de unidades producidas.

Unid.	10	20	50	100	150	200
Costo	9.4	9.2	9.0	8.5	8.1	7.4

Haciendo uso de la regresión lineal, prediga el costo por unidad para una producción de 125 unidades. Estime la varianza de dicho costo estimado. Calcule su coeficiente de correlación y diga si existe una relación lineal. Construya un intervalo de confianza para el costo unitario en un producción de 110 unidades, con un 95% de confianza.

CAPÍTULO XII

El análisis de datos categóricos

MAWIL
Publicaciones Impresas
y Digitales

www.mawil.us

12. El análisis de datos categóricos

12.1 El modelo multinomial y el estadístico X^2

La mayoría de los métodos de descripción e inferencia estadística que hemos estudiado hasta ahora no son aplicables al análisis de datos cualitativos o categóricos, al menos no explícitamente. Pero ofrecen las nociones y herramientas mínimas para aproximarse a su análisis, el cual será el objeto de este capítulo.

Una forma de modelar a los datos cualitativos es a través del **experimento multinomial** que consiste en

1. La realización de n ensayos idénticos e independientes
2. La categorización de los posibles resultados en k tipos, categorías o celdas, de acuerdo a una cualidad previamente determinada
3. El postulado de que la probabilidad, o frecuencia teórica, de que ocurra un evento en la categoría i sea p_i de tal forma que p1 + p2 + p3 + ... + pk = 1
4. El resultado, a saber, las cantidades n1, n2, n3,...,nk que representan el número de ensayos que cayeron en la categoría 1, la categoría 2, la categoria 3, y así hasta la categoría k, respectivamente, de tal forma que n1 + n2 + n3 + ... + nk = n.

La similitud con el ensayo binomial que estudiamos en los primeros capítulos, es evidente, se trata de su generalización: una prueba con más de k posibles resultados. En la tabla 12.1 se resume tal forma de abordar a los datos cualitativos.

Categoría	1	2	3	...	k	Total
Frecuencia teórica	p1	p2	p3	...	pk	1
Resultado observado	n1	n2	n3	...	nk	n

Tab.12.1

Una forma de realizar pruebas de hipótesis sobre la i-ésima frecuencia observada ni/n, es utilizando el hecho de que la esperanza del evento asociado a la categoría i es:

$$E(\text{categoría i})=n*pi$$

De forma tal que una medida de su dispersión viene dada por la diferencia:

$$ni - n*pi$$

Por la misma razón que se abordó a la regresión minimizando los errores cuadráticos, nos interesa trabajar con la diferencia al cuadrado:

$$(ni - n*pi)^2$$

Así pues, un estadístico relevante –debido a K. Pearson– para estu-

diar las desviaciones en los experimentos categóricos es:

$$X^2 = (n1 - n*p1)^2/np1 + (n2 - n*p2)^2/np2 + \ldots + (nk - n*pk)^2/npk$$

$$\mathbf{X^2 = \sum (ni - n*pi)^2/npi}$$

A dicho número se le conoce como **Estadístico Ji Cuadrado**, debido a que cuando n es mayor a 30, X^2 se distribuye ji cuadrado con k - 1 grados de libertad. Es usual que para fines de cálculo se le presente como

$$X^2 = (\sum ni^{\,2}/npi) - n$$

Sus usos son diversos. Se derivan, en general, pruebas de frecuencias, independencia y bondad de ajuste.

12.2 Prueba de bondad de ajuste

Bajo la condiciones de un experimentos multinomial con un n grande, y haciendo uso del estadístico Ji Cuadrado de Pearson, queremos probar la hipótesis sobre la coincidencia entre las proporciones teóricas y los resultados observados. A tal contraste de hipótesis se le denomina **prueba de bondad de ajuste**. Nuestra hipótesis nula sería:

$$H0: n1 = n*p1, n2 = n*p2, \ldots , nk = n*pk$$

Considerando como hipótesis alternativa:

$$H1: nj \neq n^*pj, \text{ para algún } 1 \leq j \leq k$$

El estadístico de prueba será:

$$X^2 = (n1 - n^*p1)^2/np1 + (n2 - n^*p2)^2/np2 + \ldots + (nk - n^*pk)^2/npk$$

Si consideramos un nivel de significancia α, nuestra región de rechazo;

$$RR: X^2 \geq J^2_{k-1,\,\alpha}$$

Examinemos un ejemplo. Se sabe que existen 4 partidos políticos electorales E1, E2, E3 y E4. El gobierno estima que las afiliaciones a dichos partidos son del 41%, 9%, 4% y 46%, respectivamente. Luego de un escándalo político, se teme que la distribución de simpatías y apoyos se vio modificada. Se decide realizar un experimento para verificar si el balance de fuerzas se mantiene. Para ello sea realiza una encuesta a 200 individuos y se registra su opinión para poner a prueba la hipótesis nula H0: nE1 = n*p1, nE2 = n*p2, nE3 = n*p3, nE4 = n*p4. Veamos los resultados registrados en una tabla:

	Partido E1	Partido E2	Partido E3	Partido E4
Proporciones a priori	p1=0.41	p2=0.09	p3=0.04	p4=0.46

Cantidades experimentales	nE1 = 92	nE2 = 20	nE3 = 4	nE4 = 84
			Total	200

La hipótesis nula es tal que:

H0: nE1 = 200*0.41, nE2 = 200*0.09, nE3 = 200*0.04, nE4 = 200*0.46,

Las cantidades esperadas serían entonces:

N*p1 = 200*0.41 = 82

N*p2 = 200*0.09 = 18

N*p3= 200*0.04 = 8

N*p4 = 200*0.46 = 92

Por lo que, sustituyendo para construir al estadístico de Pearson, se tiene que:

$X^2=(nE1 – N*p1)^2/N*p1 + (nE2 – N*p2)^2 /N*p2 + (nE3– n*p3)^2/N*p3 + (nE4 – N*p4)^2/N*p4$

$X^2=(92 – 82)^2/82 + (20 – 18)^2 /18 + (4 – 8)^2/8 + (84 – 92)^2/92$

$X^2=4.1374$

Notemos que hay 4 categorías, por lo que k - 1=3. Supongamos que

toleramos como probabilidad de cometer un error tipo I, es decir, aceptar la hipótesis nula dado que es falsa, a un $\alpha = 0.05$. Es decir, un nivel de significancia del 5%. Hemos de buscar en la tabla de distribución ji cuadrado el valor del cuantil de una ji cuadrado de 3 grados de libertad y cola del 0.05 en las tablas 8.2 y 8.3 Así $J^2_{4-1,\,0.05} = 7.8147$

Como no es cierto que $X^2 = 4.1374 \geq 7.8147 = J^2_{4-1,\,0.05}$, es decir, el valor del estadístico cayó fuera de la región crítica, entonces concluimos que no se puede rechazar la hipótesis nula con la información que provee el muestreo y el nivel de significancia del 5%.

Calculemos, sin embargo, el p-valor de esta prueba, para saber qué nivel de significancia en sería tal que tendríamos que rechazar la hipótesis nula. Veamos pues

$$P\left(J^2_{4-1,\,\alpha} \geq 4.1374\right) = \alpha$$

Mirando la tabla 8.3, concluimos que el valor debería estar entre 0.3 y 0.2, por lo que promediamos y concluimos que $\alpha=(0.3+0.2)/2=0.25$ aproximadamente. Así, si tolerásemos un 25% de posibilidad de cometer un error de tipo I, tendríamos que concluir que sí hubo un cambio en los apoyos a los partidos políticos tras el escándalo.

12.3 Comentarios finales

El análisis de los datos categóricos es una aplicación de la estadística ampliamente utilizada. Aunque requiera dominar herramientas sofisticadas como la distribución ji cuadrado, sus enunciados son sencillos de manejar, así como de establecer la certeza de sus conclusiones.

Otra técnica del análisis de datos cualitativos que merece una breve mención para estimular estudios posteriores, son las tablas de contingencia o prueba de independencia entre múltiples métodos de categorización de datos derivados de eventos cualitativos. El abordaje de esta técnica es similar a la prueba de bondad de ajuste, solo que esta vez se plantea en términos de una tabla con más información. Sea por ejemplo, una ampliación al ejemplos propuesto, donde se revisa el apoyo a los 4 partidos políticos de acuerdo a la región geográfica en donde residan los ciudadanos. De allí se derivaría el cuadro:

	Partido E1	Partido E2	Partido E3	Partido E4
Región rural	nE11	nE12	nE13	nE14
Región urbana	nE21	nE22	nE23	nE24

Con subtotales por fila y columna, cantidad de categorías (cuatro partidos, dos regiones) matriciales que determinarían los grados de libertad del estadístico en cuestión y, por lo tanto, la región de rechazo.

Es necesario hacer énfasis en la certidumbre en las conclusiones de las herramientas expuestas, depende fundamentalmente de que la muestra sea grande, así como también de saber determinar los grados de libertad de la distribución ji cuadrado. Así mismo, debe recordarse que la conclusión sobre que no debe rechazarse la hipótesis nula no implica que esta deba aceptarse, sino que la muestra no tiene información suficiente para ello.

Ejercicio

Los datos actuariales de un determinado país han venido registrando que un 52% de los usuarios de sus pólizas de salud lo hace en condición estable o preventiva, mientras que un 32% se encuentra en una situación seria, mientras que un 16% lo hace ya en un estado crítico de su afección. A pesar de la solidez de los estudios, un analista intuye que los porcentajes no son del todo cierto. Para verificarlo, realiza una muestra aleatoria de 300 pacientes que han utilizado su póliza de seguro en los últimos 3 meses. Así, el número de pacientes muestreados que reportan haber sido evaluado en un estado preventivo, serio y crítico son, respectivamente, 148, 92 y 60. ¿Demuestran estos datos la intuición del analista? Enuncie con detalle la hipótesis nula, construya una tabla y concluya con un nivel de significancia del 5% y del 10%. Determine además el p-valor.

TABLAS

x	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5	0,504	0,508	0,512	0,516	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,591	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,648	0,6517
0,4	0,6554	0,6591	0,Ç6628	0,6664	0,67	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,695	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,719	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,758	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,791	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,834	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,877	0,879	0,881	0,883
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,898	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,937	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,975	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,983	0,9834	0,9838	0,9842	0,9846	0,985	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,989
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,992	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,994	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,996	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,997	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,998	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,999	0,999
3,1	0,999	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997

3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

A-1. Tabla de la normal estándar Z para determinar P(Z>x) con $0 \leq x \leq 3.49$

	tp	0,4	0,3	0,2	0,1		0,05	0,025	0,01	0,005
g.l.										
1		0,3249	0,7265	1,3764	3,0777		6,3138	12,7062	31,8205	63,6567
2		0,2887	0,6172	1,0607	1,8856		2,92	4,3027	6,9646	9,9248
3		0,2767	0,5844	0,9785	1,6377		2,3534	3,1824	4,5407	5,8409
4		0,2707	0,5686	0,941	1,5332		2,1318	2,7764	3,7469	4,6041
5		0,2672	0,5594	0,9195	1,4759		2,015	2,5706	3,3649	4,0321
6		0,2648	0,5534	0,9057	1,4398		1,9432	2,4469	3,1427	3,7074
7		0,2632	0,5491	0,896	1,4149		1,8946	2,3646	2,998	3,4995
8		0,2619	0,5459	0,8889	1,3968		1,8595	2,306	2,8965	3,3554
9		0,261	0,5435	0,8834	1,383		1,8331	2,2622	2,8214	3,2498
10		0,2602	0,5415	0,8791	1,3722		1,8125	2,2281	2,7638	3,1693
11		0,2596	0,5399	0,8755	1,3634		1,7959	2,201	2,7181	3,1058
12		0,259	0,5386	0,8726	1,3562		1,7823	2,1788	2,681	3,0545
13		0,2586	0,5375	0,8702	1,3502		1,7709	2,1604	2,6503	3,0123
14		0,2582	0,5366	0,8681	1,345		1,7613	2,1448	2,6245	2,9768
15		0,2579	0,5357	0,8662	1,3406		1,7531	2,1314	2,6025	2,9467
16		0,2576	0,535	0,8647	1,3368		1,7459	2,1199	2,5835	2,9208
17		0,2573	0,5344	0,8633	1,3334		1,7396	2,1098	2,5669	2,8982
18		0,2571	0,5338	0,862	1,3304		1,7341	2,1009	2,5524	2,8784
19		0,2569	0,5333	0,861	1,3277		1,7291	2,093	2,5395	2,8609
20		0,2567	0,5329	0,86	1,3253		1,7247	2,086	2,528	2,8453
21		0,2566	0,5325	0,8591	1,3232		1,7207	2,0796	2,5176	2,8314
22		0,2564	0,5321	0,8583	1,3212		1,7171	2,0739	2,5083	2,8188
23		0,2563	0,5317	0,8575	1,3195		1,7139	2,0687	2,4999	2,8073
24		0,2562	0,5314	0,8569	1,3178		1,7109	2,0639	2,4922	2,7969
25		0,2561	0,5312	0,8562	1,3163		1,7081	2,0595	2,4851	2,7874
26		0,256	0,5309	0,8557	1,315		1,7056	2,0555	2,4786	2,7787
27		0,2559	0,5306	0,8551	1,3137		1,7033	2,0518	2,4727	2,7707
28		0,2558	0,5304	0,8546	1,3125		1,7011	2,0484	2,4671	2,7633
29		0,2557	0,5302	0,8542	1,3114		1,6991	2,0452	2,462	2,7564
30		0,2556	0,53	0,8538	1,3104		1,6973	2,0423	2,4573	2,75
40		0,255	0,5286	0,8507	1,3031		1,6839	2,0211	2,4233	2,7045
50		0,2547	0,5278	0,8489	1,2987		1,6759	2,0086	2,4033	2,6778
60		0,2545	0,5272	0,8477	1,2958		1,6706	2,0003	2,3901	2,6603
80		0,2542	0,5265	0,8461	1,2922		1,6641	1,9901	2,3739	2,6387
90		0,2541	0,5263	0,8456	1,291		1,662	1,9867	2,3685	2,6316
100		0,254	0,5261	0,8452	1,2901		1,6602	1,984	2,3642	2,6259
110		0,254	0,5259	0,8449	1,2893		1,6588	1,9818	2,3607	2,6213
120		0,2539	0,5258	0,8446	1,2886		1,6577	1,9799	2,3578	2,6174

A-2. Cuantiles de la t de Student

	Jp	0,995	0,99	0,975	0,95	0,9	0,8	0,7	0,6	0,5
g.l.										
1		0	0,0002	0,001	0,0039	0,0158	0,0642	0,1485	0,275	0,4549
2		0,01	0,0201	0,0506	0,1026	0,2107	0,4463	0,7133	1,0217	1,3863
3		0,0717	0,1148	0,2158	0,3518	0,5844	1,0052	1,4237	1,8692	2,366
4		0,207	0,2971	0,4844	0,7107	1,0636	1,6488	2,1947	2,7528	3,3567
5		0,4117	0,5543	0,8312	1,1455	1,6103	2,3425	2,9999	3,6555	4,3515
6		0,6757	0,8721	1,2373	1,6354	2,2041	3,0701	3,8276	4,5702	5,3481
7		0,9893	1,239	1,6899	2,1673	2,8331	3,8223	4,6713	5,4932	6,3458
8		1,3444	1,6465	2,1797	2,7326	3,4895	4,5936	5,5274	6,4226	7,3441
9		1,7349	2,0879	2,7004	3,3251	4,1682	5,3801	6,3933	7,357	8,3428
10		2,1559	2,5582	3,247	3,9403	4,8652	6,1791	7,2672	8,2955	9,3418
11		2,6032	3,0535	3,8157	4,5748	5,5778	6,9887	8,1479	9,2373	10,341
12		3,0738	3,5706	4,4038	5,226	6,3038	7,8073	9,0343	10,182	11,3403
13		3,565	4,1069	5,0088	5,8919	7,0415	8,6339	9,9257	11,1291	12,3398
14		4,0747	4,6604	5,6287	6,5706	7,7895	9,4673	10,8215	12,0785	13,3393
15		4,6009	5,2293	6,2621	7,2609	8,5468	10,307	11,7212	13,0297	14,3389
16		5,1422	5,8122	6,9077	7,9616	9,3122	11,1521	12,6243	13,9827	15,3385
17		5,6972	6,4078	7,5642	8,6718	10,0852	12,0023	13,5307	14,9373	16,3382
18		6,2648	7,0149	8,2307	9,3905	10,8649	12,857	14,4399	15,8932	17,3379
19		6,844	7,6327	8,9065	10,117	11,6509	13,7158	15,3517	16,8504	18,3377
20		7,4338	8,2604	9,5908	10,8508	12,4426	14,5784	16,2659	17,8088	19,3374
21		8,0337	8,8972	10,2829	11,5913	13,2396	15,4446	17,1823	18,7683	20,3372
22		8,6427	9,5425	10,9823	12,338	14,0415	16,314	18,1007	19,7288	21,337
23		9,2604	10,1957	11,6886	13,0905	14,848	17,1865	19,0211	20,6902	22,3369
24		9,8862	10,8564	12,4012	13,8484	15,6587	18,0618	19,9432	21,6525	23,3367
25		10,5197	11,524	13,1197	14,6114	16,4734	18,9398	20,867	22,6156	24,3366
26		11,1602	12,1981	13,8439	15,3792	17,2919	19,8202	21,7924	23,5794	25,3365
27		11,8076	12,8785	14,5734	16,1514	18,1139	20,703	22,7192	24,544	26,3363
28		12,4613	13,5647	15,3079	16,9279	18,9392	21,588	23,6475	25,5093	27,3362
29		13,1211	14,2565	16,0471	17,7084	19,7677	22,4751	24,577	26,4751	28,3361
30		13,7867	14,9535	16,7908	18,4927	20,5992	23,3641	25,5078	27,4416	29,336
40		20,7065	22,1643	24,433	26,5093	29,0505	32,345	34,8719	37,134	39,3353
50		27,9907	29,7067	32,3574	34,7643	37,6886	41,4492	44,3133	46,8638	49,3349
60		35,5345	37,4849	40,4817	43,188	46,4589	50,6406	53,8091	56,62	59,3347
80		51,1719	53,5401	57,1532	60,3915	64,2778	69,2069	72,9153	76,1879	79,3343
90		59,1963	61,7541	65,6466	69,126	73,2911	78,5584	82,5111	85,9925	89,3342
100		67,3276	70,0649	74,2219	77,9295	82,3581	87,9453	92,1289	95,8078	99,3341
110		75,55	78,4583	82,8671	86,7916	91,471	97,3624	101,7656	105,6323	109,3341
120		83,8516	86,9233	91,5726	95,7046	100,6236	106,8056	111,4186	115,4645	119,334

A-3.1 Cuantiles de la Ji cuadrado

Jp g.l.	0,4	0,3	0,2	0,1	0,05	0,025	0,01	0,005	0,0025
1	0,7083	1,0742	1,6424	2,7055	3,8415	5,0239	6,6349	7,8794	9,1406
2	1,8326	2,4079	3,2189	4,6052	5,9915	7,3778	9,2103	10,5966	11,9829
3	2,9462	3,6649	4,6416	6,2514	7,8147	9,3484	11,3449	12,8382	14,3203
4	4,0446	4,8784	5,9886	7,7794	9,4877	11,1433	13,2767	14,8603	16,4239
5	5,1319	6,0644	7,2893	9,2364	11,0705	12,8325	15,0863	16,7496	18,3856
6	6,2108	7,2311	8,5581	10,6446	12,5916	14,4494	16,8119	18,5476	20,2494
7	7,2832	8,3834	9,8032	12,017	14,0671	16,0128	18,4753	20,2777	22,0404
8	8,3505	9,5245	11,0301	13,3616	15,5073	17,5345	20,0902	21,955	23,7745
9	9,4136	10,6564	12,2421	14,6837	16,919	19,0228	21,666	23,5894	25,4625
10	10,4732	11,7807	13,442	15,9872	18,307	20,4832	23,2093	25,1882	27,1122
11	11,5298	12,8987	14,6314	17,275	19,6751	21,92	24,725	26,7568	28,7293
12	12,5838	14,0111	15,812	18,5493	21,0261	23,3367	26,217	28,2995	30,3185
13	13,6356	15,1187	16,9848	19,8119	22,362	24,7356	27,6882	29,8195	31,8831
14	14,6853	16,2221	18,1508	21,0641	23,6848	26,1189	29,1412	31,3193	33,426
15	15,7332	17,3217	19,3107	22,3071	24,9958	27,4884	30,5779	32,8013	34,9496
16	16,7795	18,4179	20,4651	23,5418	26,2962	28,8454	31,9999	34,2672	36,4557
17	17,8244	19,511	21,6146	24,769	27,5871	30,191	33,4087	35,7185	37,9461
18	18,8679	20,6014	22,7595	25,9894	28,8693	31,5264	34,8053	37,1565	39,4221
19	19,9102	21,6891	23,9004	27,2036	30,1435	32,8523	36,1909	38,5823	40,885
20	20,9514	22,7745	25,0375	28,412	31,4104	34,1696	37,5662	39,9968	42,3357
21	21,9915	23,8578	26,1711	29,6151	32,6706	35,4789	38,9322	41,4011	43,7751
22	23,0307	24,939	27,3015	30,8133	33,9244	36,7807	40,2894	42,7957	45,2041
23	24,0689	26,0184	28,4288	32,0069	35,1725	38,0756	41,6384	44,1813	46,6235
24	25,1063	27,096	29,5533	33,1962	36,415	39,3641	42,9798	45,5585	48,0337
25	26,143	28,1719	30,6752	34,3816	37,6525	40,6465	44,3141	46,9279	49,4354
26	27,1789	29,2463	31,7946	35,5632	38,8851	41,9232	45,6417	48,2899	50,8291
27	28,2141	30,3193	32,9117	36,7412	40,1133	43,1945	46,9629	49,6449	52,2153
28	29,2486	31,3909	34,0266	37,9159	41,3371	44,4608	48,2782	50,9934	53,5943
29	30,2825	32,4612	35,1394	39,0875	42,557	45,7223	49,5879	52,3356	54,9666
30	31,3159	33,5302	36,2502	40,256	43,773	46,9792	50,8922	53,672	56,3325
40	41,6222	44,1649	47,2685	51,8051	55,7585	59,3417	63,6907	66,766	69,6991
50	51,8916	54,7228	58,1638	63,1671	67,5048	71,4202	76,1539	79,49	82,664
60	62,1348	65,2265	68,9721	74,397	79,0819	83,2977	88,3794	91,9517	95,344
80	82,5663	86,1197	90,4053	96,5782	101,8795	106,6286	112,3288	116,3211	120,1017
90	92,7614	96,5238	101,0537	107,565	113,1453	118,1359	124,1163	128,2989	132,2556
100	102,9459	106,9058	111,6667	118,498	124,3421	129,5612	135,8067	140,1695	144,2928
110	113,1214	117,269	122,2495	129,3851	135,4802	140,9166	147,4143	151,9485	156,2304
120	123,289	127,6159	132,8063	140,2326	146,5674	152,2114	158,9502	163,6482	168,0817

A-3.2 Cuantiles de la Ji cuadrado

Bibliografía

Estadística para Administradores

www.mawil.us

Apostol, T. (2001). *Calculus. Vol. II. 2da edición.* Barcelona, España: Editorial Reverté.

Canavos, G. (1998). *Probabilidad y estadística. Aplicaciones y métodos. 1era edición.* México: McGraw-Hill Interamericana.

Galton, F. (1889). *Natural Inheritance.* London: Macmillan .

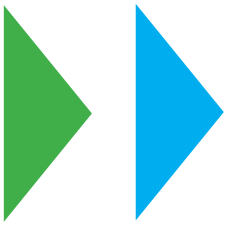
Mendenhall III, W., Wackerly, D., & Scheaffer, R. (2002). *Estadística matemática con aplicaciones. 6ta edición.* México: Thomson Editores.

Paulos, J. (1990). *El hombre anumérico. El analfabetismo matemático y sus consecuencias.* Barcelona, España: Tusquets Editores.

Ross, S. (2007). *Introducción a la estadística. 1era edición.* Barcelona, España: Editorial Reverté.

Spiegel, Murray, & Stephens, L. (2002). *Estadística. 3era edición.* México: McGraw-Hill Interamericana.

Estadística



para Administradores

MAWIL
Publicaciones Impresas y Digitales

Publicado en Ecuador
Marzo 2019

Edición realizada desde el mes de agosto del año 2018 hasta octubre del año 2018, en los talleres Editoriales de MAWIL publicaciones impresas y digitales de la ciudad de Quito.

Quito – Ecuador

Tiraje 200, Ejemplares, A5, 4 colores

Crossref

DOI: 10.26820/mawil/estadística-para-administradores

ESTADÍSTICA PARA ADMINISTRADORES

ISBN: 978-9942-787-44-6

9 789942 787446

10.26820/mawil/medicina-tropical-/978-9942-787-19-4