Universidad
Europea
del Atlántico

ESTADÍSTICA I

© TODOS LOS DERECHOS RESERVADOS

© TODOS LOS DERECHOS RESERVADOS

Universidad Europea del Atlántico

Índice

Presentación

1. Estadística descriptiva. Distribuciones de frecuencias

1.1. Etapas de una investigación 7
1.2. Tabulación de datos 8
1.3. Frecuencias Relativas-FR 10
1.4. Distribuciones de Frecuencias-DF 11
1.5. Percentiles 13
1.6. Calculando porcentajes de porcentajes 14
1.7. Tipos de estudios en ciencias de la salud y sociales 15
1.8. Aspectos a recordar 16

2. Estadística descriptiva. Medidas de centralización y de dispersión

2.1. Medidas de dispersión 21
2.2. Medidas de dispersión en la muestra 24
2.3. Cálculo de promedios en Distribuciones de Frecuencia 25
2.4. Valores Estandarizados 26
2.5. La media ponderada 27
2.6. Aspectos a recordar 28

3. Estadística descriptiva. Relación entre dos variables

3.1. Cómo se estudia la relación entre dos variables 33
3.2. Relación entre dos variables cualitativas 34
3.3. Independencia 37
3.4. Relación entre una variable cualitativa y otra cuantitativa 39
3.5. Relación entre dos variables cuantitativas 41

© Todos los derechos reservados

3.6. Variables respuesta y Factores. Respuesta y Efectos 42
3.7. Aspectos a recordar 44

4. Proporciones y probabilidad. Teorema de Bayes

4.1. Frecuencias relativas y probabilidades 47
4.2. Multiplicación de probabilidades 49
4.3. El azar y la necesidad 50
4.4. Probabilidades conjuntas, marginales y condicionadas 52
4.5. Teorema de Bayes 53
4.6. Tests diagnósticos 55
4.7. Aspectos a recordar. 55

5. Distribuciones de probabilidad: binomial, poisson y normal

5.1. Poblaciones y Muestras 59
5.2. La variabilidad e incertidumbre inherente a la extracción de una muestra 60
5.3. La Regularidad propia del Muestreo Aleatorio, MA. La distribución Binomial 61
5.4. ¿Para qué vale la regularidad propia del MA? 64
5.5. Distribución Binomial en general. Número de veces “+” en una serie de repeticiones de un FA 64
5.6. Distribución de Poisson. Binomial con N grande y π pequeña 66
5.7. Probabilidad de un intervalo de valores de la distribución Binomial 67
5.8. Media y Desviación de la distribución Binomial 68
5.9. El límite de la Binomial con N grande: distribución Normal 71
5.10. Aspectos a recordar 74

6. Inferencia estadística con una proporción

6.1. Inferencia estadística con una proporción 79
6.2. Intervalo de Confianza para proporción poblacional 81
6.3. Tests de significación con una proporción 82
6.4. Valor P del test con Binomial con resultado extremo 83
6.5. El valor P del test con resultado no extremo. P de la cola 85

© Todos los derechos reservados

6.6. ¿Por qué se usa la probabilidad de la cola? 86
6.7. Cálculo del valor P del test en Binomial con N grande.
Aproximación Normal 88
6.8. Resultados estadísticamente “significativos” y “muy significativos”.
Las barreras del 5% y del 1% 89
6.9. El valor P del Test NO es la probabilidad de que sea cierta la H0 91
6.10. La probabilidad de la H0 y el Valor P del Test 91
6.11. Riesgo de rechazar H0 equivocadamente 92
6.12. Aspectos a recordar 93

7. Distribución normal y teorema central del límite

7.1. Distribución de variables continuas 97
7.2. La distribución Normal 99
7.3. Distribución de las Medias Muestrales en el MA.
Teorema Central del Límite 104
7.4. Aspectos a recordar 107

8. Inferencia estadística con una media

8.1. Intervalo de confianza para una media 111
8.2. Test para una media con σ conocida y DN 112
8.3. Test con una media no conociendo σ 115
8.4. Test no conociendo σ y muestras pequeñas 116
8.5. Valor P a una cola y a dos colas 118
8.6. Intervalo de Confianza versus Tests Estadísticos 118
8.7. Muestras muy pequeñas 120
8.8. Ausencia de Normalidad y tests no paramétricos 120
8.9. Test de Normalidad 121
8.10. Tests no paramétricos 123
8.11. Estimadores sesgados y estimadores insesgados.
Varianza muestral = SC/(N-1) 125
8.12. Aspectos a recordar 126

© Todos los derechos reservados

9. La inferencia estadística con dos medias

9.1. IC para diferencia de dos medias 129

9.2. Test para la diferencia de dos medias 131

9.3. IC para diferencia de dos medias con datos apareados 135

9.4. Tests para diferencia de medias con datos apareados 136

9.4.1. Test para hipótesis de “no efecto poblacional” → H0: $\mu D = 0$ ó μDespués = μAntes 136

9.4.2. Test que propone cierto efecto poblacional → H0: $\mu D = \mu$Después - μAntes = Δ 136

9.5. Asunciones del test “t” para comparar dos medias 137

9.6. Tests No Paramétricos 138

9.7. Aspectos a recordar 139

10. La inferencia estadística con dos proporciones

10.1. Intervalo de Confianza para $\pi 2 - \pi 1$ 143

10.2. Test para la igualdad de dos proporciones poblacionales 144

10.3. Comparando dos proporciones con datos apareados 147

10.3.1. Intervalo de confianza para ($\pi S - \pi V$) con datos apareados 148

10.3.2. Tests para hipótesis para $\pi S = \pi V$ con datos apareados 148

10.4. Dos proporciones: datos apareados vs. datos independientes 149

10.5. Aspectos a recordar 150

11. Tamaño de muestra para estimación

11.1. Una anécdota para llamar la atención sobre un error 155

11.2. Tamaño de muestra para estimar una proporción 157

11.3. Tamaño de muestra para estimar diferencia de dos proporciones 158

11.4. Tamaño de muestra para estimar una media 160

11.5. Elementos de subjetividad e imprecisión al aplicar la fórmula para estimar una media 161

11.6. Tamaño de muestra para estimar diferencia de dos medias 161

11.7. Mitos y realidades en el tamaño de la muestra 162

11.8. Tamaños de muestra inadecuados 164

11.9. Aspectos a recordar 165

© Todos los derechos reservados

12. Tablas de contingencia

12.1. Tablas RxC: R muestras con C categorías 169
12.2. Comparación de varias proporciones 172
12.3. Colapsando tablas 174
12.4. El tipo de muestreo en las Tablas de Contingencia 176
12.5. Cálculo para test con Tablas de Contingencia 178
12.6. Una muestra con C categorías 180
12.7. Aspectos a recordar 182

13. Regresión lineal

13.1. Crecimiento y decrecimiento lineal 187
13.2. Recta más ajustada a los datos. Criterio de mínimos cuadrados 190
13.3. Recta de mínimos cuadrados en la muestra 192
13.4. Coeficiente de correlación lineal y coeficiente de determinación 194
13.5. Regresión lineal: modelo poblacional y muestreo 195
13.6. Inferencia en la regresión lineal 197
13.7. Aspectos a recordar 200

Bibliografía

Anexo

© Todos los derechos reservados

© TODOS LOS DERECHOS RESERVADOS

Presentación

El auge del Análisis Estadístico de los Datos (AED) en los últimos decenios se debe a dos hechos:

a) El creciente desarrollo tecnológico permite generar información detallada masiva sobre los más diversos temas de interés.

b) El desarrollo de la informática permite analizar esa información con rapidez, precisión y bajo costo, inimaginables hace 40 años.

De lo dicho se desprende que el AED es una herramienta necesaria para analizar la información que continuamente se genera en todas las ramas de saber y de la actividad social. Deming, padre del milagro industrial japonés, dice:

> "Ningún recurso es tan escaso en las empresas como el conocimiento estadístico. No hay conocimiento que pueda contribuir tanto a mejorar la calidad, la productividad y la competitividad de las empresas como el de los métodos estadísticos".

© Todos los derechos reservados

El AED usa valores medios

El AED opera calculando valores medios de las variables implicadas en ciertos grupos de individuos. En el AED no interesa el valor que la variable toma en cada individuo, sino la media del grupo. Así, para diagnosticar y tratar correctamente a un enfermo, el médico debe saber cuánto fuma y su tensión arterial, TA. Pero para saber si fumar aumenta la TA, el médico tiene que conocer la media de la TA en el grupo de fumadores y en el grupo de no fumadores.

Y calculando promedios de unas variables en distintos grupos de individuos el AED puede detectar relaciones no conocidas previamente.

En el caso de variables tipo "sí" o "no" se evalúa la proporción de individuos con esa característica en los grupos de interés. Así, decimos "La obesidad favorece la diabetes", aunque haya obesos sin diabetes y delgados con ella, porque la proporción de diabéticos es mucho mayor en el colectivo de obesos que en el de no obesos.

Estadística Descriptiva e Inferencia Estadística

Pero cada vez que analizando unos datos concretos se detecta en ellos una relación interesante nos preguntamos si el hallazgo encontrado en la muestra analizada es una verdad general, válida para toda la población. Por ejemplo, si en la muestra analizada se encuentra que la media de la TA

es mayor en el grupo de sedentarios que en el de deportistas, la cuestión es: ¿Ocurre eso en la población general o es una anécdota de nuestra muestra? Puesto que generalmente observamos muestras que son solo una pequeña parte de la población que intentamos conocer, esa pregunta es constante en todos los campos de la actividad científica, del análisis de encuestas y del control de calidad. Por ello en el AED se distinguen dos fases claramente diferenciadas y complementarias:

- La **Estadística Descriptiva** describe las relaciones de interés en las muestras que estamos explorando, calculando medias y proporciones en ellas.
- La **Inferencia Estadística** valora en qué medida las relaciones encontradas en nuestros datos ocurren también en la población general.

La Estadística Descriptiva descubre las relaciones en la muestra

Calcula medias y proporciones en distintos subgrupos de individuos de las muestras analizadas. Un ejemplo típico sería: Entre 600 pacientes con cierta enfermedad tratados con el fármaco "**A**" curaron el **32%** y entre otros 600 con la misma enfermedad tratados con el fármaco "**B**" curaron **50.3%**. La interpretación de estos resultados es obvia. Pero en otros casos la interpretación correcta de los resultados puede requerir reflexión cuidadosa, como se ve en estos ejemplos:

En una ciudad de 100 000 habitantes practican deporte (PD) 1000, es decir, el 1%. Tras una campaña animando a los vecinos a que PD se consigue que lo hagan otros 1000. El alcalde dice que el número de personas que PD se ha duplicado, es decir ha aumentado el **100%**. Pero la oposición dice que antes de la campaña eran sedentarias 99% y tras ella lo son 98%, es decir, el sedentarismo bajó solamente un 1%. Son dos enfoques muy diferentes de una misma realidad.

Como otro ejemplo veamos los resultados de un estudio que analiza el porcentaje de curaciones, "%+", obtenidas con cada uno de dos tratamientos, 'A' y 'B', para la misma enfermedad.

Tratamiento A

	Total	+	%+
Mujeres	100	80	80%
Varones	500	110	22%
Total	600	190	32%

Tratamiento B

	Total	+	%+
Mujeres	500	300	60%
Varones	100	2	2%
Total	600	302	50.3%

© Todos los derechos reservados

En mujeres “A” cura el 80% y “B” cura el 60%. En varones “A” cura el 22% y “B” cura el 2%. Por tanto, “A” es más efectivo que “B” tanto en mujeres como en varones. Pero si no se considera el sexo y se comparan los resultados en “personas” (última fila de la tabla), se ve que “A” es peor que “B”. Estos datos dejan confuso al investigador. Se pregunta cuál de los dos tratamientos es, según estas muestras, más efectivo, pues parece haber argumentos a favor de uno y a favor del otro. Aclarar este tipo de “contradicciones” es esencial para interpretar correctamente los datos y uno de los objetivos de la Estadística Descriptiva.

La Inferencia Estadística se basa en el cálculo de probabilidades

En biología, medicina, estudio de mercados y sociología hay pocas certezas. Las actitudes correctas no garantizan la respuesta buscada, solo aumentan la probabilidad de que ocurra. Suele pensarse que el cálculo de probabilidades implica notable complejidad matemática, pero la realidad es que en sus aplicaciones más frecuentes solo contiene cálculo matemático muy elemental y razonamiento lógico al alcance de todas las personas. Veamos algunos ejemplos de sus aplicaciones.

1. En una ciudad de 4 millones de habitantes la enfermedad “E”, que es mortal si no se diagnóstica a tiempo, afecta a 4000 de ellos. Se pone en marcha un método diagnóstico precoz que da positivo en el 99% de los enfermos, pero también da positivo (equivocadamente) en el 20% de los sanos. Se somete a ese método a todos los ciudadanos y en los que da positivo son convocados a un estudio más completo para averiguar si realmente tienen E. La mayoría de los convocados están muy asustados e incluso han hecho testamento, pues asumen que tienen alta probabilidad de tener E. Pero los pocos que conocen el cálculo elemental de probabilidades saben que su probabilidad de tener E es solamente 5 por mil.
2. Como ejemplo del Cálculo de probabilidades aplicado a la Inferencia Estadística consideremos el caso de una enfermedad para la que no hay tratamiento y cura espontáneamente en el **20%** de los casos. Un investigador cree que el producto “A” cura más del 20%, lo prueba en **8** enfermos y se curan todos.

Muchos miembros de la comunidad científica objetarán que tan buen resultado puede haberse presentado por azar, sin que “A” sea realmente efectivo, ya que es una muestra muy pequeña. Es una objeción muy razonable pero los que conocen el cálculo elemental de probabilidades enseguida comprueban que podemos tener gran confianza en que “A” es realmente efectivo.

Abreviaturas más usadas

Cada una de las 4 abreviaturas que siguen se usa cientos de veces en esta asignatura.

Abreviatura	Expresión completa, descripción y apartado en que se explica
FR	**Frecuencia Relativa** (singular o plural) Relacionan el tamaño de una parte con el tamaño total de un colectivo.

© Todos los derechos reservados

Abreviatura	Expresión completa, descripción y apartado en que se explica
DF	**Distribución de Frecuencias** (singular o plural) Dice la cantidad o la FR de individuos que tienen cada valor de una variable o están incluidos en cada intervalo.
Prob.	**Probabilidad** (singular o plural) La FR de veces que aparece cierto resultado si un fenómeno aleatorio se repite millones y millones (un número indefinidamente grande) de veces.
MA	**Muestreo Aleatorio** El hecho de extraer de una población millones y millones (un número indefinidamente grande) de muestras del mismo tamaño.

Capítulo	Objetivo particular	Resumen del capitulo	Aportación y resultado conseguido
1	Conocer la distribución de frecuencias como parte de los elementos de la estadística descriptiva.	Las etapas de una investigación, Tabulación de datos, Distribución de frecuencias, Percentiles y Porcentajes.	Se explican los elementos principales para realizar un análisis descriptivo de datos (estadística descriptiva) a través de la distribución de frecuencias, percentiles y porcentajes.
2	Conocer las medidas de centralización y de dispersión como parte de los elementos de la estadística descriptiva.	Las medidas de dispersión, promedios en distribuciones de frecuencias, los valores estandarizados y la media ponderada.	Se presenta cómo calcular las medidas de dispersión en una población y en una muestra, como parte fundamental en el análisis descriptivo de datos.
3	Establecer la relación entre dos variables como parte de la estadística descriptiva.	Definición de variables dependientes e independientes, evaluación de la relación entre dos variables (cualitativas, cuantitativas) evaluación de independencia entre las variables, factores y variables respuesta.	Se explica cómo realizar la clasificación de las variables (entre cualitativas o cuantitativas) y analizar las relaciones entre sí, para estudiar el comportamiento de una variable en los distintos niveles de la otra.
4	Definición de Probabilidad y Proporción como parte de la inferencia estadística.	Definición de Proporción, Probabilidad, Teorema de Bayes y Test diagnóticos.	Se explica los conceptos de probabilidad, el teorema de Bayes y test diagnósticos como herramientas para poder inferir la realidad de una población a partir de los datos obtenidos en una muestra.

© Todos los derechos reservados

Capítulo	Objetivo particular	Resumen del capitulo	Aportación y resultado conseguido
5	Evaluar la distribución de la probabilidad.	Definición de muestreo aleatorio, distribución binomial y distribución de Poisson. Definición de la media y desviación de la distribución binomial.	Se explica cómo analizar la distribución de la probabilidad para poder inferir de los datos de una muestra el comportamiento o los resultados esperados en una población.
6	Conocer la definición del Teorema Central del Límite y de la Distribución Normal.	Definición de la distribución normal, distribución de las medias muestrales de un muestreo aleatorio, teorema central del límite.	Se explica el concepto de distribución normal, el teorema del límite central y la distribución de medias muestrales como elementos fundamentales en el análisis de la distribución de variables continuas.
7	Conocer la inferencia estadística con una proporción.	Definición de intervalo de confianza y valor "*p*" de una proporción. Definición del valor "*p*" del test con binomial.	Se presenta cómo se puede inferir información sobre una población a partir de una muestra de la misma (proporción) a través del cálculo del intervalo de confianza y el test de significancia.
8	Conocer la inferencia estadística con una media.	Definición del intervalo de confianza y test de significancia para una media. Test de normalidad y test no paramétricos.	Se explica cómo se puede conocer el comportamiento de una media poblacional a partir de una media muestral, usando el intervalo de confianza, test de significancia con desviación estándar conocida y desconocida. También se brinda información sobre los test no paramétricos, empleados en los casos en que las variables no tienen una distribución normal.
9	Conocer la inferencia estadística con dos medias.	Definición del intervalo de confianza y test de significancia para la diferencia de medias muestrales. Test no paramétricos.	Se presenta cómo a partir de los datos muestrales se calcula el intervalo dentro del cual se tiene confianza que se encontrará la diferencia de las medias poblacionales, y el valor "p" propone un valor determinado para esa diferencia. Se muestra también el caso en el que los datos no tienen una distribución normal, donde se deberá usar los test no paramétricos.

© Todos los derechos reservados

Capítulo	Objetivo particular	Resumen del capitulo	Aportación y resultado conseguido
10	Conocer la inferencia estadística con dos proporciones.	Definición del intervalo de confianza y test de significancia para la diferencia de proporciones poblacionales. Definición de Test para la igualdad de proporciones poblacionales. Comparación de dos proporciones con datos pareados e independientes.	Se presentan las herramientas necesarias para hacer el cálculo e interpretación de los Intervalos de Confianza para la diferencia de medias poblacionales y la significancia estadística de los test que plantean hipótesis nulas (es decir que entre las medias no hay diferencia).
11	Calcular el tamaño de muestra para una estimación.	Definición de tamaño de muestra para estimar una proporción, una media y para estimar la diferencia entre dos proporciones y entre dos medias.	Se precisa cómo calcular el tamaño de una muestra para obtener un valor parecido al valor existente en toda la población.
12	Construir tablas de contingencia.	Tablas con “R” muestras con “C” categorías. Comparación de varias proporciones. Muestreo para realizar una tabla de contingencia. Cálculo de Test con tablas de contingencia.	Se presentan las Tablas de Contingencia como una forma de evaluación de la distribución de los datos de una muestra, que permita inferir que tan similares son en la población de procedencia. Se plantea como hipótesis nula que no existe diferencia entre las muestras que representan a las diversas poblaciones. Se muestra como hacer el cálculo con varias muestras y varias proporciones.
13	Conocer el análisis de datos con regresión lineal.	Criterio de mínimos cuadrados. Coeficiente de correlación lineal y Coeficiente de determinación. Regresión lineal: Muestreo, Inferencia, Intervalo de Confianza y Test de Significancia.	Se presentan los conceptos básicos para analizar el comportamiento de una variable en los diferentes niveles de la otra, en una relación lineal. Se explican los procedimientos para realizar un muestreo, el cálculo para el IC, Test de significancia e Inferencia del análisis de datos con una regresión lineal.

© Todos los derechos reservados

Estadística descriptiva. Distribuciones de frecuencias

1

Visión general

- Las etapas de una investigación.
- Tabulación de datos.
- Distribución de frecuencias.
- Percentiles y porcentajes.

Objetivo

- Conocer la distribución de frecuencias como parte de los elementos de la estadística descriptiva.

© Todos los derechos reservados

1.1. ETAPAS DE UNA INVESTIGACIÓN

En toda investigación científica pueden identificarse estas fases que siguen un orden cronológico evidente. Para llevar a cabo cada una de ellas es imprescindible que las anteriores estén correctamente realizadas.

a) Diseño del estudio.

b) Recolección de la información (fase de campo).

c) Análisis de la información:

- Tabulación de los datos.
- Estadística Descriptiva.
- Inferencia Estadística.

d) Discusión y conclusiones.

a) Diseño

Es la planificación de todo lo que se va a hacer. Se confecciona un Protocolo en el que se justifica la necesidad de esa investigación, se demuestra su viabilidad y se explican los recursos y gastos que se emplearán. La elaboración del protocolo ayuda decisivamente a clarificar y ordenar las ideas, así como a prever las dificultades y prevenir sus soluciones. Además es necesario presentarlo a los comités científicos, económicos y éticos para obtener autorización y apoyo económico.

b) Recogida de la información: recopilación de datos (fase de campo)

La calidad de los datos es requisito imprescindible para la validez de la investigación. Es fundamental la motivación adecuada de todas las personas implicadas. Los "decretos" pueden conseguir que se obtenga información, pero no que esta sea de calidad. Un ejemplo muy ilustrativo es el incremento del rendimiento del Departamento de Cálculo del Proyecto Manhattan cuando al ser encargado de ello R. Feynman consiguió motivarles informándoles del objetivo de su trabajo. Un segundo ejemplo más próximo en el tiempo es el Estudio Español de Malformaciones Congénitas (ECEMC), en el que desde hace 30 años más de un centenar de médicos registran meticulosamente información sobre 200 variables de cada bebé malformado nacido en su centro y de otros tantos controles sanos.

c) Análisis de la información

Comienza recopilando la información más relevante en forma de Tablas. Continúa haciendo Estadística Descriptiva de lo observado. Y culmina haciendo Inferencia Estadística para ver hasta qué punto los resultados encontrados en la muestra pueden extrapolarse a la población.

d) Discusión y conclusiones

Se elaboran teniendo en cuenta los conocimientos previos que había sobre el tema y los resultados obtenidos en esta investigación. Esta disciplina, Análisis de Datos, cubre el tercer punto.

1.2. TABULACIÓN DE DATOS

A lo largo de la fase de campo de un estudio la información de los distintos individuos implicados se va recogiendo en Historias Clínicas (estudios hospitalarios), en Libretas de Laboratorio (investigaciones básicas con animales o cultivos) o en Encuestas Sanitarias (estudios epidemiológicos). Las características evaluadas en cada uno de los individuos se denominan variables.

El análisis de esa información comienza con la Tabulación de los datos, que consiste en disponerlos en tablas donde cada fila corresponde a un individuo y cada columna corresponde a una variable. El siguiente esquema representa una tabla con datos de 200 individuos en los que

© Todos los derechos reservados

se ha recolectado, entre otras, las variables: sexo, grupo sanguíneo, número de caries, peso y edad.

Para cada individuo, cada variable toma un valor concreto. Por ejemplo, el primer individuo, identificado con el número 1 es un hombre (codificamos con "1" a los hombres y con "2" a las mujeres), tiene grupo sanguíneo "AB", tiene 3 caries, pesa 65.9 kg y tiene 18 años.

0BInd. nº	Sexo	Fuma	Grupo S.	...	Caries	TA	1BEdad
1	1	2	AB		3	65.9	18
4	1	2	A		3	65.0	42
8	2	2	O		2	59.4	25
---	---		---		---	---	---
200	1	2	AB		1	83.7	49

En general, las variables se clasifican en:

- **Cualitativas**, si recogen alguna cualidad no numérica del individuo. Se llaman dicotómicas si presentan solamente dos posibilidades y politómicas, si presentan varias posibilidades (ordenables o no). Los valores de este tipo de variables suelen registrarse con códigos numéricos, que no indican cantidad, sino un convenio, como por ejemplo 1 = Varón, 2 = Mujer.
- **Cuantitativas**, si recogen una información numérica. Pueden ser discretas si toman valores enteros, como número de hijos o número de caries (1, 2, 3,... caries) o continuas si pueden tomar cualquier valor dentro de un intervalo, como edad, peso, tensión arterial...

Cualitativa	**Dicotómica.** Sexo: 1 = Varón, 2 = Mujer
	Politómica. Grupo Sanguíneo: 1 = A, 2 = B, 3 = AB, 4 = O
	Politómica Ordenada. Higiene: 1 = Nada, 2 = Poca, 3 = Media, 4 = Mucha
Cuantitativa	**Discreta.** Número de caries
	Continua. Concentración de ácido úrico, Edad, TA

La clasificación de los individuos en cada variable tiene que ser exhaustiva (todo individuo debe tener un grupo en el que encaje) y excluyente (cada individuo encaja solamente en un grupo).

Las tablas así creadas contienen toda la información de interés, pero en muchos casos el investigador no está interesado en el valor que toma una variable en cada individuo, sino en la proporción de ellos con cierto valor o en la media de esa variable en un grupo. Una de las variables de nuestra tabla indica si fuman o no y la otra su tensión arterial, TA. Para diagnosticar y tratar a cada sujeto el médico necesita conocer su TA y cuánto fuma. Pero para investigar si fumar

© Todos los derechos reservados

favorece la aparición de hipertensión hay que formar dos grupos de personas, los que fuman y los que no, y en cada uno de ellos ver la media de la TA o la proporción de hipertensos.

Por ello el siguiente paso en el análisis de los datos es realizar la Estadística Descriptiva, que tiene por objeto resumir el comportamiento de cada una de las variables en el grupo estudiado y en cada subgrupo que sea de interés. Este resumen se hará de distinta forma dependiendo del tipo de variable.

En todas ellas se pueden calcular Distribuciones de Frecuencia y en las cuantitativas se puede, además, calcular medias y desviaciones.

1.3. FRECUENCIAS RELATIVAS-FR

Las frecuencias relativas, FR, relacionan el tamaño de una parte con el tamaño total de un colectivo. Si decimos que en una ciudad A hay 4.000 obesos y en la ciudad B hay 16.000, es claro que en la capital B hay muchos más casos que en la capital A, y la cantidad de recursos necesarios para asistir a estos pacientes es mucho mayor en la segunda que en la primera.

Pero la frecuencia relativa, FR, de ese problema es mayor en la ciudad A, ya que el total de población de esta ciudad es 200.000 habitantes. Por lo tanto los 4000 obesos representan el dos por cien o el veinte por mil.

$$4.000/200.000 = 2/100 = 0.02$$

En el caso de la ciudad B, que cuenta con una población de 4 millones de habitantes, los 16.000 obesos representan cuatro por cada mil.

$$16.000/4.000.0000 = 16/4.000 = 0.004$$

La tabla siguiente resume estos datos:

	Frecuencia absoluta	Núm. de habitantes	FR proporción	FR porcentaje	FR tanto por mil
Ciudad A	4000	200.000	0.02	2%	20 por 1.000
Ciudad B	16.000	4.000.000	0.004	0.4%	4 por 1.000

La FR se puede dar como proporción (por ejemplo, 0.02) o como porcentaje (2%). Pero si son FR muy bajas es más útil darlas como tanto por mil o por diez mil u otra cantidad pertinente. Por ejemplo, si en la ciudad B hay 80 hemofílicos, la frecuencia relativa es “80 entre 4 millones”, es decir:

$$80/4.000.000 = 0.00002 \text{ o } 0.002\%$$

© Todos los derechos reservados

Pero esa cifra se entiende mejor expresándola como “2 por cien mil” o “20 por millón”. El porcentaje se calcula multiplicando la proporción por 100, es decir, corriendo la coma (hoy día el punto) dos lugares a la derecha. Si de un total de 80 personas 32 son enfermos, dividiendo 32 por 80 obtenemos la proporción de enfermos, que multiplicada por 100 nos da el porcentaje:

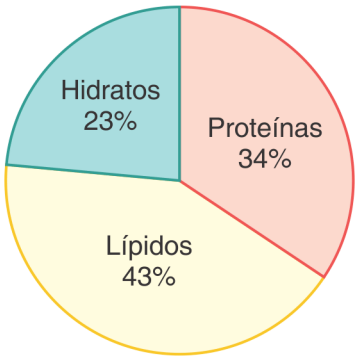
$$\text{Proporción: } 32/80 = 0.40. \text{ Porcentaje: } 0.40 \cdot 100 = 40\%$$

Tanto en Estadística Descriptiva como en la Inferencial se usa constantemente las Frecuencias Relativas. La mayoría de las veces se las refiere como “FR”.

1.4. DISTRIBUCIONES DE FRECUENCIAS-DF

El comportamiento de una variable cualitativa en un grupo de individuos se resume dando su Distribución de Frecuencias, DF, que consiste en anotar la cantidad de individuos que tienen cada valor de la variable. Por ejemplo, si de cada individuo se recoge el tipo de dieta que sigue (hay tres posibilidades) la columna del medio de la siguiente tabla da la frecuencia absoluta de cada dieta:

	Frecuencia	FR
Proteínas	68	0.34
Lípidos	86	0.43
Hidratos	46	0.23
Total	200	1



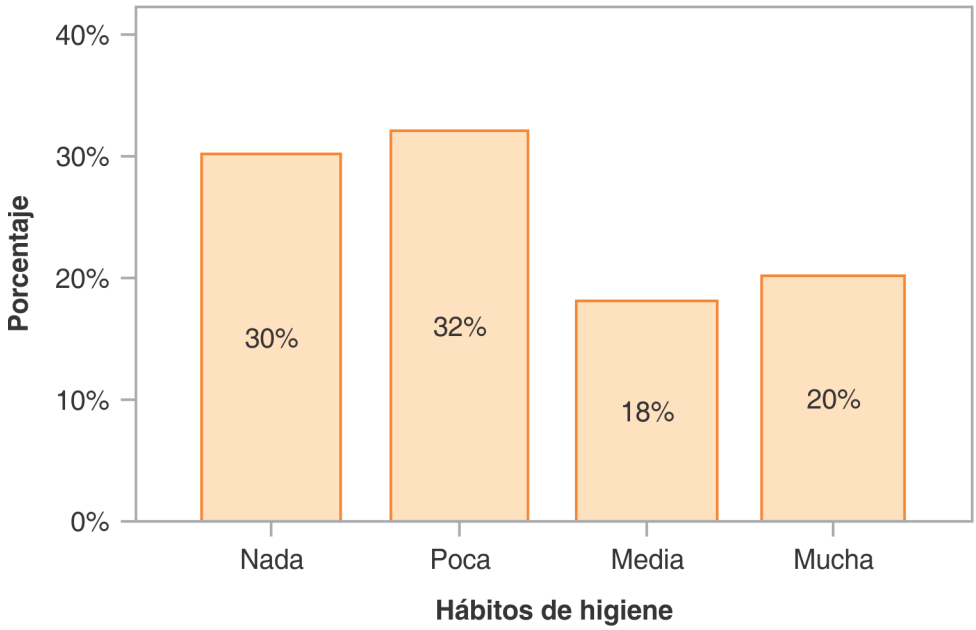
De los 200 individuos estudiados, 68 siguen dieta rica en proteínas, 86 dieta rica en lípidos y 46 dieta rica en hidratos. Estas cantidades forman la ***DF Absolutas*** (número de individuos en cada una de las categorías).

La **DF Relativas** indica la **FR** de individuos para cada valor de la variable. En nuestro ejemplo 68/200 = 0.34 o 34% sigue dieta rica en proteínas, 43% en lípidos y 23% en hidratos.

Estas cantidades pueden representarse gráficamente en diagramas de sectores, donde el ángulo de cada sector es proporcional a la frecuencia del valor que representa.

También puede presentarse en un diagrama de barras, donde la altura de cada una es proporcional a las frecuencias de cada clase.

© Todos los derechos reservados



Este diagrama de barras representa la distribución de frecuencias relativa de la variable "Hábitos de Higiene" recogida en 200 individuos: el 30% presentaba nada de higiene, el 32% poca, el 18% una cantidad media y el 20% mucha.

- Si la variable en estudio es cuantitativa con pocos valores, su comportamiento se resume también dando su DF de la misma forma que con las cualitativas. Con variables cuantitativas podemos, además, calcular las llamadas Frecuencias Acumuladas, que indican la cantidad o FR de individuos con valor igual o menor que uno dado.

  Por ejemplo:

Núm. de caries	Frecuencia absoluta	FR	FR acumulada
0	40	0.20	0.20
1	10	0.05	0.25
2	50	0.25	0.50
3	80	0.40	0.90
4	20	0.10	1

  El 25% (20+5) tiene una caries o menos, el 50% (25+25) tiene 2 caries o menos.

- Si la variable en estudio es cuantitativa discreta con muchos valores o continua se pueden hacer **intervalos** y contar el número de individuos en cada uno para hacer la correspondiente DF. Los intervalos deben cubrir todo el rango de posibles valores y no solaparse. No hay razones matemáticas ni biológicas para determinar la anchura de los intervalos. El investigador agrupa los datos como sea más útil a su estudio. Una misma variable puede ser agrupada con distintos criterios en distintos momentos. Por ejemplo,

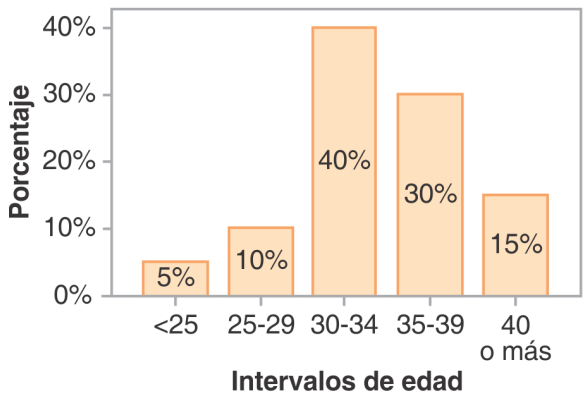
© Todos los derechos reservados

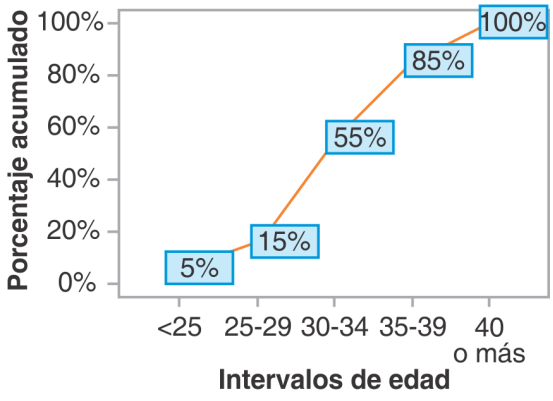
con la edad pueden formarse dos grandes grupos, niños y adultos. Y en otro momento pueden agruparse como niños, jóvenes, adultos y viejos, estableciendo los puntos de corte pertinentes. O pueden hacerse intervalos por decenios: de 0 a 9, de 10 a 19... de 90 a 99.

También se pueden representar las frecuencias absolutas y/o relativas de cada intervalo en diagramas de barras o sectores, y las frecuencias acumuladas en diagramas barras y de líneas.

La tabla que sigue da la DF de la edad de 200 individuos:

			Porcentaje
Menos de 25 años	10	5,0	5,0
25-29 años	20	10,0	15,0
30-34 años	80	40,0	55,0
35-39 años	60	30,0	85,0
40 o más años	30	15,0	100,0
Total	200	100,0	





La DF acumulada –columna a la derecha– dice que el 15% de los individuos tiene 29 o menos años, el 55% tiene 34 o menos y el 85% tiene menos de 40. Muchos investigadores tienden a tabular las variables continuas acordando una partición en intervalos y registrando para cada individuo el intervalo al que pertenece.

Ello supone una notable pérdida de información que debe evitarse. Si, por ejemplo, la edad se sustituye por la década a la que pertenece, un paciente de 41 años es indistinguible de uno de 49, lo que disminuiría la posibilidad de detectar relaciones de interés. Por ello lo razonable es tabular los valores originales y en la fase de análisis agrupar por los intervalos que interesan en cada momento.

1.5. PERCENTILES

Asociado al concepto de FR Acumulada está el de **Percentil**. En una DF de una variable cuantitativa el Percentil que le corresponde a cada valor es el porcentaje de individuos con valor igual o menor que aquel. En la DF del número de caries, 1 caries es el Percentil 25, 2 caries es el

© Todos los derechos reservados

Percentil 50. En la DF de la edad, 39 años es el Percentil 85. Si en la DF de la estatura de un colectivo, 177 cm es el Percentil 90, quiere decir que el 90% de los individuos mide 177 cm o menos.

La idea de percentil es muy útil al ubicar un individuo en el contexto del grupo al que pertenece y esta ubicación relativa suele ser más útil que el valor numérico original. A unos padres preocupados por el escaso desarrollo del niño no les ayuda saber que en un test de capacidad mental alcanzó 220 puntos. Inmediatamente preguntan cuáles son las puntuaciones propias de los niños con capacidad mental media. Lo que necesitan es saber en qué percentil está su hijo. Si resulta ser el percentil 3, es mala noticia, puesto que solo el 3% de los niños de esa edad tienen esa capacidad o menos. Si 220 puntos corresponden al percentil 47 o el 55, es una noticia tranquilizadora, puesto que el niño está en la zona media. Y si 220 puntos es el percentil 99, de cada cien niños solo uno le supera. Los percentiles se usan constantemente en la Inferencia Estadística.

A un paciente no le dice mucho saber que su tensión arterial es 120. Lo útil para él es saber a qué percentil corresponde esa cifra. Si es el percentil 97 quiere decir que de su sexo y edad, solo 3 de cada 100 le superan. Es una cifra relativamente muy alta, y por ello preocupante. En muchas ocasiones para analizar los datos hay que calcular "proporción de una proporción". Si, por ejemplo, en un colectivo son enfermos el 10% y el 50% de los enfermos han sido operados, es claro que los operados son el 5% del colectivo (la mitad del 10%), pero con cifras menos sencillas se requiere un cálculo con el que algunos investigadores no están familiarizados.

© Todos los derechos reservados

1.6. CALCULANDO PORCENTAJES DE PORCENTAJES

a) Cálculo del Total de individuos si se conoce el porcentaje

Si sabemos que de un total de 80 personas el 60% son varones, para saber la cantidad de varones se multiplica el total por la proporción:

$$\text{Cantidad de varones} = 80 \cdot 0.60 = 48 \text{ varones}$$

b) Cálculo del porcentaje de un porcentaje

El porcentaje de un porcentaje se calcula multiplicando las proporciones. Si el 60% de un colectivo son varones y el 10% de los varones tienen enfermedades de la próstata, ¿qué porcentaje de personas tiene patologías de la próstata?

$$\text{Son el 10\% del 60\%} = 0.10 \cdot 0.60 = 0.06 \text{ o } 6\%.$$

c) Proporción de enfermos debido a cierto factor

Para practicar el cálculo de proporciones averigüemos cuántos casos de la enfermedad "E" son atribuibles a la presencia de un factor, por ejemplo fumar, en una población. Para ello se calcula:

- El total de enfermos que habría en la población si nadie fumara.
- El total de enfermos que habrá teniendo en cuenta que parte de ella fuma.

La diferencia de esas dos cantidades es el exceso de enfermos debido al tabaco.

Esta cantidad depende de lo perjudicial que sea fumar (cuánto aumenta el riesgo de enfermedad) y de cuan frecuentes sean la enfermedad y el hábito de fumar.

Calculemos el exceso de enfermos debido al tabaco en una población donde:

- El número total de habitantes es de 1.000.000.
- La enfermedad afecta al 1% de los No–fumadores.
- Fuman el 10% de la población.
- Fumar multiplica por 4 el riesgo de contraer la enfermedad, es decir, la FR de E es 4 veces mayor entre los fumadores que entre los No-fumadores.

Se pide:

1. Calcular el exceso de enfermos debidos al tabaco.
2. Calcular el porcentaje de enfermos debidos al tabaco.
   - Número de enfermos que habría si nadie fumara:

     1% de 1.000.000 = 10.000.
   - Número de enfermos que habrá dado que el 10% de la población fuma:

     Si el tabaco cuadruplica el porcentaje de enfermos.

     - Enferman el 4% de los Fumadores.

     Hay 100.000 fumadores.

     - Enferman el 4% = 4.000.

     Hay 900.000 No fumadores.

     - Enferman el 1% = 9.000.

     Total de enfermos (fumadores y no fumadores) = 13.000.
   - El exceso de enfermos debido al Tabaco: 13.000–10.000 = 3.000.
   - Porcentaje de enfermos debido al tabaco: (3.000/13.000)*100 = 23%.

1.7. TIPOS DE ESTUDIOS EN CIENCIAS DE LA SALUD Y SOCIALES

La lógica de la investigación es prácticamente la misma en todas las ciencias, pero en las ciencias de la salud –somática o psíquica, individual o colectiva– concurren algunas circunstancias que le

© Todos los derechos reservados

confieren matices especiales. Los estudios en este campo pueden clasificarse en tres grandes categorías:

- **Estudios experimentales:** las unidades de experimentación son animales o cultivos de células o tejidos. También se le llama "*investigación básica*".
- **Estudios clínicos:** se estudia a pacientes o grupos aquejados de enfermedades o anomalías. Dentro de este grupo los Ensayos Clínicos aleatorizados y controlados se consideran hoy en día imprescindibles para demostrar la ventaja de un procedimiento terapéutico nuevo sobre sus competidores.
- **Estudios epidemiológicos:** se estudia típicamente, aunque no siempre, a individuos sanos, para detectar factores que favorecen o dificultan la aparición de enfermedades. Este tipo de estudios son imprescindibles en Medicina y Sociología Preventivas. Téngase en cuenta que el paradigma de la Medicina no es curar todas las enfermedades, sino evitar su aparición.

1.8. ASPECTOS A RECORDAR

- La Estadística Aplicada se encarga de analizar la información recogida en cualquier investigación. El análisis comienza recopilando la información en Tablas, continúa haciendo Estadística Descriptiva (resumen de la información) y termina con la Inferencia Estadística (intento de extrapolar a la población los resultados obtenidos en la muestra).
- Las características evaluadas en cada uno de los individuos estudiados se llaman variables. Pueden ser cualitativas (dicotómicas o politómicas) y cuantitativas (discretas o continuas).
- Distribución de Frecuencias, DF, absoluta es la cantidad de individuos en cada uno de los posibles valores de la variable o de los intervalos convenidos.
- Los intervalos de cada variable cuantitativa los decide el investigador. No hay criterios matemáticos para ello. Pero cada individuo debe pertenecer a una categoría y solo a una (clasificación exhaustiva y excluyente). Una misma variable se puede agrupar de diferentes modos en diferentes momentos.
- Las Frecuencias Relativas, FR, relacionan el tamaño de una parte con el tamaño total de un colectivo. Las FR se puede dar como proporción, o porcentaje o tanto por mil o tanto por otra cantidad.
- La DF relativas a la FR de individuos en cada uno de los posibles valores de la variable o de los intervalos convenidos.
- Las DF se pueden representar en diagramas de sectores, barras o histogramas.
- Con variables cuantitativas se calculan también frecuencias acumuladas. Si en la DF de la estatura hay, por ejemplo, 70% de individuos con estatura menor de 180 cm, se dice que 180 cm es el percentil 70.
- En muchas situaciones no interesa tanto la magnitud que cierta variable toma en un individuo como el percentil que representa, la posición que ocupa en su grupo.
- El porcentaje de un porcentaje se calcula multiplicando las proporciones.

© Todos los derechos reservados

- Para calcular la cantidad de enfermos debida a la presencia de un factor se calcula la cantidad de ellos que hay estando presente el factor y se le resta la cantidad de enfermos que habría si nadie estuviera expuesto.
- Las investigaciones en Ciencias de la Salud y sociales pueden clasificarse en tres grandes categorías: experimentales, clínicas y epidemiológicas.

© Todos los derechos reservados

© Todos los derechos reservados

Resumen

© Todos los derechos reservados

© Todos los derechos reservados

Estadística descriptiva. Medidas de centralización y de dispersión

2

Visión general

- Medidas de dispersión.
- Promedios en distribuciones de frecuencias.
- Valores estandarizados.
- Media ponderada.

Objetivo

- Conocer las medidas de centralización y de dispersión como parte de los elementos de la estadística descriptiva.

© Todos los derechos reservados

2.1. MEDIDAS DE DISPERSIÓN

Con variables cuantitativas, además de resumir la información agrupando los valores en intervalos y calculando la Distribución de Frecuencias de éstos, se pueden calcular las llamadas *medidas de posición* (o centralización) y de *dispersión*.

Las *medidas de posición* indican en torno a qué valores están los datos. Su significado es obvio y su cálculo no ofrece dificultades conceptuales. La más utilizada es la *media aritmética*, el promedio de los datos, que es la suma de todos los valores divida por el número de ellos, es decir, $M = (\Sigma X)/N$. Algunas veces se utiliza la mediana, valor de la variable con frecuencia acumulada de 50%, es decir, el percentil 50, el valor situado en el centro cuando se ordenan los datos de menor a mayor valor. La moda es el valor más frecuente de la variable, el más repetido.

Las medidas de dispersión son menos conocidas pero no menos importantes. Dan idea de lo agrupados o dispersos que están entre sí los datos. Imagine una asignatura en que la nota final es la media de dos exámenes parciales. Si las notas del estudiante A son 4.8 y 5.2, su media es 5. Si las del estudiante B son 0 y 10 también tiene media 5, pero la situación es muy diferente del

anterior. A tiene un conocimiento moderado de ambas partes de la materia, mientras que B ignora totalmente una parte y conoce perfectamente la otra. Si una variable corporal de las muchas que se miden en Medicina presenta sistemáticamente valores muy próximos a la media, cuando en un paciente se encuentra un valor moderadamente alejado de ella se considerará indicio de enfermedad, señal de alarma. Pero en una variable que presenta espontáneamente valores muy dispersos respecto a la media, un valor moderadamente alejado de ella no se considerará un indicio de enfermedad, sino expresión de la variabilidad de esa medida.

A continuación se explican las medidas de dispersión más usadas de modo que el alumno entienda lo que indican. Términos como "*Varianza*", "*Desviación Estándar*" y "*Error Típico*" no pueden ser para el investigador "índices matemáticos" o "parámetros estadísticos" de significado incierto, cuya existencia acepta como un mal menor sin entender lo que expresan. Veremos que todas son medias de cantidades que indican la dispersión de los valores respecto a la media aritmética y por tanto tienen una interpretación intuitiva muy sencilla.

Para exponer las medidas de dispersión tomemos como ejemplo dos grupos de N = 9 individuos cada uno, en los que se mide la variable cuantitativa "X", cuyos valores aparecen en la primera columna de la tabla que sigue. En ambos grupos la media es M = 10 años. El grupo del ejemplo 1º tiene mediana y moda 10 y en el otro ambas son 13 años.

En la segunda columna de cada tabla aparece la diferencia d = valor del individuo menos la media del grupo y en la tercera columna esa diferencia al cuadrado.

Así, el primer individuo del grupo 1 tiene valor X = 11 y está 1 unidad por encima (signo +) de la media de su grupo y el primer individuo del segundo grupo, que tiene valor X = 5 y está 5 por debajo (signo –) de la media de su grupo.

La tabla siguiente recoge las distintas medidas de dispersión.

Ejemplo 1º Poca dispersión		
X	**d = (X–M)**	**d^2 = $(X–M)^2$**
11	+1	1
9	–1	1
10	0	0
12	+2	4
9	–1	1
10	0	0
8	–2	4
10	0	0
11	+1	1
M = 10	$\Sigma \lvert d \rvert$ **= 8**	**SC = Σd^2 = 12**

Ejemplo 2º Mucha dispersión		
X	**d = (X–M)**	**d^2 $(X–M)^2$**
5	-5	25
13	+3	9
2	–8	64
20	+10	100
15	+5	25
4	-6	36
17	+7	49
13	+3	9
1	–9	81
M = 10	$\Sigma \lvert d \rvert$ **= 56**	**SC = Σd^2 = 398**

© Todos los derechos reservados

- **El rango** nos dice entre qué mínimo y máximo están los valores de la variable.
- **La Desviación Media**, DM, es la media de los valores absolutos de las d, es decir, la suma de ellas ignorando el signo dividida por el número de casos (la suma de las d´s con su signo es cero, y por ello no informa acerca de si son grandes o pequeñas). Si la DM de un grupo es, por ejemplo, 3 quiere decir que por término medio los valores se alejan 3 de su media, unos por encima de ella y otros por debajo, no que cada uno de ellos se aleje justamente 3 (del mismo modo que si la media, M, es 10 no quiere decir que todos los valores sean 10).

Medida		Fórmula	Grupo 1	Grupo 2
Rango o Amplitud	RA	Valores máximo y mínimo	De 8 a 12	De 1 a 20
Desviación Media	DM	$\Sigma \lvert d \rvert /N$	8/9 = **0.89**	56/9 = **6.22**
Suma de Cuadrados	SC	$\Sigma d^2 = \Sigma(X - M)^2$	**12**	**398**
Varianza	σ^2	SC/N	**1.33**	398/9= **44.22**
Desviación Estándar	σ	$(\text{Varianza})^{1/2}$	$(1.33)^{1/2}$= **1.15**	$(44.22)^{1/2}$ = **6.65**

En la siguiente lista de 4 números (7, 7, 13 y 13) con media 10, la DM es 3 y cada uno se aleja 3 de la media, los dos primeros por debajo de 10 (d = –3) y los otros por encima de 10 (d = +3).

Valores →	7	7	13	13	M = 10
Distancias →	–3	–3	+3	+3	DM = (3+3+3+3)/4 = 3

Esta otra lista también tiene M =10 y DM = 3. Unos se alejan 2 y otros 4, y la media de esas distancias es 3 (DM = 3).

Valores →	6	8	12	14	M = 10
Distancias →	–4	–2	+2	+4	DM = (4+2+2+4)/4 = 3

En esta otra lista con media 10, las distancias individuales son 6, 0, 2 y 4 y la media de estas desviaciones es 3 (DM = 3).

Valores →	4	10	12	14	M = 10
Distancias →	–6	0	+2	+4	DM = (6+0+2+4)/4 = 3

Por tanto, en una lista de valores con DM pequeña las **d** son pequeñas, pero generalmente no iguales, y si la DM es grande las **d** son en general grandes pero puede haber unas pequeñas y otras muy grandes.

© Todos los derechos reservados

- **La Suma de Cuadrados**, SC, es la suma de las d^2; es decir, de los cuadrados de las distancias de cada valor a la media del grupo (no debe confundirse con la suma de los valores originales al cuadrado).

  Para este último ejemplo es:

  SC = $-6^2+0^2+2^2+4^2 = 56$, no $4^2+10^2+12^2+14^2 = 456$.
- **La varianza**, σ^2, es la media de las d^2, es decir, la SC partido por N.
- **La desviación estándar o típica** es la raíz cuadrada de la varianza, es decir, la raíz cuadrada de la media de los cuadrados de las desviaciones. Tiene valor numérico próximo a la DM y nos informa, cuánto, por término medio, se alejan los valores de la media aritmética. En la práctica, para resumir en un solo número el alejamiento de los valores respecto a la media, se usa la Desviación Estándar, nunca la DM. Esta última se ha mencionado aquí para facilitar la comprensión de lo que indica aquella.

2.2. MEDIDAS DE DISPERSIÓN EN LA MUESTRA

Llamamos *Población* al total de individuos con ciertas características y *Muestra* al conjunto de individuos extraídos de la población que observamos y/o manipulamos en un estudio, evaluando en ellos ciertas variables que son de nuestro interés. En el capítulo 6 se comentan detenidamente estos conceptos.

Es habitual representar la media poblacional por μ y la Desviación Estándar poblacional por σ. Cuando la lista de valores que tenemos corresponde a toda la población de interés, las medidas de posición y dispersión se calculan como se dijo en el apartado anterior.

Cuando la lista de valores que tenemos corresponde a una muestra, en vez de calcular la Varianza, σ^2 = SC/N, calculamos la Cuasivarianza o Varianza Muestral: S^2 = SC/(N−1) y en vez de calcular la Desviación Estándar, σ = (SC/N)½, calculamos la Desviación Estándar Muestral: S = {SC/(N-1)}½. En el capítulo 9 se explica por qué la Desviación Estándar Muestral se calcula dividiendo la SC por N−1, en vez de dividirla por N. El resto de medidas se calculan como se vio anteriormente.

En la muestra se calcula también el *Error Estándar* o *Típico*:

ES = S/(N)½ = [SC/N (N-1)]½, que va a tener especial protagonismo cuando intentamos adquirir conocimiento sobre una población a partir de la muestra.

EJEMPLO 2.1

Se toma una muestra de N = 7 varones adultos sanos y en cada uno se mide la frecuencia cardiaca, X, obteniéndose:

© TODOS LOS DERECHOS RESERVADOS

Resultados, X	80	83	85	78	76	86	72	$\rightarrow \Sigma = 560$
d= (X–Media)	0	3	5	-2	-4	6	-8	$\rightarrow \Sigma = 0$
d^2	0	9	25	4	16	36	64	$\rightarrow \Sigma = 154$

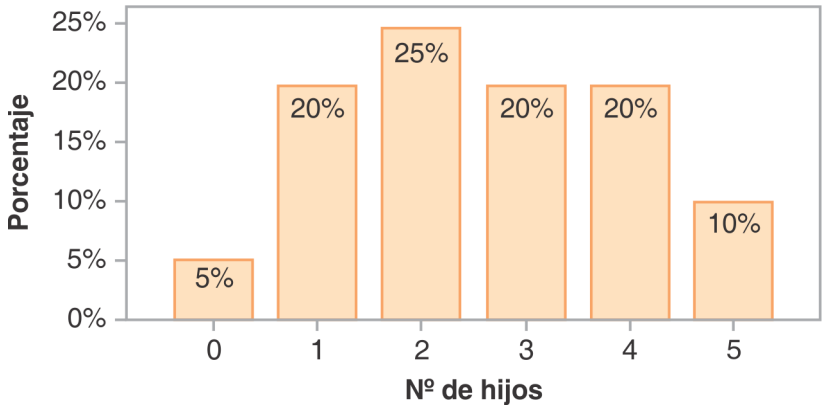
- Media = 560/7 = 80
- SC = 0+9+25+4+16+36+64 = 154
- Cuasivarianza: $S^2 = 154/6 = 25.67$
- Desviación Estándar Muestral: $S = (25{,}67)^{1/2} = 5{,}06$
- Error Estándar: $ES = S/(n)^{1/2} = 5{,}06 / 7^{1/2} = 1.91$

2.3. CÁLCULO DE PROMEDIOS EN DISTRIBUCIONES DE FRECUENCIA

Si la lista de valores sobre la que queremos calcular promedios contiene varios valores que se repiten varias veces, es más eficiente tratar esa información en forma de Distribución de Frecuencias. Las fórmulas para la Media, Suma de Cuadrados, Varianza, etc., son esencialmente las mismas que en el caso general, pero teniendo en cuenta que si, por ejemplo, el valor 3 se repite 60 veces, en vez de sumarlo esas 60 veces es más cómodo escribir y calcular 3 x 60.

EJEMPLO 2.2

En un grupo de 240 individuos se cuenta para cada uno el Nº de hijos.



© Todos los derechos reservados

Xi:	**0**	**1**	**2**	**3**	**4**	**5**	**Valores de la variable**
n_i:	12	48	60	48	48	24	Frecuencias Absolutas
f_i:	.05	.20	.25	.20	.20	.10	Frecuencias Relativas
F_i:	.05	.25	.50	.70	.90	1.00	Frecuencias Acumuladas

Medidas de posición

- **Moda.** Clase con mayor frecuencia, 2 hijos.
- **Percentil** 90 = 4 (el 90% tiene 4 hijos o menos).
- **Mediana.** Es el percentil 50, 2 hijos.

$$\text{Media aritmética: } M = \frac{\sum X_i}{N},, \; M = \frac{\sum n_i X_i}{\sum n_i},, \; M = \sum f_i X_i$$

M = (12x0 + 48x1 + 60x2 + 48x3 + 48x4 + 24x5)/240 = 624/240 = **2.6**

M =.05x0 + .20x1 + .25x2 + .20x3 + .20x4 + .10x5 = 0 + .20 + .50 + .60 + .80 +.50 = **2.6**

Para calcular el total de hijos se multiplica cada valor por el número de veces que se repite, es decir, por su frecuencia absoluta y se suman todos esos productos, es decir, 0x12 + 1x48 + 2x60 +... = 624. La media es ese total dividido por el número de personas. Pero si no se conocen las frecuencias absolutas, se puede calcular la media multiplicando cada valor por su Frecuencia Relativa y sumando esos productos (0x0.05 + 1x0.20 + 2x0.25 +...). Ahora ya no hay que dividir por el total porque ya se ha dividido cada frecuencia absoluta para calcular la relativa.

2.4. VALORES ESTANDARIZADOS

Cuando se evalúa una variable cualitativa su distribución se concreta en la proporción de casos que presentan cada valor. Esa proporción nos da la probabilidad de encontrar individuos con ese valor en las muestras. Decimos, por ejemplo, que el 20% de los individuos son del grupo sanguíneo 'A', el 18% son del 'B', etc.

Con variable cuantitativa continua no cabe hablar de proporción de cada valor porque hay infinitos valores posibles y cada uno de ellos –considerado con total precisión– tiene proporción "casi–cero". Lo que se hace es evaluar la proporción de valores comprendidos en *ciertos intervalos*. Diremos, por ejemplo, que el 27% de los individuos tienen Tensión Arterial, TA, entre 100 y 120, o que el 15% de los individuos tienen TA menor de 98, etc.

Es muy útil dar los intervalos usando los llamados "**Valores Estandarizados**", que indican el número de desviaciones estándar que se aleja de la media cada valor de la variable. Por ejemplo,

© Todos los derechos reservados

si un conjunto de valores tiene Media μ = **160** y Desviación Estándar σ = **10**, el valor 180 está 20 unidades por encima de la media, esto es, 2 desviaciones estándar por encima de la Media, y decimos que su valor estandarizado es Z = 2. Y el valor 150 está 10 unidades por debajo de la media, esto es 1 desviación estándar por debajo de la media y decimos que su valor estandarizado es Z = –1. Y el valor 168 está 8/10 = 0.8 σ por encima de μ y decimos que su valor estandarizado es Z = 0.8.

En general, el valor estandarizado se calcula:

$$Z = \frac{x - \mu}{\sigma}$$

Para X = 170 es Z = **1**, para X = 175 es Z = **1.5**, para X = 140 es Z = –**2** y para X = 165 es Z = **0.5**. Y para esta variable es equivalente decir, por ejemplo, que el 9% de los individuos tienen valor de X entre 170 y 175 a decir que el 9% de los individuos tienen valor estandarizado entre 1 y 1.5: P (170< X <175) = P (1<Z<1.5) = 0.09.

© Todos los derechos reservados

2.5. LA MEDIA PONDERADA

Para introducir este concepto consideremos los siguientes ejemplos.

Ejemplo 2.3

Un alumno ha hecho en una asignatura 5 exámenes parciales de importancia similar con estas notas 7, 7, 7, 7 y 2, y el profesor le dice que la nota media del curso es la media entre 7 y 2, es decir (7+2)/2 = 4.5. El alumno hará notar que no ha tenido un 7 y un 2, sino cuatro 7 y un 2, por lo que la media es M = (7+7+7+7+2)/ 5 = (7 · 4+2 · 1)/(4+1) = 30/5 = 6.

Decimos que la nota “7” se ponderó por 4, porque hay cuatro exámenes con ella, mientras que la nota 2 solamente se presentó una vez.

Ejemplo 2.4

En una familia de 6 personas hay 4 mujeres, cada una con US8 y 2 varones, cada uno con US1. Si alguien dice que la media de dólares por persona es (8+1)/2 = US4.5, le corregimos haciendo notar que hay 4 personas con US8 y 2 personas con US1, por lo que la media es M = (8 · 4+1 · 2)/(4+2) = 34/6= 5.67.

Decimos que la cantidad “8” se ponderó por 4, porque hay cuatro personas con ella, mientras que la cantidad “1” solamente se presentó 2 veces.

EJEMPLO 2.5

Si en una fábrica hay 20 obreros con sueldo de US100 y 5 obreros con sueldo de US300, la media de sueldo en la empresa no es (100 + 300)/2 = 200, sino M = (100 · 20+300 · 5)/(5+20) = 3 500/25 = 140.

En general se llama *media ponderada* de dos cantidades que se repiten distinto número de veces a la suma teniendo en cuenta el número de veces que se repite cada una, dividida por el número total de individuos. Es la media total de todo el colectivo.

Si lo que se evalúa en dos grupos es la proporción de individuos con cierta característica, la proporción del colectivo formado por la fusión de los dos grupos es una media de las dos proporciones, ponderada cada una por el tamaño de su grupo.

En la siguiente tabla vemos que en un grupo de 8 mujeres enfermó 1, proporción = 1/8 = 0.125 y en un grupo de 22 varones enfermaron 11, proporción = 0.50.

	Enfermos	Total		
Mujeres	1	8	→	P(Enf | Muj) = 1/8 = 0.125
Varones	11	22	→	P(Enf | Var) = 11/22 = 0.500
Todos	12	30	→	P(Enf) = 12/30 = 0.400

La proporción en el total es 12/30= 0.40 y puede escribirse como la media de las dos proporciones anteriores ponderadas por el tamaño de las muestras.

Proporción total= 12/30 = (1+11)/(8+22) = (8 · 1/8 + 22 · 11/22)/(8+22) = **(8 · 0.125 +22 · 0.50)/(8+22)** = 0.40.

En todos los ejemplos anteriores se ve, y es una verdad general, que la media ponderada de dos cantidades es siempre un valor comprendido entre ellas y más cercano al que recibe mayor peso.

Cuando las cantidades por las que se pondera son los tamaños de dos grupos implicados, la media ponderada es la media del conjunto formado por la fusión de los dos grupos en un solo.

2.6. ASPECTOS A RECORDAR

- Con variables cuantitativas, además de resumir la información agrupando los valores en intervalos y calculando la distribución de frecuencias, DF, de estos, se pueden calcular las llamadas **medidas** de **posición** (o centralización) y de **dispersión**.

© TODOS LOS DERECHOS RESERVADOS

- Las **medidas** de **posición** indican en torno a qué valores están los datos: La media aritmética nos dice *"a cuanto toca cada individuo si el total se reparte por igual entre todos ellos"*. La media se puede calcular a partir de la DF relativas, sin conocer ni las frecuencias absolutas ni el tamaño del grupo.
- Otras medidas de posición son la Moda (el valor más repetido) y la Mediana (el percentil 50).
- Las **medidas** de **dispersión** dan idea de lo agrupados o dispersos que están los datos entre sí. El Rango o Amplitud es la diferencia entre el mayor y el menor valor. La Desviación Media es la media del valor absoluto de las distancia de cada valor a la media aritmética. La Suma de Cuadrados es la suma de los cuadrados de esas distancias. La Varianza es Suma de Cuadrados dividida por el número de individuos, es decir, la media de esas distancias elevadas al cuadrado. La Desviación Estándar o Típica es la raíz cuadrada de la Varianza.
- En la muestra se calculan la Cuasivarianza, Desviación Estándar Muestral y Error Estándar.
- El valor estandarizado, Z, indica el número de desviaciones estándar que cada valor de la variable se aleja de la media.
- La media ponderada de dos valores es un valor comprendido entre esos dos valores.
- La media de dos medias muestrales ponderadas por el tamaño de las muestras es la media del conjunto formado por la fusión de las dos muestras en un solo grupo.
- La media de dos proporciones muestrales ponderadas por el tamaño de las muestras es la proporción del conjunto formado por la fusión de las dos muestras en un solo grupo.

© Todos los derechos reservados

© Todos los derechos reservados

Resumen

© Todos los derechos reservados

© Todos los derechos reservados

Estadística descriptiva. Relación entre dos variables

3

Visión general

- Definición de variables dependientes e independientes.
- Evaluación de la relación entre dos variables (cualitativas, cuantitativas).
- Evaluación de independencia entre las variables.
- Factores y variables respuesta.

Objetivo

- Establecer la relación entre dos variables como parte de la estadística descriptiva.

© Todos los derechos reservados

3.1. CÓMO SE ESTUDIA LA RELACIÓN ENTRE DOS VARIABLES

En los capítulos anteriores estudiábamos el comportamiento de una variable en un grupo de individuos resumiéndolo mediante medias o proporciones. Pero muy raramente el análisis de una investigación se limita a estudiar el comportamiento de cada variable por separado. Lo habitual es estudiar relaciones entre variables, viendo cómo los valores de una son distintos en los individuos con diferentes valores de otra.

Comenzaremos por explicar en qué consiste el estudio de la **relación entre dos variables**. Se trata de un principio fundamental en el análisis de datos y todo investigador debe tenerlo perfectamente claro antes de adentrarse en los modos concretos de ejecutarlo.

La idea básica es **siempre** la misma: estudiar el comportamiento de una de las variables (mediante medias o proporciones) en los distintos niveles de la otra. Si el comportamiento de una de las variables es el mismo en los distintos niveles de la otra, decimos que los valores de la primera no dependen de la otra, o que las variables son Independientes. Si el comportamiento de

una de las variables no es el mismo en los distintos niveles de la otra, decimos que los valores de la primera dependen de la otra, o que esas variables son dependientes.

Esta idea tan sencilla preside continuamente el análisis de la información, cualquiera que sea la técnica concreta que se use y el nivel de complejidad de la misma. Como ejemplo sencillo podemos considerar el estudio de la relación entre sexo y cáncer de pulmón (CP) en un colectivo. Describir dicha relación es reportar la FR de CP en los varones y en las mujeres. Si esa FR es la misma en ambos géneros, decimos que la incidencia de CP es independiente del sexo. Si el CP es más frecuente en un sexo decimos esa incidencia depende del sexo.

Ejemplo más sofisticado podría ser un "Análisis de supervivencia con regresión múltiple de Cox". Esta técnica contiene cierta complejidad conceptual y matemática.

Pero el análisis termina siempre diciendo cuanto mayor es cierto coeficiente (el "Hazard" de Cox) en un sexo que en otro, o en jóvenes que en viejos o en los pacientes con un tratamiento que en los que no lo han recibido. Es decir, aquí también se cumple que **analizar la relación entre dos variables es estudiar el comportamiento de una de ellas en los distintos niveles de la otra**. En la práctica esto se hará de distinta forma dependiendo del tipo de variables implicadas.

Veremos que:

- Si las dos variables son cualitativas construiremos lo que llamaremos *Tablas de Contingencia*, y en ellas calcularemos y compararemos las *Distribuciones de Frecuencias Conjuntas, Marginales y Condicionadas*.
- Si una variable es cualitativa y la otra cuantitativa compararemos las medias y desviaciones de la cuantitativa para cada categoría de la cualitativa.
- Si las dos variables son cuantitativas haremos nubes de puntos y cálculos de *Regresión* y *Correlación*.

© Todos los derechos reservados

3.2. RELACIÓN ENTRE DOS VARIABLES CUALITATIVAS

Veremos esta situación sobre un ejemplo concreto. De cada uno de los individuos de un grupo de 2.000 pacientes con cáncer de pulmón, sabemos si tuvo o no fiebre inicial (Fiebre = 'F+' y No fiebre = 'F–') y si aparecieron o no metástasis a otros órganos (Sí metástasis = 'M+' y No metástasis = 'M–'). Esa información estaría inicialmente recogida en una tabla como la de la izquierda:

Nº de individuo	Fiebre	Metástasis
1	No	Sí
2	Sí	No
...	...	...
2000	Sí	No

DF de la fiebre		
Valor	n	FR
Sí	600	0.30
No	1400	0.70
Total	2000	1

DF de las metástasis		
Valor	n	FR
Sí	1100	0.55
No	900	0.45
Total	2000	1

Ya sabemos cómo resumir la información de cada variable en este grupo: calculando su DF. Así, en la tabla del centro vemos que de los 2 000 pacientes, 600 tuvieron fiebre (un 30%) y 1400 no la tuvieron (un 70%) y en la tabla de la derecha vemos que 1100 presentaron metástasis (55%) y 900 no lo presentaron (45%). Pero esta información, aunque útil, no dice si fiebre y metástasis están o no relacionadas. La relación se ve calculando la FR de casos con metástasis en los individuos con fiebre y en los individuos sin fiebre. Para ello comenzamos por anotar los recuentos en una tabla de dos dimensiones:

DF conjuntas y marginales absolutas

La **DF Conjunta absoluta** se hace dando en una tabla de doble entrada la cantidad de individuos que hay en cada casilla. Sumando las frecuencias conjuntas se obtienen las **DF Marginales absolutas**.

	M+	M–	Total
F+	120	480	**600**
F–	980	420	**1400**
Total	**1100**	**900**	2000

DF conjuntas y marginales relativas

La **DF Conjunta y Marginales Relativas**, se obtienen dividiendo la frecuencia absolutas por el total de individuos estudiantes (2000 pacientes).

	M+	M–	Total
F+	0.06	0.24	**0.30**
F–	0.49	0.21	**0.70**
Total	**0.55**	**0.45**	1

Vemos que de los 2 000 pacientes, 120 tuvieron fiebre y metástasis (un 6%), 480 tuvieron fiebre y no metástasis (24%), 980 tuvieron metástasis y no fiebre (49%) y 420 no tuvieron ninguna de las dos (21%). Estas 4 cantidades forman la DF conjunta. Hubo 600 pacientes con fiebre (con o sin metástasis), lo que supone el 30% del total, y 1.400 (70%) no la tuvieron, ambas forman la DF marginal de la fiebre. De la misma forma, 1.100 pacientes (55%) no presentaron metástasis y 900 (45%) sí la presentaron, lo que forma la DF marginal de metástasis.

DF de fiebre condicionada a metástasis

Dividiendo las frecuencias conjuntas por las marginales de Metástasis obtenemos las **Frecuencias Condicionadas a Metástasis**.

	M+	M–
F+	120/1100 = **0.11**	480/900 = **0.53**
F–	980/1100 = **0.89**	420/900 = **0.47**
	1	1

Así, habían tenido fiebre inicial el 11% de los pacientes que hicieron metástasis y, por tanto, no tuvieron fiebre el 89% de ellos. De los que no hicieron metástasis, tuvieron fiebre el 53% y no la tuvieron el 47%.

© Todos los derechos reservados

DF de metástasis condicionada a fiebre				
Dividiendo las conjuntas por marginales de Fiebre obtenemos **Frecuencias Condicionadas a Fiebre**.		M+	M–	
	F+	120/600 = **0.20**	480/600 = **0.80**	1
	F–	980/1400 = **0.70**	420/1400 = **0.30**	1

Así, hicieron metástasis el 20% de los que habían tenido fiebre y, por tanto, no hicieron metástasis el 80% de ellos. Hicieron metástasis el 70% de los que no habían tenido fiebre y no hicieron metástasis el 30% de ellos.

Continuamos estudiando la relación entre dos variables tomando como ejemplo la presencia o no de dos enfermedades 'A' y 'B'. Debemos tener en cuenta que conocer P (A) y P (B) y la proporción de casos con cada una de las enfermedades:

I. **No** permite conocer cómo es la relación entre A y B, es decir, si la presencia de una favorece o estorba la presencia de la otra,

II. **No** permite conocer P (A y B), es decir, la proporción de individuos que tienen ambas enfermedades.

Veamos ambos puntos con ejemplos:

I. Si sabemos que en un colectivo de 2.000 personas 600 padecen la enfermedad A (el **30%**) y 800 padecen la enfermedad B (el **40%**), no podemos deducir si B es más frecuente en las personas con A que en las que no tienen A. Véalo en estos 4 supuestos, todos con P (A) = **0.30** y P (B) = **0.40**.

1º	A	A–			DF de A Condicionada a B
B	500	300	800	→	P(A | B) = 500/800 = 0.625
B–	100	1100	1200	→	P(A | B–) = 100/1200 = 0.083
	600	1400	2000		**P(A) = 600/2000 = 0.30**

Tienen A el 62.5% de los que tienen B y el 8.3% de los que no tienen B. Es **más** probable tener A cuando se tiene B. Fuerte dependencia directa.

2º	A	A–			DF de A Condicionada a B
B	100	700	800	→	P(A | B) = 100/800 = 0.125
B–	500	700	1200	→	P(A | B–) = 500/1200 = 0.417
	600	1400	2000		**P(A) = 600/2000 = 0.30**

Tienen A el 12.5% de los que tienen B y el 41.7% de los que no tienen B. Es **menos** probable tener A cuando se tiene B. Fuerte dependencia inversa.

© Todos los derechos reservados

3º	A	A–			DF de A Condicionada a B
B	240	560	800	→	P(A | B) = 240/800 = 0.30
B–	360	840	1200	→	P(A | B–) = 360/1200 = 0.30
	600	1400	2000		**P(A) = 600/2000 = 0.30**

Tienen A el 30% de los que tienen B y el 30% de los que no tienen B. Es **igual** de probable tener A cuando se tiene que cuando no se tiene B. Decimos que hay **Independencia** entre ellas.

4º	A	A–			DF de A Condicionada a B
B	0	800	800	→	P(A | B) = 0/800 = 0
B–	600	600	1200	→	P(A | B–) = 600/1200 = 0.50
	600	1400	2000		**P(A) = 600/2000 = 0.30**

Tienen la enfermedad A el 0% de los individuos que tienen B y el 50% de los que no tienen la B. No es posible tener A si se tiene B. Son características Incompatibles (máxima dependencia inversa).

En todos estos ejemplos se ve que el valor de P(A) = 0.30 está comprendido entre P (A| B) y P (A| B–). En la tercera tabla en que es P (A| B) = P (A| B–) el valor de P (A) es igual a los dos condicionados. Esto es una verdad general: el valor marginal está comprendido entre los condicionados y cuando estos son iguales entre sí, el marginal es igual a ellos.

Vemos que la dependencia se puede estudiar comparando las DF de A condicionadas a B, o las DF de B condicionadas a A, llegando a la misma conclusión.

II. Para ver, que conocer la P (A) y P (B) **no** permite conocer la P (A y B), es decir, la proporción de individuos que tienen ambas enfermedades, calculemos P (A y B) en los cuatro colectivos de anteriores:

1º → P(A y B) = 500/2000 = 0.25	2º → P(A y B) = 100/2000 = 0.05
3º → P(A y B) = 240/2000 = 0.12	4º → P(A y B) = 0/2000 = 0

Vemos que en los cuatro colectivos es P (A)= 0.30 y P (B) = 0.40 y, sin embargo, no es igual P (A y B).

3.3. INDEPENDENCIA

En la tercera tabla de arriba veíamos que la proporción de enfermos de A es la misma entre las personas con B que entre las sin B, en este caso 0.30. Es decir,

$$P(A|B) = P(A|B-)$$

A eso se llama Independencia e implica también que la proporción de enfermos de B es la misma entre las personas con A que entre las que no tienen A, en este caso 0.40.

© Todos los derechos reservados

$$P(B|A) = P(B|A-)$$

En la independencia ocurren estos dos hechos:

I. Siempre que hay independencia, y solamente cuando la hay, la P (A) marginal es igual a las P (A) condicionadas a B+ y a B–:

$$P(A) = P(A|B) = P(A|B-)$$

Es decir, la FR de casos con A en todo el colectivo es igual a la FR de casos con A entre los que tiene B y a la FR de casos con A entre los que no tiene B. Ello se deduce del hecho de que la P (A) marginal siempre es un valor comprendido entre los valores de las dos condicionadas.

Si la FR marginal es un valor entre las dos condicionadas, cuando estas son iguales entre sí (lo que llamamos independencia) tiene que ser igual a ese valor único de las condicionadas. Por la misma razón es:

$$P(B) = P(B|A) = P(B|A-)$$

II. Siempre que hay independencia, y solamente cuando la hay, la P (A y B) es el producto de la P (A) marginal por la P (B) marginal:

$$P(A \text{ y } B) = P(A) \cdot P(B)$$

Es decir, la FR de individuos en los que se dan las dos enfermedades es el producto de las FR con las que se da cada una de ellas. Para demostrar eso recordemos que en todos los casos, con dependencia o con independencia, la FR conjunta puede verse como "*proporción de una proporción*".

Por ejemplo, si del total de un colectivo tienen B el 20% y el 50% de los que tienen B también tienen A, tendrán A y B el 50% del 20%, es decir, $0.50 \cdot 0.20 = 0.10$ del total.

En general si P (B) es la FR de aquellos que tienen B, y entre los que tienen B, P(A|B) es la FR de ellos que, además, tienen A, la FR de los que tienen A y B es el producto de esas dos FR:

$$P(A \text{ y } B) = P(A|B) \cdot P(B)$$

Como vimos en el punto I, si hay independencia, y solo cuando la hay, es:

$$P(A) = P(A|B)$$

Cambiando en la expresión anterior queda:

$$P(A \text{ y } B) = P(A) \cdot P(B)$$

En el ejemplo de la tercera tabla arriba, con características independientes, era P (A) = 0.30 y P (B) = 0.40 por lo que P (A y B) tiene que ser $0.30 \cdot 0.40 = 0.12$.

© Todos los derechos reservados

En efecto, en la casilla superior izquierda se ve que tienen ambas enfermedades 240 personas, que sobre el total de 2.000 son 240/2.000 = 0.12 ó 12%.

EJEMPLO 3.1

En un colectivo fuman (F) el 40% y son asmáticos (A) el 70%. Con esa información no podemos saber qué proporción del total son fumadores y asmáticos.

Podrían tener ambas características desde solamente el 10% hasta un máximo del 40%. Tampoco podemos saber si el asma es más frecuente en los fumadores o en los no fumadores.

Podría, por ejemplo, ocurrir esto (asma más frecuente en no fumadores):

	A	A–			DF de A Condicionada a F
F	0.12	0.28	0.40	→	P(A | F) = 0.12/0.40 = **0.30**
F–	0.58	0.02	0.60	→	P(A | F–) = 0.58/0.60 = **0.967**
	0.70	0.30			

Son fumadores y asmáticos el 12%.
Tienen asma el 30% de los fumadores y el 96.7% de los no fumadores.

Pero también podría ocurrir esto (asma independiente de fumar):

	A	A–			DF de A Condicionada a F
F	0.28	0.12	0.40	→	P(A | F) = 0.28/0.40 = **0.70**
F–	0.42	0.18	0.60	→	P(A | F–) = 0.42/0.60 = **0.70**
	0.70	0.30			

Son fumadores y asmáticos el 28%.
Tienen asma el 70% de los fumadores y el 70% de los no fumadores.

3.4. RELACIÓN ENTRE UNA VARIABLE CUALITATIVA Y OTRA CUANTITATIVA

En una empresa se desea estudiar si el nivel de agresividad de 24 mujeres (clasificado en 4 categorías: nada, poco, medio o mucho) está relacionado con su rendimiento en el trabajo (medido en una escala de 0 a 60). La información se recoge en la tabla siguiente a la izquierda. El comportamiento de cada una de las variables se resume:

© TODOS LOS DERECHOS RESERVADOS

Agresividad →→ DF para cada categoría (hay 25% de mujeres en cada una).

Rendimiento →→ Media y Desviación Estándar.

Pero esta información no permite saber si el rendimiento está relacionado con la agresividad, es decir, si a mayor agresividad corresponde mayor o menor rendimiento.

Nº de individuo	Rendimiento	Nivel de agresividad
1	15	Nada
2	26	Poco
---	---	---
24	36	Mucho

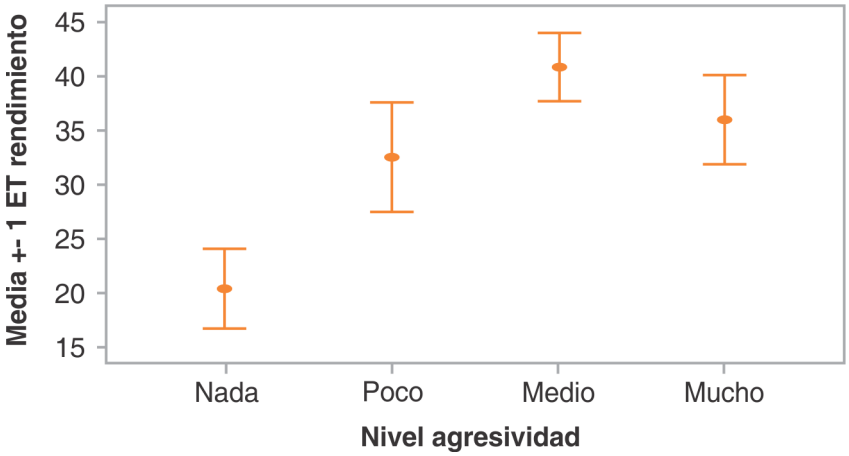
→

D. de F. de agresividad		
Nada	6	0.25
Poco	6	0.25
Medio	6	0.25
Mucho	6	0.25

Resumen de rendimiento
media = 33
DS = 11.6
ES = 2.4

Para ver la relación entre estos dos caracteres debemos comparar las medias y desviaciones del rendimiento en cada uno de los niveles de agresividad. Se puede resumir esta información representando en una gráfica la media de agresividad más/menos el Error Estándar.

Nivel de agresividad	Rendimiento en el trabajo						Media	S	ES
Nada	15	26	9	14	30	28	**22**	8.8	3.6
Poco	25	33	29	38	17	54	**33**	12.7	5.2
Medio	34	47	53	38	41	33	**41**	7.8	3.2
Mucho	35	53	38	31	23	36	**36**	9.8	4.0



Vemos que hay clara dependencia: en las tres primeras categorías a mayor nivel de agresividad hay mayor rendimiento, pero en las mujeres muy agresivas el rendimiento es algo menor que en el nivel medio. Hablaríamos de independencia si las cuatro medias fueran similares.

© Todos los derechos reservados

Esta forma de estudiar la relación entre una variable cualitativa y otra cuantitativa es la más común, pero hay otras posibilidades:

- Si la variable cuantitativa es discreta con pocos valores puede ser tratada como cualitativa. Se estudia la relación entre ella y la cualitativa haciendo Tablas de Contingencia y comparando las DF Condicionadas.
- Si la variable cuantitativa es continua se pueden agrupar sus valores en intervalos, y hacer Tablas de Contingencia cruzándola con la cualitativa. Hay que tener en cuenta que todo agrupamiento en intervalos supone una pérdida de información, ya que todos los individuos con valores de la variable en un mismo intervalo son considerados equivalentes. Por ejemplo, si se agrupan las edades por quinquenios, serán equivalentes todos aquellos que tengan 20, 21, 22, 23 y 24 años, ya que pertenecerían al mismo quinquenio. Es preferible mantener la edad como tal.

3.5. RELACIÓN ENTRE DOS VARIABLES CUANTITATIVAS

© Todos los derechos reservados

Para estudiar la relación entre dos variables cuantitativas se pueden agrupar los valores de ambas en intervalos y tratarlas como dos cualitativas (apartado 3.2) o se puede agrupar los valores de una en intervalos y ver la relación con la otra como en el apartado 3.4. Pero recordemos que esto supone cierta pérdida de información.

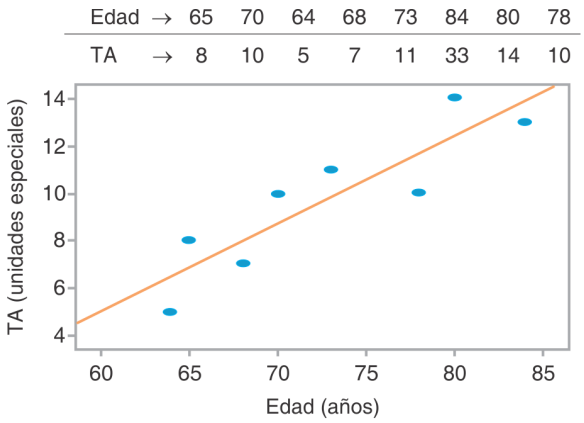
Otra opción es hacer un gráfico de dispersión o "nube de puntos" con los valores originales de las variables, sin agrupar en intervalos. Por ejemplo, para estudiar la relación entre la Tensión Arterial, TA, y la edad, en el eje horizontal se representan los valores posibles de edad y en el vertical los valores de TA.

Cada individuo está representado por un punto con su edad y su TA.

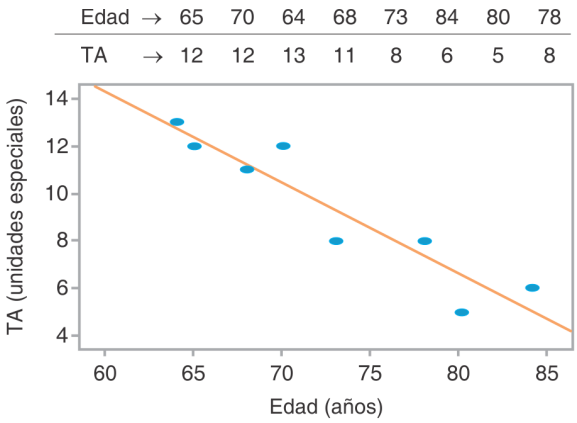
X ≡ Edad (años) Y ≡ T.A. (unid. especiales)

En cada uno de los 4 ejemplos siguientes aparecen los datos de 8 individuos.

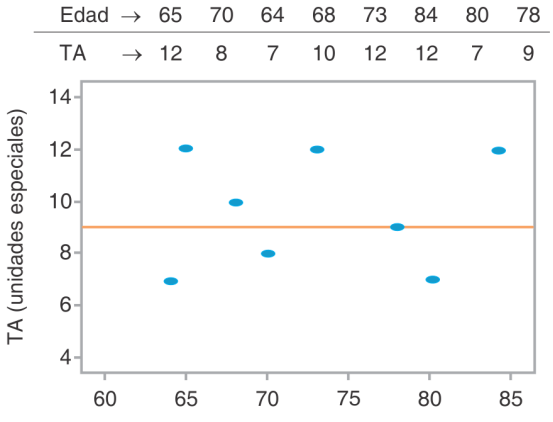
En el primero de ellos, el primer individuo tenía 65 años y TA de 8, el segundo 70 años y TA de 10... Para cada uno de los ejemplos, la representación de la nube de puntos nos permite ver la relación entre edad y TA.



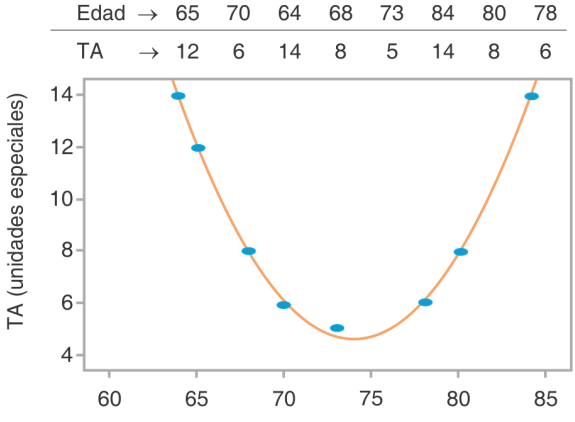
La TA aumenta linealmente con la edad: a mayor edad, mayor TA.



La TA aumenta linealmente con la edad: a mayor edad, mayor TA.



No hay relación de TA y edad: a medida que aumenta la edad, la TA no varía.



Relación parabólica: la TA disminuye hasta los 75 años y luego aumenta.

© Todos los derechos reservados

3.6. VARIABLES RESPUESTA Y FACTORES. RESPUESTA Y EFECTOS

Al estudiar la relación entre dos variables, aquella por la que se forman los estratos se suele llamar "factor", mientras que la otra, de la que se calcula la media o la FR para cada estrato de la anterior, se suele llamar "variable respuesta".

- Se le suele llamar **Respuesta** a la proporción o media de la variable respuesta en cada nivel del factor implicado.
- Se le suele llamar **Efecto** a la diferencia (o cociente) entre dos respuestas. Si, por ejemplo, se estudia la FR de enfermos en fumadores (F+) y en no fumadores (F-), el tabaco es el factor y la enfermedad es la variable respuesta. Si los resultados son los de esta tabla ("+" son enfermos), vemos que de 50 fumadores, 35 están enfermos (70%) y de 80 no fumadores, 24 están enfermos (30%):

F+		
Total	+	%+
50	35	**0.70**

F-		
Total	+	%+
80	24	**0.30**

Efecto de F
0.70–0.30 = 0.40
Sube **40** puntos

Decimos que la respuesta ha sido 30% en los no fumadores y 70% en los fumadores. Y decimos que el efecto del tabaco es incrementar la respuesta, es decir, el porcentaje de enfermos, en 40 (70–30) puntos.

Para saber cómo un factor influye en una variable no es suficiente conocer la respuesta en el grupo expuesto al factor. Esa respuesta solo es valorable al compararla con la respuesta en el grupo no expuesto al factor. Por ejemplo, para ver si la medicina "G" ayuda a curar la tuberculosis no basta con saber que en un grupo de 80 tuberculosos tratados con G curaron 32, el 40%. Esa cifra solo es informativa al compararla con el por ciento de curaciones en los no tratados. Si en un grupo de 80 no tratados curaron 60 (75%), el efecto de G es disminuir el por ciento de curados en 35 puntos.

G+		
Total	+	%+
80	32	**0.40**

G-		
Total	+	%+
80	60	**0.75**

Efecto de G
0.75–0.40 = 0.35
Baja **35** puntos

Pero si entre los 80 no tratados curaron 8 (10%), el efecto de G es incrementar el por ciento de curados en 30 (40–10) puntos.

G+		
Total	+	%+
80	32	**0.40**

G-		
Total	+	%+
80	8	**0.10**

Efecto de G
0.40–0.10 = 0.30
Baja **30** puntos

El efecto del factor también puede expresarse como cociente entre las respuestas en individuos con y sin factor, en vez de usar las diferencias. En este último ejemplo se puede decir que el por ciento de curaciones es 4 veces mayor en el grupo con G, o que G multiplica por 4 la proporción de curaciones (0.40/0.10 = 4).

Los dos modos de expresar el efecto, con diferencias o con cocientes, son igualmente válidos. En la vida cotidiana también usamos los dos modos. Si, por ejemplo, un sueldo pasa de 1.000 a 1.500 dólares, decimos que se incrementó en 500 dólares o que se multiplicó por 1.5. En adelante, salvo que expresamente se diga lo contrario, los efectos serán comentados como *diferencia* de respuesta.

Como ejemplo de variable respuesta cuantitativa supongamos que la media de la Tensión Arterial, TA, en un grupo de jóvenes es MJOV = **70** y en un grupo de viejos es MVIE = **95**. Aquí el factor es la edad y la variable respuesta es la TA. La respuesta es **70** en jóvenes y **95** en viejos. El efecto de la edad es aumentar la respuesta en **25** (95–70).

© Todos los derechos reservados

3.7. ASPECTOS A RECORDAR

- Estudiar la relación entre dos variables es estudiar el comportamiento de una de las variables (dando su media o proporción) en los distintos niveles de la otra.
- Si el comportamiento de una de las variables es el mismo en los distintos niveles de la otra, decimos que son variables independientes; en caso contrario decimos que son dependientes.
- Si las dos variables son cualitativas (o cuantitativas agrupadas en intervalos) se construyen Tablas de Contingencia. Se calcula la DF Conjunta (frecuencia de individuos con las dos características sobre el total de estudiados), las Marginales (DF de cada variable, sin tener en cuenta la otra) y las Condicionadas (DF de una característica en los distintos niveles de la otra). Si estas últimas son iguales, decimos que las características son independientes.
- En el caso de dos variables dicotómicas, A y B, la independencia consiste en que $P(A|B) = P(A|\bar{B})$, lo que implica que $P(B|A) = P(B|\bar{A})$.
- Cuando hay independencia, y solamente cuando la hay, ocurre que $P(A) = P(A|B) = P(A|\bar{B})$ y $P(B) = P(B|A) = P(B|\bar{A})$ $P(A \text{ y } B) = P(A) \cdot P(B)$.
- Si una variable es cualitativa o cuantitativa agrupada en intervalos, y la otra cuantitativa, se calculan medias y desviaciones de la segunda para cada categoría de la cualitativa. Si las medias son iguales, decimos que las variables son independientes.
- Si ambas variables son cuantitativas se representa la nube de puntos y se hacen cálculos de Regresión y Correlación. Si no existe una cierta tendencia a variar una a medida que aumenta o disminuye la otra, decimos que son independientes.
- La variable por la que se divide en categorías o estratos se le suele llamar "**factor**" y a la otra, "**variable respuesta**". Llamamos **Respuesta** a la media o proporción de la variable respuesta y llamamos **Efecto** a la diferencia (o cociente) entre dos respuestas.

© Todos los derechos reservados

Resumen

© Todos los derechos reservados

© Todos los derechos reservados

Proporciones y probabilidad. Teorema de Bayes

4

Visión general

- Definición de proporción.
- Definición de probabilidad.
- Teorema de Bayes.
- Test diagnóticos.

Objetivo

- Definición de probabilidad y proporción como parte de la inferencia estadística.

© Todos los derechos reservados

4.1. FRECUENCIAS RELATIVAS Y PROBABILIDADES

Los cuatro primeros capítulos desarrollaban la Estadística Descriptiva, que enseña a analizar la muestra obtenida en la investigación, describiendo en ella el comportamiento de cada variable y la relación entre distintas variables. A partir de aquí comenzamos a estudiar la Inferencia Estadística, IE, que es el intento de conocer acerca de las poblaciones a partir de lo encontrado en las muestras.

En la vida común y en el contexto del análisis de datos se emplea continuamente el término probabilidad y muchas veces se le da un número. En este capítulo veremos que ese número es siempre una proporción y que los cálculos y las ideas implicadas pueden ser entendidos y usados correctamente sin tener conocimientos matemáticos.

Sabemos que la FR de individuos con cierta característica se puede expresar como proporción o porcentaje o tanto por mil o por diez mil u otra cantidad pertinente.

Si de un colectivo donde son “E” (“E” como inicial de “enfermo” o de cualquier otra característica) el 90% de los individuos, nos disponemos a tomar un individuo al azar (es decir, que todos los

individuos tienen las mismas opciones de salir), no podemos saber si va a ser E o sano. Esa incertidumbre es inevitable. En esta situación hay consenso en decir que al extraer un individuo "La probabilidad de que sea E es 0.90" o resumidamente "P (E) = 0.90".

Pero ¿qué quiere decir esa frase? Aclaremos en primer lugar que ella no evita la incertidumbre respecto a cómo va a ser el individuo que vamos a extraer. Tras enunciarla seguimos sin saber si el individuo que vamos a extraer va a ser o no E.

Entonces, ¿qué indica esa expresión?

La respuesta es que no hace referencia a lo que ocurre cuando se toma un individuo de la población, sino a lo que ocurre si la extracción se repite millones de veces con reposición (el individuo obtenido en cada extracción es repuesto a la población antes de extraer el siguiente). "P (E) = 0.90" quiere decir que si se hacen millones de extracciones, serán E muy aproximadamente 90 de cada 100 extraídos.

Si en otra población la proporción de E es 0.80, diremos que al extraer un individuo al azar la probabilidad de que sea E es P (E) = 0.80, lo que indica que si se hacen millones de extracciones, serán E muy aproximadamente 80 de cada 100 extraídos. Si en esta última población repetimos millones de veces la extracción de un individuo es claro que en la primera de ellas podrá ser E (de hecho, es más fácil que sea E a que no lo sea). También puede salir E en la segunda extracción. Y en la tercera y la cuarta.... Es decir, puede salir E en cada una de las extracciones y por ello pueden serlo en todas las de esa larga serie. Por la misma razón también podrían salir no–E en todas las extracciones.

Sin embargo, la experiencia muestra que en la práctica "nunca" ocurre que sean E todos los individuos extraídos, ni que todos sean no–E, ni ningún otro resultado muy extremo. "Siempre" ocurre que si en una población hay 80% de E y se hacen millones de extracciones, son E muy aproximadamente el 80% de los extraídos. Este hecho se expresa diciendo que al tomar un individuo al azar de ese colectivo la probabilidad de que sea E es 0.80, es decir, P (E) = 0.80.

Esto es un hecho empírico, observado por el ser humano durante milenios de evolución. Se le llama Ley de la regularidad de las grandes series estadísticas o Ley empírica del azar y es de importancia capital en la vida común y en la investigación científica.

Repitamos una vez más esta idea fundamental: decir que la probabilidad de que un individuo extraído de una población sea E es 0.90 no evita la incertidumbre sobre si el individuo extraído será o no E. Estamos diciendo a qué valor se aproxima mucho la FR de individuos E si la extracción se repite millones de veces.

Note la diferencia entre la proporción de E que hay en un grupo y la probabilidad de que sea E un individuo sacado de ese grupo.

a) La proporción de E en un colectivo es un hecho físicamente evaluable contando el número total de individuos y el de los que son E. Su cociente nos da la proporción. Esta se mantiene fija mientras no se añadan o sustraigan individuos.

© Todos los derechos reservados

b) El término probabilidad se refiere a la FR de veces en que sale E cuando la extracción al azar de un individuo se repite millones de veces. La probabilidad es el valor al que se aproximará la FR a largo plazo. Ayudará a tener clara esa diferencia considerar una urna donde hay 60% de bolas blancas y 40% de negras. Si todas las bolas son del mismo tamaño, peso y textura, decimos que la probabilidad de que una bola extraída al azar sea blanca es 0.60.

Con ello queremos decir que si la extracción se repite millones de veces, en 60 de cada 100 extracciones la bola es blanca. Pero si las blancas son de doble tamaño que las negras, al tomar una bola al azar la probabilidad de que sea blanca es mayor. Si nos dicen, por ejemplo, que es P (Blanca) = 0.74, quiere decir, que al hacer millones de extracciones 74 de cada 100 bolas son blancas. Por tanto, en la urna son blancas 60 de cada 100, y al hacer extracciones lo son 74 cada 100.

4.2. MULTIPLICACIÓN DE PROBABILIDADES

© Todos los derechos reservados

Si decimos que la probabilidad de que un individuo tomado al azar de un colectivo tenga la característica "C" es, por ejemplo, P(C) = 0.6, la ley de la Multiplicación de Probabilidades dice que si tomamos, por ejemplo, tres individuos, la probabilidad de que los tres tengan C es P (C, C, C) = 0.6 x 0.6 x 0.6 = 0.63 = **0.216**.

1. ¿Qué quiere decir esa cifra "0.216"?

P = 0.216 es "216 de cada 1.000" y quiere decir que si se repite millones de veces ese proceso, es decir, si se obtienen millones de tríos, en 216 de cada mil tríos los tres individuos tendrán C.

2. ¿Por qué es la respuesta correcta?

Si sacamos millones de tríos, el primer individuo de cada trío será C, por definición, en el 60% de los tríos.

El segundo individuo será C en el 60% de los tríos cuyo primer individuo es C. Es decir, el primer y el segundo son C en el 60% del 60%, que es 0.6 x 0.6 = 0.36 o 36%.

En el 60% de los tríos en que son C el primero y el segundo lo es también el tercero. Respecto al total representa el 60% del 36%, es decir, 0.36 x 0.6 = 0.216.

Si en vez de tres el grupo es, por ejemplo, de 8, la probabilidad de que los 8 individuos sean C es el 60% del 60% del... 60%, ocho veces, es decir, 0.6 x 0.6 x...0.6= 0.68 = 0.0168. Es decir, en 168 de cada 10.000 muestras de N = 8, los ocho son C.

En general, si la probabilidad de que un individuo sea C es π, la probabilidad de que sean C todos los individuos de una muestra de N es π^N.

4.3. EL AZAR Y LA NECESIDAD

La "multiplicación de probabilidades" que acabamos de ver permite, entre otras muchas cosas, calcular la cantidad de sujetos que esperamos tengan éxito en una serie de pruebas sucesivas cuando un elevado número de ellos lo intentan.

Muchos de los hechos y/o personas que consideramos extraordinarios no son más que ejemplos de lo que ocurre *necesariamente* cuando se repite un fenómeno aleatorio un número elevado de veces.

Considere una enfermedad seria y frecuente y un tratamiento que aplicado correctamente tiene éxito solamente en el 40% de los casos. Se hace un estudio europeo que incluye a 10.000 médicos y cada uno trata un paciente cada semana.

Veamos la FR de médicos que obtienen éxito reiteradamente a lo largo de varias semanas si todos aplican correctamente el tratamiento. En la primera semana obtienen éxito unos **4.000** (el 40% de 10.000). Entre esos 4.000 tiene éxito también en la segunda el 40%, de modo que con éxito en la primera y segunda hay el 40% del 40%, es decir, $0.4 \cdot 0.4 =$ **0.16** del grupo inicial. Tienen éxito en las tres primeras semanas el 40% de los anteriores, $0.16 \cdot 0.4$, o lo que es lo mismo $0.43 =$ **0.064** del grupo inicial. Y así sucesivamente. La tabla siguiente da la FR y el número aproximado de médicos que han obtenido éxito en las primeras semanas.

Semana	Proporción de médicos con éxito en todos los pacientes tratados hasta ese momento	Número aproximado de médicos con éxito en todos los pacientes tratados hasta ese momento
1ª	0.4000	4000
2ª	$0.40^2 = 0.1600$	1600
3ª	$0.40^2 = 0.0640$	640
...	...	...
7ª	$0.40^2 = 0.0016384$	16

Por tanto, tras tratar a 7 pacientes habrá en torno a 16 médicos que han tenido éxito en todos ellos, aunque apliquen el mismo tratamiento y de la misma manera que los demás. En la oficina de seguimiento de los resultados se sabe desde el principio que si todos los médicos aplican el mismo tratamiento, va a haber entorno a 16 profesionales con éxito en los primeros 7 pacientes. No se sabe quiénes van a ser y no depende de ninguna característica de los médicos en cuestión (otra cosa sería que algunos médicos aplicaran un tratamiento mejorado y por ello obtuvieran más éxitos que otros, pero ahora no estamos contemplando esa opción). Aquí se trata de que por puro azar entorno a 16 médicos obtienen éxito en los 7 primeros pacientes.

Para subrayar que es una cuestión de azar consideremos una urna con 40% de bolas blancas y 60% negras, todas de igual tamaño, peso y textura. Si extraen 7 bolas al azar 10.000 personas, 0.40 de ellas tiene blanca 1ª bola, $0.4 \cdot 0.4 = 0.16$ del total tienen blanca las 1ª y 2ª bolas, $0.43 = 0.064$ tienen blanca las 1ª, 2ª y 3ª bolas... y en torno a 16 les salen blancas las 7 bolas. Y ello no

© Todos los derechos reservados

depende de ninguna característica de esas personas. Si el mismo grupo de 10.000 vuelve a hacer la misma operación de extraer 7 bolas al azar, de nuevo en torno a 16 de ellas les salen blancas las 7 bolas, pero estas 16 personas no serán las mismas que las de antes y no es posible saber a quién le va a ocurrir eso antes de que ocurra. Lo que sí sabemos es que será una cantidad próxima a 16, como en el caso de los médicos que van a tener éxito en los 7 primeros pacientes.

En el entorno de cada médico que obtuvo 7 éxitos seguidos se le verá como algo excepcional, puesto que lo esperado es que obtuvieran éxito en el 40% de sus pacientes, que es $7 \cdot 0.4 = 2.8$, es decir, 2 o 3 pacientes de los 7 tratados. Y se tenderá a atribuir esos buenos resultados a las más exóticas causas.

Pero en la oficina de seguimiento de los 10.000 médicos se sabía que eso era lo previsto si ninguna causa externa interviene.

Por otra parte, la FR de médicos que fracasan en los 7 primeros pacientes es $0.67 = 0.0280$ y por ello entre los 10.000 habrá entorno a 280 médicos con tan mal resultado. También en sus respectivos entornos tenderá a atribuirse esos fracasos a causas externas más o menos pintorescas. Pero nosotros sabemos que habrá aproximadamente esa cantidad de médicos con 7 fracasos, por el hecho de que el tratamiento fracasa en el 60% de los casos y hay 10.000 médicos implicados.

El ciudadano de a pie observa admirado la vertiginosa ascensión en la feria de las vanidades de algunos personajes del mundo de las finanzas. Algunos de esos triunfadores tienen especial habilidad en el oficio y/o en delinquir, pero cierto número de ellos no son genios ni delincuentes. Son solamente la consecuencia necesaria de las leyes del azar. Supongamos un millar de jóvenes ejecutivos que teniendo un capital inicial de 100.000 dólares y mucha ambición deciden arriesgarlo todo en operaciones donde hay probabilidad 0.50 de duplicar su dinero en un año y probabilidad 0.5 de perderlo todo. En el primer intento aproximadamente 500 de ellos perderán su capital inicial y los otros 500 lo duplican. Si estos últimos lo invierten todo en otra operación similar, 250 perderán todo y otros 250 tendrán 400.000 dólares. Si estos 250 repiten el proceso, aproximadamente 125 tendrán ahora 800.000 dólares. Hasta ahora los "supervivientes" se han arriesgado tres veces. Si continúan haciéndolo hasta 8, la probabilidad de salir airoso en las 8 es $0.58 \approx 0.004$. Es decir, de los 1.000 iniciales dispuestos a perder todo o duplicar en cada intento, aproximadamente 4 habrán tenido éxito en los 8 intentos, y su capital inicial se habrá multiplicado por $28 = 256$, de modo que será **25.600.000** dólares. Son los triunfadores del momento. Los otros 996 estarán arruinados. Cayeron en acto de servicio y yacen en el anonimato definitivo.

Hay que insistir en que esos casos no dependen de habilidades personales, protecciones divinas ni influencias astrológicas. Si la disyuntiva entre perder todo en cada operación o duplicarlo se dilucidara tirando una moneda bien balanceada, el resultado habría sido el mismo. Tras repetir 8 veces el juego, unos 4 jugadores entre los mil participantes habrían obtenido "cara" en todas las tiradas. Todos los demás habrían obtenido cruz en alguna y lo perdieron todo. La tabla resume estos ejemplos:

© Todos los derechos reservados

	π Prob. de éxito en cada intento	N Núm. de intentos	Prob. de éxito en los n intentos $p=\pi^N$	K Cantidad inicial de sujetos	Núm. esperado de sujetos con éxito en los n intentos E = P · K
Médicos	0.4	7	$0.4^7 = 0.0016$	10000	16
Financieros	0.5	8	$0.5^8 = 0.0039$	1000	4

4.4. PROBABILIDADES CONJUNTAS, MARGINALES Y CONDICIONADAS

Una aplicación del cálculo de probabilidades muy útil en prácticamente todas las ciencias se conoce con el nombre de Teorema de Bayes (TB). Los libros suelen explicarlo con una expresión matemática que el investigador tiene dificultad para entender. Aquí lo expondremos sin recurrir a más conocimientos previos que el concepto y cálculo de "proporción de una proporción".

Comencemos repasando los conceptos de probabilidades conjuntas, marginales y condicionadas en esta tabla que relaciona las frecuencias de las enfermedades A y B en un grupo de 2.000 personas.

	A → P(A) = 0.30	**A– → P (A–) = 0.70**	**Total**
B	**n** = 500 **P (A y B)** = 500/2000 = 0.25 **P(A | B)** = 500/800 = 0.625 **P(B | A)** = 500/600 = 0.833	**n** = 300 **P (A– y B)** = 300/2000 = 0.15 **P(A– | B)** = 300/800 = 0.375 **P(B | A–)** = 300/1400 = 0.214	**n** = 800 **P(B)** = 800/2000 = 0.40
B–	**n** = 100 **P (A y B–)** = 100/2000 = 0.05 **P(A | B–)** = 100/1200 = 0.083 **P(B– | A)** = 100/600 = 0.167	**n** = 1100 **P (A– y B–)** = 1100/2000 = 0.55 **P(A– | B–)** = 1100/1200 = 0.917 **P(B– | A–)** = 1100/1400 = 0.786	**n** = 1200 **P(B–)** = 1200/2000 = 0.60
Total	**n** = 600 **P(A)** = 600/2000 = 0.30	**n** = 1400 **P(A–)** = 1400/2000 = 0.70	**n** = 2000

Con n denotamos el número de individuos en cada una de las casillas. Así, hay 500 individuos con las dos enfermedades (A y B), 300 tienen solo B (A– y B), 100 tienen solo A (A y B–) y 1.100 no tienen ninguna (A– y B–). Por lo tanto, hay 800 personas que tienen la enfermedad B (unos con A y otros sin A) y 600 tienen la A (unos con B y otros sin B).

Los valores de la celda superior izquierda, por ejemplo, nos dicen que:

- Hay 500 personas con ambas enfermedades y ello es el 25% del total.

  Si tomamos al azar una persona, la probabilidad de que tenga ambas enfermedades, A y B, es P (A y B) = 0.25.

- Las 500 personas con A y B son el 62.5% de las personas que tienen B.

© Todos los derechos reservados

Si tomamos al azar una persona con B, la probabilidad de que tenga también A es P (A|B) = 0.625.

- Las 500 personas con A y B son el 83.3% de las que tienen A.

  Si tomamos al azar una persona con A, la probabilidad de que tenga también B es P (B|A) = 0.833.

4.5. TEOREMA DE BAYES

El Teorema de Bayes, TB, se aplica cuando:

a) Conocemos la DF marginal de la variable 'A' y la DF de la variable 'B' condicionada a cada valor de 'A'.

b) Queremos saber la DF marginal de 'B' y la DF de 'A' condicionada a cada valor de 'B'.

Como ejemplo, supongamos que respecto a las enfermedades A y B sabemos esto:

	P(A) = 0.30	**P (A–) = 0.70**	**Total**
B	P(B | A) = 0.833	P(B | A–) = 0.214	
B–	P(B– | A) = 0.167	P(B– | A–) = 0.786	
	1	**1**	

Es decir, el 30% de una población tienen la enfermedad A (el 70% no la tiene). Es la DF marginal de A. Además, tienen B el 83.3% de los que tienen A (no tiene B el 16.7% de los que tienen A, 1–.833 =.167). Tienen B el 21.4% de los que no tienen A (no tienen B el 78.6% de los que no tienen A). Es la DF de B condicionada a A. Y queremos saber la FR de individuos que tienen la enfermedad B, y la FR que tienen A entre los que tienen B y entre los que no tienen B.

	P(A) = 0.30	**P (A–) = 0.70**	**Total**
B	P(B | A) = 0.833 P(A | B) = **?**	P(B | A–) = 0.214 P(A– | B) = **?**	P(B) = **?**
B–	P(B– | A) = 0.167 P(A | B–) = **?**	P(B– | A–) = 0.786 P(A– | B–) = **?**	P(B–) = **?**

Para encontrar la solución empezamos calculando la FR de individuos con A y B (proporción de una proporción): si tienen A el 30% de los individuos y el 83.3% de los que tienen A también tienen B, la FR de individuos con A y B es 0.30 x 0.833 = 0.25, es decir, el 25% del total tienen A y B. Del mismo modo se calculan las demás frecuencias conjuntas.

© Todos los derechos reservados

	P(A) = 0.30	**P (A–) = 0.70**	**Total**
B	P(B | A) = 0.833 P(A y B) = P(A) x P(B | A) = = 0.30 x 0.833 = **0.25**	P(B | A–) = 0.214 P(A– y B) = P(A–) x P(B | A–) = = 0.70 x 0.214 = **0.15**	P(B) =
B–	P(B– | A) = 0.167 P(A y B–) = P(A) x P(B– | A) = = 0.30 x 0.167 = **0.05**	P(B– | A–) = 0.786 P(B– y A–) = P(A–) x P(B– | A–) = = 0.70 x 0.786 = **0.55**	P(B–) =

Sumando las conjuntas se obtiene las marginales de B. Así, si 25% tienen A y B, y otro 15% tienen B sin A, en total hay 40% que tienen B (con o sin A).

	P(A) = 0.30	**P (A–) = 0.70**	**Total**
B	P(B | A) = 0.833 P(A y B) = P(A) x P(B | A) = = 0.30 x 0.833 = **0.25**	P(B | A–) = 0.214 P(A– y B) = P(A–) x P(B | A–) = = 0.70 x 0.214 = **0.15**	P(B) = P(A y B) + P(A– y B) = 0.25 + 0.15 = **0.40**
B–	P(B– | A) = 0.167 P(A y B–) = P(A) x P(B– | A) = 0.30 x 0.167 = **0.05**	P(B– | A–) = 0.786 P(B– y A–) = P(A–) x P(B– | A–) = = 0.70 x 0.786 = **0.55**	P(B–) = P(A y B–) **+** P(A– y B–) = 0.05 + 0.55 = **0.60**

Una vez conocidas las marginales de B, se calculan las FR de A condicionadas a B, dividiendo las conjuntas por las marginales de B. Así, si de cada cien individuos, 40 tienen B, y de ellos 25 tienen también A, la FR de casos con A entre los que tienen B es 25/40 = 0.625. Todas las FR de A condicionadas a B se recogen en esta tabla:

	P(A) = 0.30	**P (A–) = 0.70**	**Total**
B	P(B | A) = 0.833 P(A y B) = P(A) x P(B | A) = = 0.30 x 0.833 = **0.25** P(A | B) =.25/.40 = **0.625**	P(B | A–) = 0.214 P(A– y B) = P(A–) x P(B | A–) = = 0.70 x 0.214 = **0.15** P(A– | B) =.15/.40= **0.375**	P(B) = P(A y B) + P(A– y B) = 0.25 + 0.15 = **0.40**
B–	P(B– | A) = 0.167 P(A y B–) = P(A) x P(B– | A) = 0.30 x 0.167 = **0.05** P(A | B–) =.05/.60 = **0.083**	P(B– | A–) = 0.786 P(B– y A–) = P(A–) x P(B– | A–) = = 0.70 x 0.786 = **0.55** P(A– | B–) =.55/.60 = **0.917**	P(B–) = P(A y B–) **+** P(A– y B–) = 0.05 + 0.55 = **0.60**

La información inicial dice que tienen A el 30% del total del colectivo, y hemos encontrado que tienen A el 62.5% de los que tienen B y solo el 8.3% de los que no tienen B. Si se toma un individuo al azar de la población general, la probabilidad de que tenga A es 30% (es lo que se llama probabilidad *a priori*), pero si el individuo tiene la enfermedad B, su probabilidad de tener A es 62.5%, mientras que si no tiene B, la probabilidad de que tenga A es 8.3% (son las llamadas probabilidades *a posteriori*).

© Todos los derechos reservados

4.6. TESTS DIAGNÓSTICOS

Se llama **Test Diagnóstico**, TD, para una enfermedad "E", a una prueba diagnóstica de aplicación fácil y barata que da "positivo" en la mayoría de las personas enfermas, "E+", y "negativa" en la mayoría de los no enfermos, "E–".

Lo ideal es que diera positiva en *todos* los "E+" y negativa en todos los "E–", pues de ese modo nos diría con toda seguridad si el individuo tiene o no la enfermedad. Se denomina **Sensibilidad** del test al porcentaje de resultados positivos entre los que realmente tienen la enfermedad, y **Especificidad** del test al porcentaje de resultados negativos entre los que no tienen la enfermedad. Un test diagnóstico es tanto mejor cuanto mayores son su Sensibilidad y su Especificidad.

Los cálculos implicados en los TD son una aplicación directa del Teorema de Bayes donde las dos características a estudiar son la enfermedad y el resultado del test: conocemos la distribución *a priori* de la prevalencia de la enfermedad y la distribución del resultado del test condicionada a la enfermedad (Sensibilidad y Especificidad), y el TB nos permitirá calcular el porcentaje de tests con resultado positivo o negativo (que serán las probabilidades totales) y la distribución de frecuencias de la enfermedad condicionada al resultado del test (distribución *a posteriori* de la enfermedad). Esta última nos da la probabilidad que tienen un individuo de estar enfermo o no enfermo dependiendo de si el test le dio positivo o negativo.

Algunos de los porcentajes o probabilidades utilizados suelen denominarse con los términos que a continuación aparecen:

- **Probabilidad de Falsos Positivos:** FR de tests Positivos entre los sanos.

  Es 1– Especificidad.
- **Probabilidad de Falsos Negativos:** FR de tests Negativos entre los enfermos.

  Es 1– Sensibilidad.
- **Valor Predictivo Positivo:** FR de Enfermos entre los que dieron test Positivo.
- **Valor Predictivo Negativo:** FR de Sanos entre los que dieron test Negativo.

Un test diagnóstico con un alta especificidad (pocos falsos positivos) es muy útil para confirmar una enfermedad, mientras que un test con alta sensibilidad (pocos falsos negativos) es útil para descartar la enfermedad.

4.7. ASPECTOS A RECORDAR

- Fenómeno Aleatorio, FA, es aquel en que no se puede saber si ocurrirá o no cierto resultado cada vez que se repite, pero se sabe la proporción de veces que ocurrirá ese resultado si el FA se repite muchas veces. Llamamos Probabilidad a esa FR de veces que ocurre el resultado al repetir muchas veces el FA.

© Todos los derechos reservados

- La probabilidad de que ese resultado ocurra N veces seguidas es la probabilidad de que ocurra una vez elevada a N.
- Los valores de la probabilidad informan sobre lo que ocurre al repetir un fenómeno aleatorio muchas veces o tratar con grandes conjuntos de individuos, no al hacer una sola repetición del FA o tratar con un solo individuo.
- El Teorema de Bayes, TB, surge cuando se estudia la relación entre dos variables, digamos A y B. Conocidas la DF marginal de la variable A (llamada distribución *a priori*) y la DF de B condicionada a A, el TB nos permite obtener la DF marginal de B (probabilidades totales) y la DF de A condicionada a B (llamada distribución *a posteriori*).
- En los cálculos del TB hay que obtener la DF conjunta, calculando la proporción de una proporción.
- El TB se aplica en los **Tests Diagnósticos** donde se relaciona la presencia o ausencia de enfermedad y el resultado del test, positivo o negativo.
- **Sensibilidad** del test es la FR de individuos realmente enfermos en los que el test da positivo, y **Especificidad** del test es la FR de individuos realmente sanos en los que el test da negativo.
- Conocida la prevalencia de cierta enfermedad y la Sensibilidad y Especificidad (resultado del test condicionado a la enfermedad), se obtiene el porcentaje de tests con resultado positivo y negativo y la DF de la enfermedad condicionada al resultado del test (distribución *a posteriori* de la enfermedad).
- La distribución *a posteriori* de enfermedad nos da la probabilidad que tiene un individuo de estar enfermo, dependiendo de si el test le dio positivo o negativo.
- En el contexto de los tests diagnósticos, se utilizan los términos:
  - Probabilidad de falsos positivos = FR de test positivo entre los sanos.
  - Probabilidad de falsos negativos = FR de test negativo entre los enfermos.
  - Valor predictivo positivo = FR de enfermos entre los que dieron test positivo.
  - Valor predictivo negativo = FR de sanos entre los que dieron test negativo.

© Todos los derechos reservados

Resumen

© Todos los derechos reservados

© TODOS LOS DERECHOS RESERVADOS

Distribuciones de probabilidad: binomial, poisson y normal

5

Visión general

- Definición de muestreo aleatorio.
- Distribución binomial y Distribución de Poisson.
- Definición de la media.
- Desviación de la distribución binomial.

Objetivo

- Evaluar la distribución de la probabilidad.

© Todos los derechos reservados

5.1. POBLACIONES Y MUESTRAS

Los investigadores conviven constantemente con el problema de la IE. Puesto que observan muestras que son solo una pequeña parte de la población que intentan conocer, una pregunta es constante en su actividad científica, cualquiera que sea su campo de investigación: ¿Es el hallazgo encontrado en la muestra analizada una verdad general, válida para toda la población? ¿Qué resultados son mera anécdota particular de la muestra y cuáles son válidos más allá de ella?

En la investigación científica se estudian muestras y a partir de lo observado en ellas se intenta adquirir conocimiento sobre las poblaciones a las que pertenecen.

Población: es el total de individuos con ciertas características. Ejemplos:

- La población de varones adultos sanos de la ciudad X: 16 millones.
- La población de enfermos de úlcera de estómago: 4 millones, en todo el mundo.
- La población de ratas a las que se hubiera extirpado el páncreas y dado tres miligramos de Insulina por kilo de peso cada día: el tamaño sería el número de ratas de todo el mundo sometidas a ese tratamiento.

Muestra: es el conjunto de individuos extraídos de la población que observamos y/o manipulamos en un estudio, evaluando en ellos ciertas variables que son de nuestro interés. Ejemplos:

- Para intentar conocer cómo son los hábitos de higiene bucal en la población de varones adultos sanos de la ciudad X, estudiamos una muestra de 1234 varones adultos.
- Para intentar conocer cómo es la mucosa gástrica en los enfermos de úlcera de estómago, estudiamos las biopsias extraídas en una muestra de 327 pacientes.
- Para intentar conocer si con tres miligramos por kilo de peso diarios de insulina se puede compensar la ausencia de páncreas en ratas, a una muestra de 20 se les extirpa el páncreas y se les administra esa dosis diaria de insulina.

El investigador intenta conocer acerca de las poblaciones, pero solamente puede estudiar el comportamiento de las muestras. Si prueba, por ejemplo, un nuevo método docente en un grupo de niños con la esperanza de que mejore su aprendizaje, los resultados del método se observan en ese grupo concreto de chicos, pero el interés del investigador se centra en saber si ese método es efectivo en general, en la población de niños de esa edad y circunstancias.

Y si prueba una nueva medicina en una muestra de 200 enfermos de neumonía, los resultados observados en esos pacientes tienen interés para la comunidad científica en la medida en que sean extrapolables a la generalidad de los pacientes con esa enfermedad.

En general se llaman:

- **Parámetros:** la media, varianza, porcentaje... de una variable en la población.
- **Estadísticos:** la media, cuasivarianza, porcentaje... de esa variable en una muestra.

Los parámetros y estadísticos más habitualmente analizados son:

	Población	Muestra
Variable Cuantitativa	μ, σ^2	M, S^2
Variable Cualitativa	π	p

El investigador intenta conocer los valores de los parámetros, pero solamente conoce los valores de los estadísticos. Se llama **Inferencia Estadística**, **IE**, a la disciplina que ayuda a conocer acerca de los *Parámetros* (μ, σ^2, π,...) a partir de los *Estadísticos* (M, S^2, p,...) observados en la muestra.

5.2. LA VARIABILIDAD E INCERTIDUMBRE INHERENTE A LA EXTRACCIÓN DE UNA MUESTRA

¿Hasta qué punto es la muestra parecida a la población? La respuesta a esta pregunta es clave en la **IE**. Consideremos la situación más sencilla en que en cada individuo se evalúa si tiene o no una característica dicotómica, digamos "E". Si en la población son E, por ejemplo, el 30% de los individuos, ¿cuántos de ellos lo serán en una muestra de, por ejemplo, N= 10?

© Todos los derechos reservados

Puede que en la muestra haya justamente 3 E, dando FR muestral = 3/10 = 0.3, igual a la poblacional. Pero la experiencia muestra que el valor del estadístico de la muestra puede no ser exactamente igual al del parámetro que queremos conocer.

El número de E en la muestra puede ser cualquier otro valor de 0 a 10. En el capítulo anterior decíamos que al tomar al azar **un** individuo de esa población, podía ser E o no serlo y es imposible predecir si será o no será E. Esta incertidumbre es inherente al hecho de tomar **un** individuo y no hay modo de evitarla.

Pero sabemos que si repetimos millones de veces la extracción de **un** individuo, en 30 de cada 100 extracciones este será E. Ese conocimiento lo resumimos diciendo que la prob. de que el individuo extraído sea E es 0.30: P (E) = 0.30.

Ahora se trata de tomar una muestra de 10 individuos. El número de E en la muestra puede ser cualquier valor 0, 1,... 9, 10 y es imposible predecir cuál será esa cantidad al sacar una muestra. También esta incertidumbre es inherente al hecho de tomar una muestra y no hay modo de evitarla.

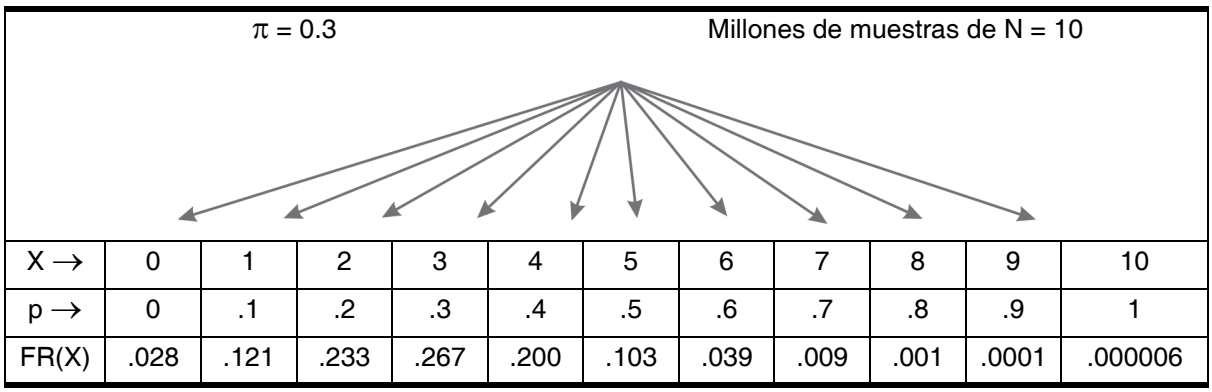
© Todos los derechos reservados

5.3. LA REGULARIDAD PROPIA DEL MUESTREO ALEATORIO, MA. LA DISTRIBUCIÓN BINOMIAL

Pero aquí también opera la Ley de la regularidad de las grandes series estadísticas o Ley empírica del azar y conduce a que si la extracción de una **muestra** de 10 individuos la repetimos **millones de veces**, resulta que:

- En 28 de cada 1000 muestras habrá 0 individuos con E (p muestral = 0.00).
- En 121 de cada 1000 muestras habrá 1 individuos con E (p muestral = 0.10).
- En 233 de cada 1000 muestras habrá 2 individuos con E (p muestral = 0.20).

En la siguiente tabla se da la FR de muestras con cada cantidad posible de E:



X →	0	1	2	3	4	5	6	7	8	9	10
p →	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
FR(X)	.028	.121	.233	.267	.200	.103	.039	.009	.001	.0001	.000006

Al hecho de extraer de una población muchos millones de muestras del mismo tamaño, se le llama hacer **Muestreo Aleatorio, MA**. Si en cada muestra calculamos un determinado estadístico,

tendremos tantos valores de él como muestras hayamos tomado. Para hacer IE es imprescindible conocer cómo se distribuyen estos valores del estadístico al hacer MA.

La Variabilidad y Regularidad propias del **MA** (millones de muestras) implica:

a) No podemos saber el número de E que va a tener cada muestra. Puede ser cualquier cantidad entre 0 y 10 (en general, entre 0 y N).

b) Sabemos, por ejemplo, que 200 cada mil muestras van a tener 4 individuos E y 103 cada mil muestras van a tener 5 individuos E..., con el ejemplo anterior.

La última fila de la tabla anterior es un ejemplo de lo que se llama **Distribución Binomial**. En general, la Distribución Binomial nos da la probabilidad de que haya X individuos "positivos" (es decir, con cierta característica) en una muestra de tamaño N, siendo π la FR de individuos positivos en la población. X puede ser 0, 1, 2... hasta N. La prob. de cada valor de X se calcula con esta fórmula:

$$P(X) = \frac{N!}{X!(N-X)!}\pi^{X}(1-\pi)^{N-X}$$

Por ejemplo, es P (4) = (10!/4! 6!) * 0.3^4 * 0.7^6 = **0.20**, es decir, el 20% de las muestras de 10 individuos tomadas de una población donde el 30% son E, tendrán 4 individuos con E (serán muestras con p = 4/10 = 40% de E).

Si se extraen millones de muestras de N = 10, en la primera de ellas puede haber 4 individuos E, y en la segunda muestra también puede haber 4 E y puede haber 4 E en cada una de las restantes muestras, por lo que podría ocurrir que saliera 4 en *todas* las muestras de esa larga serie de repeticiones. Con más motivo podría ocurrir que *ninguna* muestra tuviera 4 E. Pero la experiencia muestra que, en la práctica, la FR de muestras con 4 E "siempre" toma valores muy próximos a 0.20 y la FR de muestras con cualquier otro número de E toma valores muy próximos a los dados por esa fórmula.

Como un segundo ejemplo de distribución Binomial consideremos una población con 40% de diabéticos (π = 0.40). Si tomamos una muestra aleatoria de N = 5 individuos, teóricamente esperamos encontrar **2** diabéticos, porque 2 es el 40% de 5. A esa cantidad se la llama **Valor Esperado**, **E**, y en general se calcula como E = $\pi \cdot$ N (nótese que el símbolo "E" aquí es la inicial de "Esperado", no de "Enfermo").

Pero la cantidad de diabéticos en la muestra puede ser cualquiera entre 0 y 5. A continuación se representan los 6 tipos de muestras que pueden darse. La primera columna de la izquierda simboliza las muestras con 0 diabéticos, es decir, los 5 sanos, S. La siguiente columna indica las muestras con 1 diabético, etc. La antepenúltima *fila* recoge la cantidad de diabéticos en la muestra, X, la penúltima indica la FR de diabéticos en la muestra, p = X/N, y la última indica la FR de cada tipo de muestra en el MA, P(X), y es lo que se llama **distribución Binomial** (0.4, 5).

© Todos los derechos reservados

Tipos de muestras de N = 5 si es π = 0.40

	S	D	D	D	D	D
	S	S	D	D	D	D
	S	S	S	D	D	D
	S	S	S	S	D	D
	S	S	S	S	S	D
X:	0	1	2	3	4	5
p = X/N	0	.20	.40	.60	.80	1
P(X)	**.078**	**.259**	**.3046**	**.230**	**.077**	**.010**

S= Sano
D= Diabético

Por ejemplo, P (4) = P (X = 4) = P (p = 0.8) = (5!/4! 1!) $\cdot$ $0.40^4 \cdot 0.60^1$ = **0.077**.

EJEMPLO 5.1

a) ¿Qué es más probable: obtener dos o cuatro diabéticos?

b) ¿Qué es más probable: obtener cuatro diabéticos o cinco diabéticos?

c) Es P (2) = (5!/2! 3!) $\cdot$ $0.4^2 \cdot 0.6^3$ = 0.346 $\rightarrow$ Interprete esa cantidad.

d) Es P (4) = 0.077 $\rightarrow$ Interprete esa cantidad.

Solución:

Es una Binomial con π = 0.40 y N = 5.

a) Es más probable obtener 2 diabéticos que 4.

b) Es más probable obtener 4 diabéticos que 5 porque 4 está más cerca del valor esperado.

c) Si de una población donde hay 40% de diabéticos se toman muchas muestras aleatorias de N = 5 individuos, en 346 de cada 1.000 muestras hay 2 diabéticos y solo 2.

d) Si de una población donde hay 40% de diabéticos se toman muchas muestras aleatorias de N = 5 individuos, en 77 de cada 1.000 muestras hay 4 diabéticos y solo 4.

Las **Tablas de Binomial**, al final de la asignatura, dan los cálculos para varios valores de π y tamaños de muestra de 2 a 10. Para tamaños de muestra mayores de 10 y π distintos de los que allí aparecen, hay que calcular las probabilidades con la fórmula.

Para valores de π mayores de 50% tenga en cuenta que si, por ejemplo, en una población el 60% de los individuos son miopes, el 40% serán no miopes, y que una muestra de N = 8 con 3 miopes tiene 5 no miopes. Por ello, la prob. de que una muestra de N = 8 tenga 3 miopes es la de que tenga 5 no-miopes, es decir, P(X = 5) en una Binomial con N = 8 y π = 0.4. En la tabla se encuentra P = 0.1239.

© TODOS LOS DERECHOS RESERVADOS

5.4. ¿PARA QUÉ VALE LA REGULARIDAD PROPIA DEL MA?

Usted como estudiante se preguntará para qué vale conocer esas FR que aparecen al hacer muestreo repetido, si en la práctica el investigador toma y analiza **una** muestra, no millones de ellas. En esa **única** muestra calcula un estadístico y a partir de él intenta conocer acerca del correspondiente parámetro poblacional. ¿Por qué necesita conocer la distribución de los valores que toma ese estadístico si se extraen **millones** de muestras? Parecería que el razonamiento estadístico tiene peculiaridades extrañas muy distintas del razonamiento lógico general. Pero pronto veremos que en la vida común se procede del mismo modo y que el hecho de que la incertidumbre presente cuando sacamos *una muestra* se torne en regularidades cuando sacamos *millones de muestras* es fundamental al hacer **IE**.

EJEMPLO 5.2

En un examen Lucía ha respondido bien 20 preguntas. Si nos dice que el examen tenía 100 preguntas, el resultado parece malo. Si nos dice que tenía 22 preguntas, el resultado parece bueno. ¿Y si el examen tenía 40 preguntas? ¿Qué información complementaria necesitamos para valorar esa puntuación?

Solución:

Si el examen tenía 40 preguntas, Lucía contestó la mitad. Necesitaríamos la distribución del número de aciertos del resto de los examinados. Así podríamos conocer el percentil que corresponde a 20 aciertos.

En muchas situaciones de la vida necesitamos conocer la puntuación de todos los individuos de un colectivo para valorar la puntuación de uno de ellos. El conjunto de esas puntuaciones es lo que se llama la distribución de esa medida en ese colectivo. De la misma forma, en la IE para interpretar el valor que un estadístico tomó en una muestra necesitamos conocer la distribución de ese estadístico al hacer MA, es decir, los valores que toma si se extraen millones de muestras.

5.5. DISTRIBUCIÓN BINOMIAL EN GENERAL. NÚMERO DE VECES "+" EN UNA SERIE DE REPETICIONES DE UN FA

Hemos visto que la distribución Binomial se aplica cuando se toman al azar N individuos de un colectivo y se cuenta el número de ellos, "X", que tienen cierta característica, por ejemplo, ser enfermo.

© TODOS LOS DERECHOS RESERVADOS

Ello es un caso particular de la situación general en la que se repite N veces un Fenómeno Aleatorio y se cuenta el número de ellas en que ocurrió cierto *Suceso*.

En la situación anterior el Fenómeno Aleatorio era la extracción de cada individuo y el *Suceso* era tener o no la característica.

EJEMPLO 5.3

Se tira una moneda bien balanceada 6 veces. Cada vez que se tira, la probabilidad de que salga cara es π = 0.5. En la serie de 6 tiradas esperamos que salgan 3 caras, E = 3, pero puede salir más o menos. Veamos la probabilidad de que salgan 0, 1... 6.

$$\text{Bin}(N,\ \pi) = B(6,\ 0.5): P(X) = \frac{6!}{X!(6-X)!}\ x\ 0.5^{X}(1-0.5)^{6-X}$$

X = 0	P(0) = {6!/[0! · 6!]} 0.5^0 0.5^6	0.016
X = 1	P(1) = {6!/[1! · 5!]} 0.5^1 0.5^5	0.094
X = 2	P(2) = {6!/[2! · 4!]} 0.5^2 0.5^4	0.234
X = 3	P(3) = {6!/[3! · 3!]} 0.5^3 0.5^3	0.313
X = 4	P(4) = {6!/[4! · 2!]} 0.5^4 0.5^2	0.234
X = 5	P(5) = {6!/[5! · 1!]} 0.5^5 0.5^1	0.094
X = 6	P(6) = {6!/[6! · 0!]} 0.5^6 0.5^0	0.016

EJEMPLO 5.4

Se tira un dado homogéneo 12 veces. Cada vez que se tira, la probabilidad de que salga la cara 5 es π = 1/6 = 0.167. En una serie de 12 tiradas esperamos que salgan N · π = 12 · (1/6)= 2 cincos. Pero pueden salir más o menos de 2. Calculemos la probabilidad de que salgan 0, 1, 2, ...12 cincos:

$$\text{Bin}(N,\ \pi) = B(12,\ 1/6):\ P(X) = \frac{12!}{X!(12-X)!}(1/6)^{X}(5/6)^{12-X}$$

© TODOS LOS DERECHOS RESERVADOS

X = 0	$P(0) = \{12!/[0! \cdot 12!]\} (1/6)^0 (5/6)^{12}$	0.112
X = 1	$P(1) = \{12!/[1! \cdot 11!]\} (1/6)^1 (5/6)^{11}$	0.269
X = 2	$P(2) = \{12!/[2! \cdot 10!]\} (1/6)^2 (5/6)^{10}$	0.296
...	...	...
X = 9		0.000012
X = 12		0.0000000007

5.6. DISTRIBUCIÓN DE POISSON. BINOMIAL CON N GRANDE Y π PEQUEÑA

Distintas Binomiales con N grande y π pequeña e igual valor esperado ($E = \pi N$) dan valores de probabilidad muy parecidos, y a esa misma probabilidad se llega usando la fórmula de Poisson, que solo usa el valor de E, no N ni π. ($e = \lim_{n\to\infty} (1+1/n)^n \cong 2.71$).

$$P(X) = \frac{e^{-E} \bullet E^X}{X!}$$

© Todos los derechos reservados

Ejemplo 5.5

Veamos, por ejemplo, P(2), P(5) y P(10) en varias Binomiales de diferente N y π, pero con el mismo $E = N \cdot \pi$, en este caso E = 5:

N	π	E	P(2)	P(5)	P(10)
100	0.05	5	0.0812	0.1810	0.0177
500	0.01	5	0.0836	0.1764	0.0179
1000	0.005	5	0.0839	0.1759	0.0180
2000	0.0025	5	0.0841	0.1757	0.0181
5000	0.001	5	0.0842	0.1756	0.0181
Poisson $P(X) = e^{-E} E^X / x! = e^{-5} 5^X / X!$			**0.0842** $e^{-5} 5^2/2!$	**0.1755** $e^{-5} 5^5/5!$	**0.0181** $e^{-5} 5^{10}/10!$

Por ejemplo, en la Binomial de N = 100 y π = 0.05 es P(2) = {100!/[2! · 98!]} 0.052 0.9598 = **0.0812**. Del mismo modo se calculan las demás probabilidades.

Observamos que P(2) toma valores cada vez más próximos entre sí en las Binomiales con mayor N y menor π y ese valor se aproxima cada vez más al de la última fila, calculado con la fórmula de Poisson. Lo mismo ocurre con P(5) y P (10). Incluso en la Binomial con N = 100 y π = 0.05, la aproximación de las tres probabilidades a las de la última fila son bastante buenas para la mayoría de las aplicaciones prácticas.

EJEMPLO 5.6

Veamos P(1) y P(6) en varias Binomiales de diferente N y π, pero con la misma E = 2:

N	π	E	P(1)	P(6)
50	0.04	2	0.2706	0.0108
400	0.05	2	0.2707	0.0119
1000	0.002	2	0.2707	0.0120
Poisson $P(X) = e^{-E} E^{X}/ x! = e^{-2} 2^{X}/ X!$			**0.2707** $e^{-2} 2^{1}/1!$	**0.0120** $e^{-2} 2^{6}/6!$

Ante una Binomial con N grande y π pequeña, las probabilidades de cada valor posible: 0, 1, 2,..., es decir, la FR de muestras en el MA con 0, 1, 2... individuos "positivos", se pueden calcular con la fórmula de Poisson, que da resultados próximos a los de la Binomial y es más sencilla. Esta aproximación de la Distribución Binomial a la Poisson es bastante buena si N > 100 y π < 0.10.

Las Tablas de Poisson al final de la asignatura dan prob. de X = 0, 1, 2... para algunos valores de E.

5.7. PROBABILIDAD DE UN INTERVALO DE VALORES DE LA DISTRIBUCIÓN BINOMIAL

En muchas ocasiones al hacer **IE** hay que calcular no ya el valor de una probabilidad de la distribución Binomial, sino de cientos o miles de valores. Un ejemplo típico sería el caso de una epidemia que llega a una ciudad de N = 10.000 habitantes y cada uno tiene probabilidad π = 0.5 de caer enfermo. El número esperado de enfermos es E = 10.000 · 0.5 = 5.000. Pero en la práctica puede haber algunos más o menos de esa cantidad.

© TODOS LOS DERECHOS RESERVADOS

Si la sanidad pública tiene tratamiento para, por ejemplo, 5049 afectados, ¿cuál es la probabilidad de que haya más afectados que tratamientos y por ello algún enfermo no pueda ser tratado? La probabilidad de que haya 5050 afectados se calcula con la fórmula de la Binomial, P(X = 5050)= {10.000!/[5050! · 4950!]} · 0.55050 · 0.54950= 0.0048. Pero lo que necesitamos es la probabilidad de que haya **5050 o más** enfermos y para conocerla hay que calcular 4951 binomiales, desde X = 5050 hasta X= 10.000.

$P(X \geq 5050) = P(X = 5050) + P(X= 5051) + P(X = 5052) +...+ P(X = 10.000)$.

El resultado de tan pesado cálculo es $P(X \geq 5050)$= **0.1611**.

En Inferencia Estadística es muy frecuente este tipo de situaciones que implican cálculo muy tedioso. Pero esos cálculos se reducen a otros mucho más sencillos gracias a una propiedad de la Binomial que los matemáticos han descubierto y cuya aplicación se extiende a prácticamente todas las situaciones de la investigación, tanto en biomedicina como en el resto de las ciencias experimentales. Para exponer esa propiedad fundamental hay que empezar explicando qué son y cómo se calculan la Media y la Desviación Estándar de la distribución Binomial.

5.8. MEDIA Y DESVIACIÓN DE LA DISTRIBUCIÓN BINOMIAL

Lo explicaremos usando un ejemplo sencillo. En una población tiene Herpes, H, el 25% de las personas, $\pi = 0.25$. Hacemos MA de N = 8, y en cada muestra llamamos:

- "X" al número de individuos con H en la muestra.
- "p" a la FR de individuos con H en la muestra: p = X/8.

La tabla siguiente recoge los resultados obtenidos en las primeras muestras. Así, en la primera muestra de 8 individuos, los individuos 3º, 4º y 8º tiene H (son +), por lo que X = 3 y p = 3/8 = 37.5%; en la segunda muestra, es X = 1 y p = 1/8 = 12.5%...

© Todos los derechos reservados

	Nº de H en cada muestra										
	1º	2º	3º	4º	5º	6º	7º	8º		X	p= X/8
Muestra 1ª	–	–	+	+	–	–	–	+	→	3	.375
Muestra 2ª	–	+	–	–	–	–	–	–	→	1	.125
Muestra 3ª	+	–	–	–	–	–	+	+	→	3	.375
Muestra 4ª	–	–	–	+	–	–	+	–	→	2	.250
Muestra 5ª	–	+	+	–	+	–	+	+	→	5	.625
...	...	...	...	...	...	...	...	...		...	...
...	...	...	...	...	...	...	...	...		...	...
										↓	↓
										$\mu_X = 2$	$\mu_p = 0.25$
										$\sigma_X = 1.5$	$\sigma_p = 0.023$
										$\sigma_X = 1.22$	$\sigma_p = 0.15$

Debemos distinguir claramente las dos poblaciones implicadas:

1º **Población de individuos** → En cada persona se evalúa si tiene o no H, variable cualitativa.

2º **Población de muestras de N = 8** → En cada muestra de 8 personas se cuenta el número de ellas con H y el resultado es un valor entre 0 y 8. Por tanto, es una variable cuantitativa discreta.

Pues bien, ahora se trata de calcular cuánto vale la Media, μX, y la Desviación Estándar, σX, de esa variable cuantitativa (número de individuos con H) en esa población formada por millones de muestras de N = 8.

1. **Media y varianza de X (número de H en cada muestra)**

   El valor esperado de X es 2, ya que es el 25% de 8. En muchas muestras X será justamente 2, en otras será mayor y en otras menor de 2.

   Se demuestra que la media de X en el MA es 2. En general, $\mu X = N \cdot \pi$.

   Por tanto, en el MA la cantidad $\mu X = E = N \cdot \sigma$ es tres cosas:

   a) El valor esperado de X, que da FR muestral de H igual a la poblacional.

   b) El valor que aparece con más frecuencia en el MA. En esta distribución Binomial es P(1) = 0.2670, P(2) = 0.3115, P(3) = 0.2076...

   c) La media de X calculada sobre todas las muestras, μX.

   La σ de X en el MA es $\sigma X = \{\pi(1-\pi)N\}^{1/2} = \{0.25 \cdot 0.75 \cdot 8\}^{1/2} = 1.5^{1/2} = 1.22$.

   Es decir, si se hace MA, la media de individuos con H por muestra será 2 y la desviación estándar entre esas cantidades será 1.22.

© Todos los derechos reservados

2. **Media y varianza de p = X/N (FR de H en cada muestra)**

   Pero si en cada muestra dividimos X por N, obtenemos p y ese valor es también una variable cuantitativa de la que podemos calcular Media, μp, y desviación, σp.

   El valor esperado de p = X/N es 0.25, ya que es el de la población. En muchas muestras p será justamente 0.25, en otras será mayor y en otras menor de 0.25.

   Se puede demostrar que la media de p en el MA es 0.25. En general $\mu p = \pi$.

   Por tanto, en el MA la cantidad π es tres cosas:

   a) El valor esperado de p.

   b) El valor de p que aparece con más frecuencia en el MA.

   c) La media de p calculada sobre todas las muestras, μp.

   La σ de p en el MA es $\sigma p = \{\pi (1-\pi)/N\}^{1/2} = \{0.25 \cdot .75/8\}^{1/2} = 0.023^{1/2} = 0.15$.

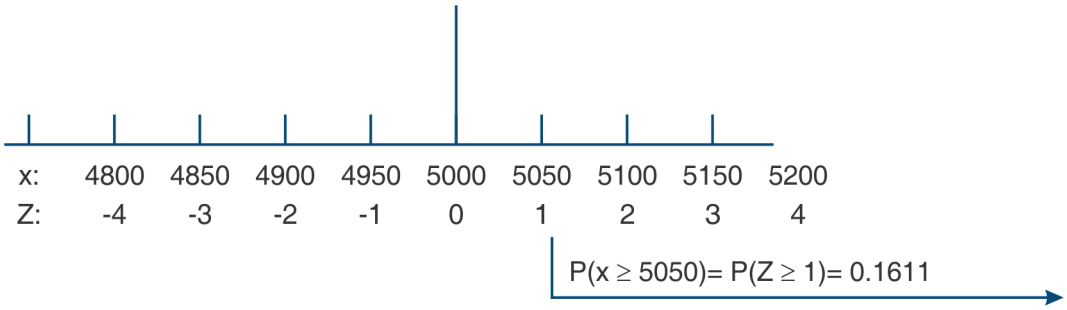
   Es decir, si se hace MA, la media de la proporción muestral de individuos con H será 0.25 y la desviación estándar entre esas cantidades será 0.15.

Estas relaciones van a ser decisivas en la Inferencia Estadística, como iremos viendo en este y los siguientes capítulos.

Como en toda variable cuantitativa, sus valores se pueden expresar como **valores estandarizados**, dando el número de desviaciones que se aleja ese valor de la media. Por ejemplo, en la Binomial de $\pi = 0.5$ y N = 10.000, es $\mu X = 5.000$ y $\sigma x = [\pi \cdot (1-\pi) \cdot N]^{1/2} = [0.5 \cdot 0.5 \cdot 10.000]^{1/2} = 50$.

Si una muestra tiene, por ejemplo, X = 5100 casos positivos, son 100 más que el valor medio, es decir, Z = (5100–5000)/50 = 2. Al valor X = 4950 le corresponde Z = –1. Al valor X = 5075 le corresponde Z = 1.5, etc.

En el apartado anterior veíamos que $P(X \geq 5050) = P(X = 5050) + P(X = 5051) + P(X = 5052) + \ldots + P(X = 10000) = 0.1611$. Por tanto, $P(X \geq 5050) = P(Z \geq 1) = 0.1611$.



© Todos los derechos reservados

5.9. EL LÍMITE DE LA BINOMIAL CON N GRANDE: DISTRIBUCIÓN NORMAL

Conocer la Media y Desviación Estándar de la distribución Binomial permite expresar la cantidad de individuos "positivos" en cada muestra como valor estandarizado y ello permitió a los matemáticos descubrir una ley definitivamente útil en Inferencia Estadística:

La probabilidad de que el número de individuos "positivos" esté en cualquier intervalo que nos interese depende únicamente de cuantas desviaciones se alejan de la media los extremos de ese intervalo.

La explicaremos usando ejemplos concretos:

Como decíamos en los apartados anteriores, si en una Binomial con $\pi = 0.5$ y N = 10.000 queremos calcular $P(X \geq 5050) = P(X = 5050) + P(X = 5051) + P(X = 5052) + \ldots + P(X = 10.000)$ tenemos que calcular 4951 binomiales.

Y si en una Binomial con $\pi = 0.1$ y N = 40.000 queremos saber $P(X \geq 4060) = P(X = 4060) + P(X = 4061) + P(X = 4062) + \ldots + P(X = 40000)$ tenemos que calcular casi 36000 binomiales.

Y si en una Binomial con $\pi = 0.2$ y N = 90.000 queremos saber $P(X \geq 18120) = P(X = 18120) + P(X = 18121) + P(X = 18122) + \ldots + P(X = 90.000)$ tenemos que calcular casi 72000 binomiales.

Resumamos en una tabla las probabilidades que queríamos conocer y el resultado:

π	N	Buscamos	Tenemos que calcular	Resultado
0.5	10000	$P(X \geq 5050)$	¡¡4951 binomiales!!	0.1611
0.1	40000	$P(X \geq 4060)$	¡¡35940 binomiales!!	0.1606
0.2	90000	$P(X \geq 18120)$	¡¡71880 binomiales!!	0.1589

Observamos que las tres probabilidades buscadas tienen valores muy parecidos. Podría ser una curiosa coincidencia. Pero si calculamos la media, μX, y la desviación, σX, de esas Binomiales, vemos que en los tres supuestos el número de casos positivos implicado se aleja una desviación de la media de ellos en el MA, es decir, su valor estandarizado es Z = 1. La tabla que sigue resume estos datos.

π	N	μ_X	σ_X	
0.5	10000	**5000**	**50**	$X = 5050 \rightarrow Z = 1$
0.1	40000	**4000**	**60**	$X = 4060 \rightarrow Z = 1$
0.2	90000	**18000**	**120**	$X = 18120 \rightarrow Z = 1$

© Todos los derechos reservados

Los matemáticos demuestran que en toda Binomial con N grande la probabilidad de obtener una muestra con número de casos "positivos" igual al valor esperado más una desviación estándar o aun mayor, se aproxima mucho a 0.1587, es decir, $P(X \geq \mu X + 1\sigma X) = P(Z \geq 1) \rightarrow$ se aproxima mucho a 0.1587.

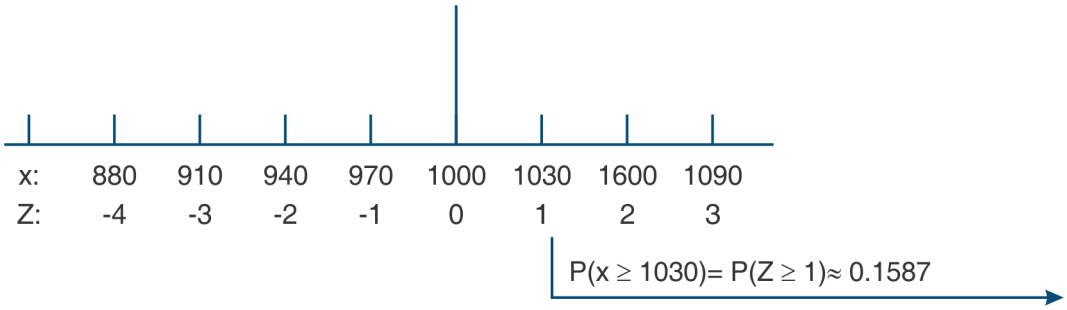
Pero esa convergencia de la probabilidad de toda binomial hacia un mismo valor para $P(Z \geq 1)$ ocurre para cualquier otro valor de Z. Por ejemplo, con N grande, en todas las Binomiales ocurre que:

$P(Z \geq 2)$ se aproxima a 0.0228 y $P(Z \geq 3)$ se aproxima a 0.0013.

Lo vemos en la tabla siguiente para $\pi = 0.50$ y N = 400, N = 10.000 y N = 40.000.

π	N	μ_X	σ_X	$P(Z \geq 2)$	$P(Z \geq 3)$
0.5	400	**200**	**10**	$P(X \geq 220) = 0.0251$	$P(X \geq 230) = 0.0016$
0.5	10000	**5000**	**50**	$P(X \geq 5100) = 0.0233$	$P(X \geq 5150) = 0.0015$
0.5	40000	20000	100	$P(X \geq 20200) = 0.0231$	$P(X \geq 20300) = 0.0014$
				$\rightarrow\rightarrow$ 0.0228	$\rightarrow\rightarrow$ **0.0013**

Ello permite saber muy aproximadamente la FR de muestras con número de casos "positivos" comprendido en cualquier intervalo, sin tener que calcular cientos o miles de veces la fórmula binomial. Por ejemplo, si en una población son enfermos el 10% y tomamos una muestra de N = 10.000, la cantidad esperada de enfermos en ella es E = 1.000. Si queremos saber la probabilidad de que en la muestra haya más de 1030 enfermos habría que aplicar la Binomial 8971 veces y sumar los resultados. Pero la desviación de esa Binomial es $(0.1 \cdot 0.9 \cdot 10000)^{1/2} = 30$, por lo que a X = 1030 corresponde Z = 1. $P(X \geq 1030) = P(Z \geq 1) \approx 0.1587$.



Y como a X= 1060 corresponde Z = 2, la probabilidad de que en la muestra haya 1060 enfermos o más es $P(X \geq 1060) = P(Z \geq 2) \approx$ **0.0228** (nos evitamos aplicar la Binomial 8941 veces).

Y como a X = 1090 corresponde Z = 3, la probabilidad de que en la muestra haya 1090 enfermos o más es $P(X \geq 1090) = P(Z \geq 3) \approx$ **0.0013** (nos evitamos aplicar la Binomial 8911 veces).

© Todos los derechos reservados

Pero aun hay más. Lo mismo ocurre para cualquier otro valor de Z. La tabla que sigue da los valores a los que se aproximan todas las Binomiales con N grande para algunos valores de Z.

Z →	0.10	0.20	0.25	0.50	0.75	**1.00**	**1.50**	**2.00**	**2.50**	**3.00**
P más allá de Z	.4602	.4207	.4013	.3085	.2266	.1587	.0668	.0228	.0062	.0013

Por ejemplo, en la Binomial con π= 0.1 y N = 10.000 (μX = 1000 y σx = 30), a X = 1003 corresponde Z = (1003–1000)/30 = 0.1 y la probabilidad de que en la muestra haya 1003 o más enfermos es P (X ≥ 1003) = P (Z ≥ 0.1) ≈ 0.4602. Y a X = 1075 corresponde Z = 75/60 = 2.5 y la probabilidad de que en la muestra haya 1075 o más enfermos es P (X ≥ 1075) = P (Z ≥ 2.5) ≈ 0.0062.

A cada valor de Z corresponde una probabilidad, que está especificada en la llamada Tabla de la distribución Normal (final de la asignatura), para valores Z de 0 a 6.

Debemos tener en cuenta que esas probabilidades son simétricas respecto a la media, es decir, la probabilidad de que salgan muestras con número de casos "positivos" por debajo de una desviación bajo la media es igual a la probabilidad de que salgan muestras con número de casos "positivos" por encima de una desviación sobre la media: P (Z ≤ –1) = P (Z ≥ 1). Ello implica que la probabilidad de muestra con número de casos "positivos" por encima de la media es 0.5. Y, por tanto, P (0 ≤ Z ≤ 1) = 0.5–P (Z ≥ 1).

Veamos todo esto con ejemplos concretos:

© Todos los derechos reservados

Ejemplo 5.7

Si de una población con 25% obesos se toma una muestra de N= 300, calcular la probabilidad de que en ella haya a) Más de 89 obesos, b) menos de 65 obesos, c) entre 60 y 75, ambos incluidos y d) entre 60 y 80.

Respuesta: μX = N · π = 0.25 · 300 = 75, σX = {0.25 · 0.75 · 300}½ = 7.5. Es decir, si se hace MA, la media de individuos obesos por muestra será 75 y la desviación estándar entre esas cantidades será 7.5.

a) P (X ≥ 90) = P (Z ≥ 2) = 0.0228.

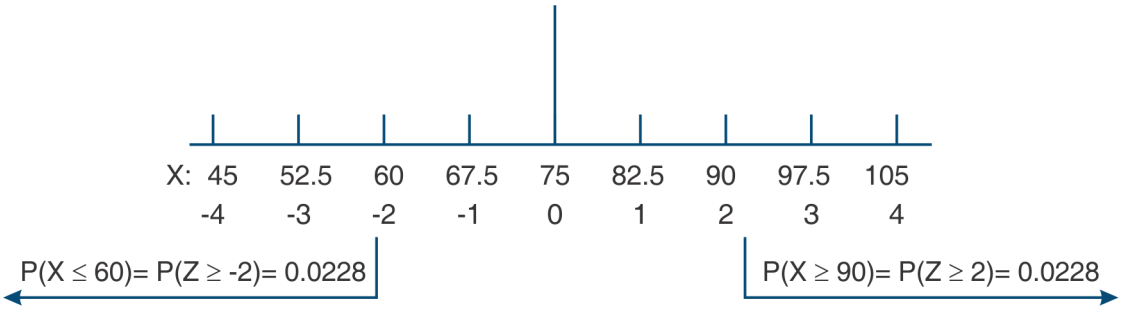
b) P (X ≤ 64) → Z = (64–75)/7.5 = –1.46 → P (Z ≤ –1.46) = P (Z ≥ 1.46)) = 0.0721.

c) Entre 60 y 75, ambos incluidos: P (60 ≤ X ≤ 75) Para X = 60 es Z = (60–5)/7.5 = –2 → P (Z ≤ –2) = P (Z ≥ 2) = 0.0228 → Como la probabilidad de 75 o menos es 0.50, tiene que ser P (60 ≤ X ≤ 75) 0.5–0.0228 = 0.4772.

d) Entre 60 y 80: Ente 60 y 75 hay 0.4772. Busquemos P (X ≥ 80) Para x = 80 es Z = 5/7.5 = 0.66 En las tablas de la distribución Normal vemos P(Z ≥ 0.66) = 0.2546. Entonces es P

$(75 \leq X \leq 80) = 0.5 - 0.2546 = 0.2454$. Por tanto: $P(60 \leq X \leq 80) = P(-2 \leq Z \leq 0.66) = P(-2 \leq Z \leq 0) + P(0 \leq Z \leq 0.66) = 0.4772 + 0.2454 = 0.7226$.

e) Entre 55 y 65: $P(55 \leq X \leq 65)$. A $X = 55$ corresponde $Z = -20/7.5 = -2.66$ y a $X = 65$ corresponde $Z = -10/7.5 = -1.33$. Por tanto $P(55 \leq X \leq 65) = P(-2.66 \leq Z \leq 1.34) = P(Z \leq 1.34) - P(Z \leq 2.66) = 0.0901 - 0.0039 = 0.0861$.



En los próximos capítulos se ve gran cantidad de cálculos de este tipo.

5.10. ASPECTOS A RECORDAR

- Población es el conjunto –generalmente indefinidamente grande– de todos los individuos con ciertas características. Muestra es el conjunto de individuos extraídos de la población para ser observados por el investigador.
- La IE intenta saber acerca de las poblaciones a partir del conocimiento de las muestras, es decir, intenta conocer acerca de los **parámetros** a partir del conocimiento de los **estadísticos**.
- Si tomamos una muestra no podemos asumir que el parámetro tenga exactamente el mismo valor que el estadístico encontrado en la muestra.
- Para hacer IE hay que conocer cómo se distribuyen, es decir, qué valores toman, los estadísticos muestrales en el **Muestreo Aleatorio**.
- La **distribución Binomial** nos da la FR de muestras de N individuos que en el MA tendrán X individuos "positivos".
- En general, se aplica **distribución Binomial** cuando se repite N veces un Fenómeno Aleatorio y se cuenta el número en que ocurrió cierto Suceso.
- La Distribución de Poisson es una distribución Binomial con N grande (>100) y π pequeña (< 0.10).
- Si de una población con FR de enfermos π se hace MA de tamaño N:
  - La media del número de enfermos por muestra, X, es $\mu X = E = N \cdot \pi$.
  - La varianza de las X es $\sigma 2_X = \pi (1-\pi) N$.

© Todos los derechos reservados

- La media de la proporción de enfermos por muestra, p, es $\mu p = \pi$.
- La varianza de las p es $\sigma 2_p = \pi\ (1-\pi)/\ N$.

- Ejemplo: si de una población con 40% de E se hace MA de N = 6 es:
  - $\mu p = 0.40$, $\sigma 2$.
  - $p = 0.04$, $\mu X = E = 2.40$, $\sigma 2$.
  - $X = 1.44$.
- Con N grande la probabilidad de obtener X o más casos en la muestra depende solamente de Z, número de σX que se aleja X de la media: $Z = (X-\mu X)/\sigma X$.
- $P\ (X \geq \mu X + 1\sigma X) = P\ (Z \geq 1) \rightarrow$ se aproxima mucho a 0.1587.
- $P\ (X \geq \mu X + 2\sigma X) = P\ (Z \geq 2) \rightarrow$ se aproxima mucho a 0.0228.
- $P\ (X \geq \mu X + 3\sigma X) = P\ (Z \geq 3) \rightarrow$ se aproxima mucho a 0.0013.
- Las tablas llamadas de **distribución Normal** dan los valores de $P\ (Z \geq 0.02)$, $P\ (Z \geq 0.04)$, $P\ (Z \geq 0.06)$... hasta $P\ (Z \geq 4)$, $P\ (Z \geq 5)$, $P\ (Z \geq 6)$.

© Todos los derechos reservados

© Todos los derechos reservados

Resumen

© Todos los derechos reservados

© Todos los derechos reservados

Inferencia estadística con una proporción

6

Visión general

- Definición de la distribución normal.
- Distribución de las medias muestrales de un muestreo aleatorio.
- Teorema central del límite.

Objetivo

- Conocer la definición del Teorema Central del Límite y de la Distribución Normal.

© Todos los derechos reservados

6.1. INFERENCIA ESTADÍSTICA CON UNA PROPORCIÓN

Los investigadores conviven constantemente con el problema de la Inferencia Estadística, IE ¿Es el hallazgo encontrado en la muestra analizada una verdad general válida para toda la población? ¿Qué resultados son mera anécdota particular de la muestra y cuáles válidos más allá de ella? Puesto que observa muestras que son solo una pequeña parte de la población que intenta conocer, esa pregunta es constante en su actividad científica, cualquiera que sea su campo de investigación.

Desde comienzos del siglo XX la Estadística pone al servicio de los investigadores dos herramientas para ayudarle en ese intento de extrapolar de la muestra a la población: los **Tests Estadísticos** y los **Intervalos de Confianza**. Pero el beneficio que podían proporcionar esos nuevos recursos se vio mermado por:

a) El escaso uso de los Intervalos de Confianza.

b) El uso excesivo y muchas veces incorrecto de los Tests.

En los Tests Estadísticos se usa el llamado valor P del test y la mayoría de los investigadores tienen notables dificultades para entender lo que indica ese valor y para usarlo correctamente.

Pero aunque el cálculo de ese valor P y de los intervalos de confianza es una cuestión matemática, el tema puede ser entendido y aplicado correctamente usando únicamente la lógica común del hombre de la calle. Aquí lo presentamos sin utilizar ninguna herramienta matemática.

En el capítulo anterior veíamos que conocer la FR de casos positivos en una población no permite saber la FR de ellos que habrá en una muestra sacada de ella. En este capítulo lidiamos con esa incertidumbre en sentido inverso: a partir de la FR observada en una muestra no podemos conocer la FR en la población de la que fue extraída.

Veámoslo en ejemplos muy sencillos.

El PIOS –Partido Independentista de Orejuela del Sordete– obtuvo en las últimas elecciones generales el 30% de votos (la FR poblacional en ese momento fue 0.30). En su intento de incrementar el número de seguidores deciden cambiar el nombre a PIROS –Partido Independentista Revolucionario de Orejuela del Sordete– y tras ese gran cambio encargan una encuesta sobre la intención de voto. En una muestra aleatoria de N = 100 personas resultó que 34 manifestaron intención de votarle si hubiera nuevas elecciones. Cuando el líder se dispone a celebrar que sus seguidores subieron del 30% al 34% se le hace notar que es una muestra pequeña y ese 34% de seguidores en ella puede no corresponderse con la FR de seguidores en la población general.

Recaban la asistencia de un estadístico y les dice, que en efecto, ese dato no permite concluir que ha aumentado la FR de seguidores, pues:

a) El Intervalo de Confianza al 95% es aproximadamente de **26% a 42%**.
b) El test estadístico correspondiente da **P = 0.15**.

El líder sigue haciendo cambios en el nombre y en la siguiente encuesta con N = 100 personas se encuentra que 46 se declaran dispuestos a votarles. De nuevo piden asistencia de un estadístico y este les dice que el resultado apoya fuertemente la creencia en que ha aumentado realmente la FR de seguidores, pues:

a) El Intervalo de Confianza al 95% es aproximadamente de **38%** a **54%**.
b) El test estadístico correspondiente da **P = 0.00003**.

Está muy comprobado y documentado que más del 98% de los profesionales con estudios superiores no entiende la información que el estadístico ha dado y aceptan las conclusiones que este establece como dogmas de fe de obligado acatamiento.

En este capítulo se explican estos conceptos y el cálculo de esos valores y se muestra que no son entidades matemáticas inaccesibles, sino ideas básicas al alcance de todo profesional.

Si en una muestra de tamaño **N** se encuentra que hay **X** individuos con cierta característica (como votar a un partido) la FR en la muestra es **p = X/N**. Conocer el valor de p en una muestra no permite conocer el valor de π en la población, pero a partir de p podremos calcular:

© Todos los derechos reservados

a) El llamado **Intervalo de Confianza**, IC, para π, de modo que tendremos gran confianza en que π esté dentro de él. El investigador puede elegir el nivel de confianza, pero cuanto mayor sea la confianza más ancho es el intervalo.

b) El valor **P del Test de Significación**, TS, que puede ayudar a concluir que π no tiene cierto valor en que estamos interesados. En principio queremos saber cuánto vale π, pero a partir de una muestra eso no es posible. Pero sí es posible saber que π no vale cierta cantidad, lo cual puede ser muy útil.

Los test estadísticos ayudan a saber cuánto **NO** vale el valor poblacional en que estamos interesados.

6.2. INTERVALO DE CONFIANZA PARA PROPORCIÓN POBLACIONAL

Un modo de hacer Inferencia Estadística es calcular un Intervalo de Confianza, IC, para el valor del parámetro que estamos estudiando. No hay razones estadísticas ni biológicas para elegir uno u otro nivel de confianza para el intervalo. Lo más habitual es calcularlos con una confianza del 95% o del 99%, pero el investigador decide la confianza que quiere tener y cuanta más confianza quiere, más ancho sale el intervalo.

La fórmula utilizada es:

$$IC(\pi) \equiv p \pm Z \cdot \sqrt{p(1-p)/N}$$

Donde Z es el valor que determina la confianza que vamos a tener en que el valor de la proporción poblacional esté dentro del intervalo calculado con esta fórmula.

Para IC al 95%, es Z = 1.96, y para confianza de 99% es Z = 2.58. En el capítulo siguiente se explica cómo elegir los valores de Z para otros niveles de confianza.

EJEMPLO 6.1

en una muestra de N = 100 personas hay 60 alérgicas a la aspirina, AA. La FR de AA en la muestra es p = **0.60**.

Calculamos el IC para π:

- *IC95%* = 0.60 ± 1.96 · (0.6 · 0.4/100)½ = 0.60 ± 0.096 = 0.504 y 0.696.
- *IC99%* = 0.60 ± 2.58 · (0.6 · 0.4/100)½ = 0.60 ± 0.126 = 0.474 y 0.726.

© TODOS LOS DERECHOS RESERVADOS

Tenemos confianza 95% en que la proporción poblacional de AA esté entre 0.504 y 0.696 y tenemos 99% de confianza en que esté entre 0.474 y 0.726. Confianza 95% quiere decir que de cada 100 intervalos calculados con esa fórmula, 95 contienen realmente el valor de π, y en los otros 5 el valor del parámetro está fuera del intervalo.

En la vida cotidiana manejamos continuamente los intervalos de confianza y hacemos intervalos más anchos si queremos tener más confianza. Uno de los ejemplos, es decir, cuando no conocemos la edad de una persona, que "probablemente tiene entre 40 y 50 años", y si queremos un intervalo dentro del cual tengamos mayor confianza en que se encuentre su edad decimos, por ejemplo, que "está comprendida entre 35 y 55 años".

Esta fórmula, basada en la aproximación de la distribución Binomial a la Normal, da valores próximos a los exactos cuando el número de casos "positivos" (con la característica C), y el número de casos "negativos" (sin C) son mayores de 5. Cuando no se cumple esta condición hay que recurrir a cálculos mucho más pesados, pero la interpretación es la misma.

6.3. TESTS DE SIGNIFICACIÓN CON UNA PROPORCIÓN

Para mostrar el mecanismo lógico que se usa en los Tests de Significación, (tanto en IE como en la vida común) supongamos que en Alemania se estudió a cada individuo de la población para averiguar si era o no alérgico a la aspirina, AA, y se encontró que lo son el **20%** de la población: π_A **= 20%**.

Se sospecha que la FR poblacional de AA es mayor de 20% en Bélgica, Córcega y Dinamarca. No pudiendo estudiar a todos los individuos de estas poblaciones, en cada país contaremos el número de AA en una muestra aleatoria de N = **1000**, y calcularemos la FR de AA en cada muestra, pB, pC y pD, respectivamente (pondremos la inicial del país como subíndice en la FR poblacional, π, y en la muestral, p). Para cada país consideremos la hipótesis:

"El porcentaje de AA es en la población 20%", es decir, la diferencia respecto a Alemania es nula. Se la llama **Hipótesis Nula** y se simboliza como H0. Se escribe **H0: π = 20%**.

¿Qué postura razonable se debe tomar, a partir de lo observado en cada muestra, acerca de la FR de AA en cada país πB, πC, πD?

1. Si entre los 1000 belgas se encuentran 970 AA $\rightarrow$ pB = **97%**, concluimos que:
   a) La πB no es 20% (rechazamos H0).
   b) La πB puede ser 20% (aceptamos H0 como posible).
   c) La πB es 20% (afirmamos que H0 es cierta).
2. Si entre los 1000 corsos se encuentran 210 AA $\rightarrow$ pC = **21%**, concluimos que:
   a) La πC no es 20% (rechazamos H0).
   b) La πC puede ser 20% (aceptamos H0 como posible).

© Todos los derechos reservados

c) La πC es 20% (afirmamos que H0 es cierta).

3. Si entre los 1000 daneses se encuentran 200 AA $\rightarrow$ pD = **20%**, concluimos que:

   a) La πD no es 20% (rechazamos H0).

   b) La πD puede ser 20% (aceptamos H0 como posible).

   c) La πD es 20% (afirmamos que H0 es cierta).

4. ¿Qué p debe salir en la muestra para que sea razonable elegir la opción "c"?

Resumamos el razonamiento que hemos hecho en los supuestos anteriores para elegir, correctamente, 1–a, 2–b, 3–b y "ningún valor de p lleva a la opción c":

1. Para cada país conocemos la FR de AA en la muestra, p, (92% en Bélgica, 21% en Córcega y 20% en Dinamarca) y queremos saber si la FR poblacional π puede ser 20% en cada uno de ellos, como lo es en Alemania.
2. Para cada país se plantea la llamada **Hipótesis Nula**, **H0**, que dice que π es 20%, como en Alemania (es decir, la diferencia respecto a Alemania es nula).
3. Si H0 es cierta **esperamos** que en la muestra sea p = 20% o un valor próximo a ese. Decimos que el Valor Esperado de la p es 20% o, lo que es lo mismo, que el Valor Esperado de X, número de AA en la muestra, es 200 (puesto que 200/1000 = 0.20 o 20%).
4. Si X se aleja mucho del 200 (es decir, p se aleja de mucho de 0.20), la muestra constituye fuerte evidencia contra la H0 y nos inclinamos por rechazarla.

   Si X es próximo a 200, la muestra no constituye evidencia contra la H0, no la rechazamos y la aceptamos como posible.

   - Hipótesis Nula, H0: π = **20%**
   - Valor esperado de p si es cierta H0 = **20%**

   Si encontramos **p** mucho mayor que 20%, rechazamos H0, será π > 20%.

   Si encontramos **p** próxima a 20%, no rechazamos H0, puede ser π = 20%.

En todos los casos de IE se sigue un razonamiento idéntico a este. Con valores p muy alejados o muy próximos al π de la H0 no hay que hacer ningún cálculo complementario. Pero si la **p** toma valores moderadamente alejados de la π de la H0, no es claro si se debe rechazar H0 o debe aceptarse como posible. En estos casos puede ayudar conocer el llamado **valor P** del test.

6.4. VALOR P DEL TEST CON BINOMIAL CON RESULTADO EXTREMO

Sigamos con el ejemplo en que la π de AA en Alemania es π = **20%**.

1. Sospechamos que también en España la π es mayor que en Alemania, es decir, πE > 20%. Se estudia una muestra de **N = 10** españoles y se encuentra que todos ellos son AA, es decir, **X = 10**, **p = 100%**. Este resultado sugiere que es πE > 20%. Pero al ser una muestra

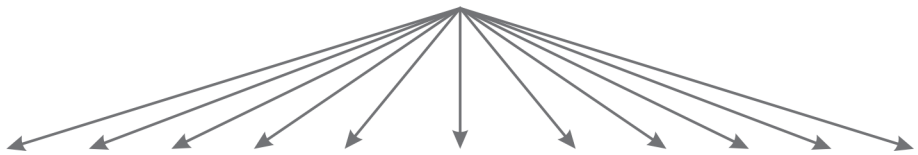
© Todos los derechos reservados

pequeña, el investigador piensa que podría ser $\pi E = 20\%$ y que $pE = 100\%$ puede haber aparecido en la muestra p. Puede ayudar a salir de la duda calcular el valor P del test, que en este caso es la prob. de que salga una muestra con $pE = 100\%$ de AA si en la población es $\pi E = 20\%$.

La distribución Binomial nos da estas probabilidades (capítulo 6).

- **Hipótesis Nula, H0:** en la población española son AA el 20%, $\pi E = 20\%$.
- **Valor Esperado** de pE bajo la H0: 20% (valor esperado de X, número de AA, es 2).
- **Valor Observado:** pE = 100% (10 AA).
- **Valor P del Test** = probabilidad de sacar una muestra con 10 AA si es $\pi E = 20\%$.
- π **= 20%** MA de N = 10.

$\pi = 20\%$ MA de N= 10



X →	0	1	2	3	4	5	6	7	8	9	10
p →	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
FR(X)	**.107**	**.268**	**.302**	**.201**	**.088**	**.026**	**.006**	**.001**	**.0001**	**.000004**	**.0000001**

$P(p_E = 100\% \; 1\pi_E = 20\%)$ ó $P(X = 101\pi_E = 20\%) =$ **0.000.0001**.

Este valor de P quiere decir que si en España son AA el 20% de la población y se hace MA de tamaño N = 10, solamente en una muestra de cada 10 millones son AA los 10 individuos. Es decir, es muy infrecuente que en el MA de N = 10 aparezca una muestra con todos AA si en la población son AA el 20%. Este resultado (los 10 españoles AA) es muy fuerte evidencia contra la H0 y a favor de que es $\pi_E > 20\%$.

2. Sospechamos que en Francia la π es menor que en Alemania, es decir, $\pi_F < 20\%$. Se estudia una muestra de 10 franceses y se encuentra que ninguno es AA, **X = 0**, **p = 0%**. Este dato sugiere que es $\pi_F < 20\%$. Pero también podría ser $\pi_F = 20\%$ y el p = 0% haber aparecido en la muestra por azar.

   En el esquema de la distribución Binomial vemos que:

   Valor P del Test = $P(p_F = 0\% \; 1 \; \pi_F = 20\%) = P(X = 0 \; 1 \; \pi_F = 20\%) =$ **0.107**.

   Si en Francia son AA el 20% de la población y se hace MA de N = 10, más de 10 muestras de cada 100 tienen ningún AA. Es frecuente que aparezca una muestra de N = 10 con ningún AA si en la población francesa son AA el 20%.

   El resultado (pF = 0%) no es evidencia contra la H0. No queda demostrado que sea $\pi_F < 0.20$. El dato es compatible con que sea $\pi_F = 0.20$.

© Todos los derechos reservados

En estos ejemplos, el valor P del test es la probabilidad de obtener una muestra como la obtenida en nuestro estudio si en la población son AA el 20%.

6.5. EL VALOR P DEL TEST CON RESULTADO NO EXTREMO. P DE LA COLA

Sigamos con el ejemplo en que la π de AA en Alemania es π **= 20%**.

1. Sospechamos que en Georgia la π es mayor que en Alemania, es decir, π_G > 20%. Se estudia una muestra de N = 10 georgianos y se encuentra que hay 5 AA, es decir, X = 5, p_G = 50%. ¿Es este resultado compatible con que en la población de Georgia haya 20% de AA, es decir, π_G = 20%?

   En la distribución Binomial vemos que:

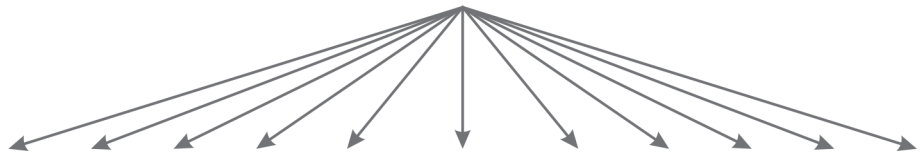
   P(pG = 50% 1 π_G = 20%) = **P**(X = 5 1 π_G = 20%) = **0.026**.

   Si en Georgia son AA el 20% de la población y se hace MA de tamaño N = 10, en 26 muestras de cada 1000 muestras será p_G = 50%. Pero cuando el resultado no es el dato más extremo, como lo era en los dos ejemplos anteriores (0 ó 10 AA en la muestra de N = 10), se toma como valor P del test no la FR de muestras con ese porcentaje de AA, sino la FR de ellas con ese % *o cualquier otro más alejado del esperado* bajo la H0.

   En este ejemplo en que el por ciento esperado en la muestra es 20%, porcentajes más alejados del esperado que el obtenido en la muestra son p_G = 50%, es decir, 5 o más AA en la muestra de N = 10.

   En la distribución Binomial obtenida en el MA con π = 20%, añadimos una última fila que da la FR de muestras con **X o más** AA, si es X mayor que 2 y da la FR de muestras con **X o menos** AA si es X menor que 2. Estos valores, llamados "Probabilidad de la cola" son los que se usan en los Tests Estadísticos.

π = 20% MA de N = 10



X →	0	1	2	3	4	5	6	7	8	9	10
p →	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
FR(X)	**.107**	**.268**	**.302**	**.201**	**.088**	**.026**	**.006**	**.001**	**.0001**	**.000004**	**.0000001**
P_{COLA}	**.107**	**.375**		**.322**	**.121**	**.033**	**.0071**	**.0011**	**.0001**	**.000004**	**.0000001**

© Todos los derechos reservados

Valor P del Test = $P(p_G \geq 50\%\ 1\pi_G = 20\%) = P(X \geq 5) = P(X = 5) + P(X = 6) + P(X = 7) = 0.026 + 0.006 + 0.001 =$ **0.033**. Se desprecian P(X = 8), P(X = 9) y P(X = 10) por su pequeño valor.

El motivo por el que se toma como valor P del test la probabilidad de encontrar una muestra como la de nuestro experimento o con estadístico aun más alejado del valor esperado bajo H0 se ve en el próximo apartado.

2. Sospechamos que en Hungría la π es menor que en Alemania, es decir, $\pi_H < 20\%$. Se estudia una muestra de N = 10 húngaros y se encuentra que uno es AA, es decir, X = 1, $p_H = 10\%$. La P del test es P ($p_H = 10\%$), pues 10% es menor que 20%.

   Valor P del Test= $P(p_H \leq 10\% \mid \pi_H = 20\%) = P(X \leq 1 \mid \pi_H = 20\%) = P(X=0) + P(X=1) = 0.107 + 0.268 =$ **0.375**.

   Con este valor P no se rechaza H0, el 10% de AA obtenido en la muestra de húngaros es compatible con que en la población de Hungría sean AA el 20%.

6.6. ¿POR QUÉ SE USA LA PROBABILIDAD DE LA COLA?

También se sospecha que en Italia, en Japón, en Kenia y en Lituania la π de individuos con AA es mayor que 20%. Para cada país se toma una muestra de **N = 1000**.

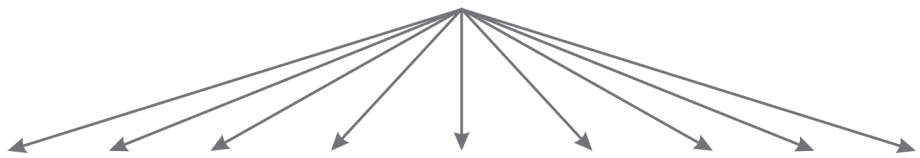
1. En **Italia** se encontró que entre las 1000 personas eran AA 203, $p_I = 20.3\%$. Todos los investigadores consideran este dato compatible con H0, es decir, que no es difícil que salga, $p_I = 20.3\%$ si es $\pi_I = 20\%$, puesto que la diferencia entre lo esperado bajo la H0 y lo encontrado es solamente de 3 individuos.

   Pero en la tabla siguiente vemos que $P(p_I = 20.3\% \mid \pi_I = 20\%) = P(X = 203 \mid \pi_I = 20\%) = 0.03$, un valor bastante bajo que podría considerarse moderada evidencia contra la H0. Pero la probabilidad de la cola es:

   Valor P del Test= $P(p_I \geq 20.3\% \mid \pi_I = 20\%) = P(X \geq 203 \mid \pi_I = 20\%) =$ **0.4187**

   Si es $\pi_I = 20\%$, de cada cien muestras de N = 1000, más de 42 muestras tendrán 203 o más AA. Este valor no invita a rechazar H0.

π = 20% MA de N = 1000

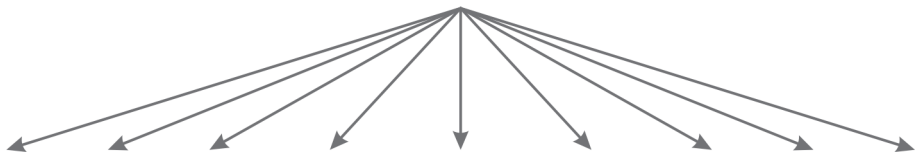


X →	50	180	190	200	203	220	250	300	400
p →	5%	18%	19%	20%	20.3%	22%	25%	30%	40%
P(x)	**10^{-42}**	**.0091**	**.023**	**.032**	**.030**	**.0090**	**.00002**	**10^{-14}**	**10^{-48}**
P_{COLA}	**10^{-41}**	**.0602**	**.2274**		**.4187**	**.0628**	**.00007**	**10^{-13}**	**10^{-47}**

© Todos los derechos reservados

2. En **Japón** se tomó una muestra de N = 10.000 personas y eran AA 2.001, p_J = 0.2001 o 20.01%. La siguiente tabla da las probabilidades de algunos valores de X para π = 20% y MA de N = 10.000. Vemos que P (X = 2001) = **0.00996**. Pero ello no quiere decir que una muestra con 2001 es difícilmente compatible con que en la población sea π = 0.20. El sentido común sugiere que es fácil obtener muestras con X próxima a 2001 si es π = 0.20. La P_{COLA} es P (2001)+P (2002) + P (2003)+...+ P (9999) + P(10000) = **0.4940**. Nos dice que en el MA aparecen muchas muestras con X = 2001 o mayor, es decir, que ese valor muestral es totalmente compatible con π = 20%.

π = 20% MA de N= 10000



X →	0	1900	1950	**2000**	2001	2050	2100	2150	2200
p →	5%	19%	19.5%	20%	20.1%	20.5%	21%	21.5%	25%
P(x)	**$2 \cdot 10^{-8}$**	**.0006**	**.0046**	**.00997**	**.00996**	**.0046**	**.0004**	**.00001**	**10^{-35}**
P_{COLA}	**10^{-194}**	**.0007**	**.1077**		**.4940**	**.1082**	**.0067**	**.00009**	**10^{-36}**

3. En **Lituania** se encontró que entre las 10000 personas eran AA 2000, p_L = 20%, justamente lo que teóricamente se espera tener si es cierta H0. En estos casos *no tiene sentido calcular el valor del test*, puesto que este indica la probabilidad de obtener un resultado que *se aleje* de lo propuesto por H0 tanto como se aleja el de nuestra muestra o más. Pero el valor de la muestra *NO se aleja* del propuesto por H0, sino que coincide exactamente con él. Ese resultado es el más compatible con H0, el que menos invita a rechazarla. Pero, por supuesto, no permite afirmar que es cierta.

   En la tabla vemos que P (p_L = 20% |π_L = 20%) = P (X = 2000 |π_L = 20%) = **0.00997**. Pero ese valor no constituye evidencia en contra de que sea π_L = 20%.

   Con Binomiales de N muy grande o con variables continuas la probabilidad de que en las muestras el estadístico tenga un valor determinado se aproxima a cero, tanto si este valor es muy lejano al esperado bajo la H0 como si es muy cercano a él, o incluso coincide con él. Por eso la probabilidad de que en la muestra haya cierto valor concreto **no** informa sobre la compatibilidad de ese resultado con la H0. Por ello se usa la P de la cola para cuantificar la compatibilidad o incompatibilidad de los datos con la H0. Esto se ve aun más claramente en el siguiente ejemplo.

© Todos los derechos reservados

6.7. CÁLCULO DEL VALOR P DEL TEST EN BINOMIAL CON N GRANDE. APROXIMACIÓN NORMAL

Calcular el valor de "PCOLA" en Binomiales con N grande exige aplicar la fórmula binomial docenas, cientos o miles de veces. Así, en la Bin (10000, 0.2) que recoge la tabla anterior vemos que para conocer P ($X \geq 2050$)= P (2050)+P (2051)+P (2052) +....+ P (10000) = **0.1082**, hay que aplicar la fórmula binomial 7951 veces. Pero veíamos en al capítulo 6 que la distribución Normal, DN, nos permite encontrar una buena aproximación a ese valor, con mínimo esfuerzo.

Puesto que esta Binomial tiene media E = 2000 y desviación estándar $\sigma X = \{\pi \cdot (1-\pi) \cdot N\}^{1/2} = \{0.2 \cdot 0.8 \cdot 10000\}^{1/2} = 40$, 2050 se aleja de 2000 en 50 unidades, es decir, Z = (2050–2000)/40= 1.25 desviaciones. En las tablas de la DN vemos que P ($Z \geq 1.25$)= **0.1056**, valor muy próximo al obtenido con la Binomial.

Como segundo ejemplo, vemos que P ($X \leq 1900$)= **0.0007**. Para llegar a ese valor hubo que calcular 1901 binomiales y sumarlas. Pero 1900 se aleja de 2000 en 100, es decir, Z = (1900–2000)/40 = –**2.5** desviaciones. En las tablas de la DN vemos que P ($Z \geq 2.5$) = P ($Z \leq -2.5$) = **0.0062**, valor suficientemente próximo al obtenido con la Binomial.

En general, el cálculo es:

$$Z = \frac{|E - X|}{\sigma_X} = \frac{|E - X|}{\sqrt{\pi \cdot (1-\pi) \cdot N}}$$

Donde **E** es la cantidad esperada de casos bajo la H0, $E = \pi \cdot N$ y el valor P del test se obtiene mirando en las tablas de la DN la FR para Z mayor o igual que el obtenido.

Dividiendo numerador y denominador por N se puede escribir también como:

$$Z = \frac{|\pi - p|}{\sigma_p} = \frac{|\pi - p|}{\sqrt{\pi \cdot (1-\pi)/N}}$$

Nota 1. El valor P encontrado con la aproximación de la DN se aproxima más al verdadero de la distribución Binomial si se aplica la llamada "*corrección por continuidad*", **cpc**, que consiste en ampliar el intervalo de interés en 0.5 unidades, de modo que las fórmulas quedarían de esta forma, que el estudiante verá en los textos:

$$Z = \frac{|\pi - p| - 0.5/N}{\sqrt{\pi \cdot (1-\pi)/N}} \qquad Z = \frac{|E - X| - 0.5}{\sqrt{\pi \cdot (1-\pi) \cdot N}}$$

© Todos los derechos reservados

En general la aproximación es mejor aplicando cpc, pero en nuestros ejercicios no la aplicaremos para simplificar el cálculo.

Nota 2. Con valores de π muy pequeños, digamos menores de 0.01, $(1-\pi)$ se aproxima mucho a 1 y la expresión Z = 1 E – X 1/[$\pi \cdot (1-\pi) \cdot$ N]½ se aproxima mucho a Z = 1 E–X 1/($\pi \cdot$ N)½ =1 E–X 1/ E ½, que también puede escribirse como:

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

Donde X, núm. de individuos "positivos" en la muestra, aquí se indica por "O".

EJEMPLO 6.2

En la población general son hipertensos el 0.8% (π = 0.008) de los jóvenes y en una muestra de 250 jóvenes sedentarios son hipertensos 7. ¿Es razonable concluir que el sedentarismo produce incremento de la FR de hipertensos?

Hipótesis nula, H_0: π_{SED} = 0.008, E bajo H_0 = 0.008 · 250 = 2. Valor Observado: O = 7 $\rightarrow$ Z^2 = (2–7)2/2 = 12.5 $\rightarrow$ Z = 3.5 $\rightarrow$ P = **0.0002**. El resultado (7 hipertensos de 250 jóvenes) es difícilmente compatible con que π_{SED} = 0.008. Se rechaza H0.

Concluimos que entre los sedentarios hay más de 0.8% de hipertensos.

6.8. RESULTADOS ESTADÍSTICAMENTE "SIGNIFICATIVOS" Y "MUY SIGNIFICATIVOS". LAS BARRERAS DEL 5% Y DEL 1%

Es evidente que con, por ejemplo, P = 0.0000001 rechazamos la H0, porque el dato es muy difícilmente compatible con ella, y con P = 0.20 no rechazamos H0, porque el dato es compatible con ella. ¿Pero qué postura tomamos con, por ejemplo, P = 0.03, que no es ni tan pequeño como el primero ni tan grande como el segundo?

No hay ningún criterio matemático, biológico ni sociológico, para establecer un valor de P frontera tal que por debajo de él se rechaza la H0 y por encima de él se acepta como posible. Cuanto menor es el valor P, más evidencia constituye contra la hipótesis nula, pero no hay una cifra que separe la zona de aceptación de la H0 de la zona de rechazo. El paso de una a otra es gradual, con una zona de transición en la que el valor P no invita claramente a tomar ninguna de las dos posturas. Muchos investigadores creen que el valor P = 0.05 es la barrera que separa la zona de rechazo de H0 de la zona de aceptación. Pero ello es una simplificación excesiva. Los valores P =

© TODOS LOS DERECHOS RESERVADOS

0.06 o P = 0.04, por ejemplo, nos dicen prácticamente lo mismo acerca de nuestra sospecha de que H0 es falsa y ninguno de los tres garantiza que lo sea.

Solamente en una tesitura de *Toma de Decisiones* fácticas (infrecuente en la investigación científica) debemos convenir un valor P por debajo del cual se *ejecuta una acción* y por encima de él se ejecuta otra. Pero cuando el objetivo no es tomar decisiones fácticas sino adquirir conocimiento (lo habitual en investigación, análisis de encuestas y estudios de mercado), no hay valor P frontera.

Es habitual decir que el resultado es "**estadísticamente significativo**" cuando es **P ≤ 0.05.** Esto puede ser un convenio útil si se sabe lo que indica. Pero si no se entiende ese convenio puede generar más confusión que ayuda. Un error muy frecuente es creer que "estadísticamente significativo" indica que queda probado que la hipótesis nula es falsa. Pero acabamos de ver que "estadísticamente significativo" es otro modo de decir "P= 0.05", lo cual apunta hacia que H_0 no es cierta, pero no lo garantiza. Al rechazar la H0 cuando se encuentra "P ≤ 0.05" corremos un riesgo de 5% de rechazar hipótesis nulas que son ciertas. Es decir, si rechazamos las H_0 cuando encontramos P ≤ 0.05, a la larga rechazaremos, equivocadamente, 5 de cada cien H_0 ciertas.

La evidencia contra H_0 aumenta al disminuir el valor P. Cuando se obtiene **P ≤ 0.01** se suele hablar de resultado "**estadísticamente muy significativo**". También puede ser un convenio útil siempre que esté claro que "muy significativo" es otro modo de decir "P ≤ 0.01". Ello implica notable sospecha en que la H_0 es falsa, pero no seguridad definitiva. Al rechazar la H_0 cuando se encuentra P ≤ 0.01 corremos riesgo de 1% de rechazar hipótesis nulas que son ciertas. Por ejemplo, si un dado homogéneo en el que realmente cada cara sale a la larga un sexto de las veces, se tira tres veces, la prob. de que salgan tres "6" es 0.004, menor de 0.01 y todos los jugadores de parchís saben que ese suceso no es tan infrecuente.

© Todos los derechos reservados

Resumiendo

1. La disyuntiva no está entre rechazar H_0 si P es pequeña y *afirmar que es cierta* si P es grande. Ningún valor de P permite concluir que H_0 es cierta. Rechazamos H_0 si P es muy pequeño, por ejemplo, P = 0.000 001 y no la rechazamos si P es grande, por ejemplo, P = 0.30.
2. **No** hay un valor P frontera, que lleve a rechazar H_0 con valores P menores que esa frontera y a no rechazarla con valores P mayores que ella.
3. "**Estadísticamente significativo**" es sinónimo de P ≤ 0.05, que implica cierta evidencia contra la H_0, pero sin darnos seguridad de que es falsa.
4. "**Estadísticamente muy significativo**" es sinónimo de P ≤ 0.01, que supone notable evidencia contra la H_0, pero sin darnos seguridad total de que es falsa.
5. Si se rechazan las H_0's cuando es P ≤ 0.05, a la larga se rechazarán erróneamente 5 de cada 100 H_0's ciertas. Y si se rechazan las H_0's cuando es P ≤ 0.01, a la larga se rechazará erróneamente 1 de cada 100 H_0's ciertas. Ejemplo: un laboratorio busca medicamentos útiles para cierta enfermedad y prueba con muchos. Para cada uno de ellos plantea como H_0 que no es útil. Si la rechaza cuando aparece P < 0.05, a la larga declarará

como útiles 5 de cada 100 medicamentos inútiles. Y si rechaza H_0 cuando aparece P < 0.01, a la larga declarará como útiles a 1 de cada 100 medicamentos inútiles.

6.9. EL VALOR P DEL TEST NO ES LA PROBABILIDAD DE QUE SEA CIERTA LA H_0

Ya sabemos que el Valor P del test nos dice la probabilidad de obtener cierto tipo de muestras cuando es cierta H_0. El error más frecuentemente cometido por los investigadores es creer que esa P indica la probabilidad de que la H_0 sea cierta. Pero esa creencia es totalmente equivocada.

6.10. LA PROBABILIDAD DE LA H_0 Y EL VALOR P DEL TEST

En la mayoría de los estudios biomédicos no se puede calcular la probabilidad de que la hipótesis planteada sea cierta. Veremos un ejemplo muy sencillo de otra área en el que **sí** se puede calcular esa probabilidad. Ello ayuda a distinguir esa probabilidad de la P del test y a entender cómo esta última se utiliza para elaborar conclusiones.

© Todos los derechos reservados

Ejemplo 6.3

Sabemos que el 3% de las monedas de una ciudad son falsas y aunque su aspecto es indistinguible de las legales, en las falsas la probabilidad de salir "cara" es 0.8. En las monedas legales la probabilidad de salir "cara" es 0.5.

1º Hemos comprado una moneda en la ciudad. No sabemos si es legal o falsa. Consideremos $H_0 \equiv$ *La moneda es legal*. La probabilidad de que esta H_0 sea cierta es 0.97 Es decir, si tomamos muchas monedas de esa ciudad, serán legales 97 de cada 100.

2º Para intentar saber si nuestra moneda es falsa o legal hacemos 14 tiradas. Si el número de veces que sale "cara" es próximo a 7, consideramos el resultado compatible con la H_0 y si salen muchas caras aumentará nuestro temor de que sea falsa. El resultado es que sale cara en las 14 tiradas.

Un cálculo sencillo nos dice que haciendo millones de series de 14 tiradas cada una con una moneda legal, solamente en 6 cada 100 000 series aparecen todo caras. Valor P del test = 0.00006 Distingamos claramente entre:

- **P = 0.97** Es la probabilidad de que la moneda sea legal (que H_0 sea cierta).
- **P = 0.00006** Es la probabilidad de obtener 14 caras **si** la moneda es legal. Es la probabilidad de obtener el resultado obtenido en nuestro estudio si es cierta H_0.

Son dos valores muy diferentes, 0.97 y 0.00006, referidos a hechos muy diferentes:

La probabilidad de que la H_0 sea cierta, es decir, que la moneda sea legal es P= 0.97. De cada 100 monedas compradas en esa ciudad, 97 serán legales. Pero esa cifra no nos dice si nuestra moneda es una de las legales o una de las falsas. Intentamos saber más acerca de ella haciendo esos 14 lanzamientos y el resultado es que sale cara en todos ellos. La P del test es la probabilidad de que salga todo caras en 14 lanzamientos con una moneda legal.

Si no conociéramos la FR de monedas legales, es decir la probabilidad de que la moneda comprada fuera legal, el único dato sería el valor P del test. En la llamada *Inferencia Clásica o frecuentista*, que se explica en esta asignatura y se aplica en la gran mayoría de los análisis, se asume que no se puede evaluar la probabilidad de que H_0 sea cierta o sea falsa, lo que en este ejemplo equivale a no conocer la FR de monedas falsas en la ciudad. El investigador toma una muestra (equivale a tirar la moneda 14 veces) y calcula la probabilidad de que aparezca por azar el tipo de resultado encontrado en la muestra, si es cierta la H_0.

En la llamada *Inferencia Bayesiana* **sí** se evalúa la probabilidad de que la H_0 sea cierta antes de conocer el resultado de nuestro estudio y se ve cómo se modifica al tener en cuenta dicho resultado. En nuestro ejemplo la Inferencia Bayesiana implica averiguar la probabilidad inicial de que la moneda sea legal y ver cómo se modifica esa probabilidad al tener en cuenta que ha salido cara en las 14 tiradas.

Aunque algunos autores presentan estos dos enfoques, el clásico y el bayesiano, como alternativas contrapuestas, una reflexión serena muestra que son técnicas complementarias que pueden colaborar en un mismo análisis.

© Todos los derechos reservados

6.11. RIESGO DE RECHAZAR H_0 EQUIVOCADAMENTE

Consideremos de nuevo el caso de una enfermedad que sin tratamiento cura en el 20% de los casos, π= 0.20. Sospechando que el nuevo tratamiento "B" cura más de 20%, se hace un estudio tratando N= 10 enfermos con B y se plantea la H_0: B es inútil, πB = 0.20. Si en la muestra de 10 tratados aparecen 10 curaciones, el test da P = 0.0000001, lo que lleva a concluir que B cura más de 20%. Si en la muestra hay 3 curaciones, es P = 0.3222, el resultado es compatible con que B cure el 20%, no queda demostrado que sea útil.

π = 20%										MA de N= 10000	
X →	0	1	2	3	4	5	6	7	8	9	10
p →	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
FR(X)	**.107**	**.268**	**.302**	**.2013**	**.0881**	**.0264**	**.0055**	**.0008**	**.0001**	**.000004**	**.0000001**
P_{cola}	**.107**	**.375**		**.3222**	**.1209**	**.0328**	**.0064**	**.0009**	**.0001**	**.000004**	**.0000001**

Si hay 5 curaciones, es P = 0.0328, lo que sugiere que B es útil pero no permite asegurarlo. Podemos seguir el convenio establecido y decir que el resultado es "estadísticamente

significativo" porque es $P \leq 0.05$. Si lo hacemos así calificaremos como "estadísticamente significativo" los resultados en que hay 5 o *más curaciones* (ya que $P(X \geq 5) = 0.0328$ que es menor que 5% y $P(X \geq 4) = 0.1209$ que ya es mayor de 5%).

En la tabla vemos que con un medicamento inútil (con él curan 20%) 5 o más curaciones aparecen por azar en el 3.28% de las veces que se repite el experimento, es decir, 3 o 4 veces cada 100 repeticiones. Si hacemos este estudio (probar en una muestra de 10) con muchos medicamentos, 3 o 4 de cada 100 inútiles darán 5 o más curados por azar, será $P < 0.05$, el resultado será calificado como "significativo" y el medicamento calificado como "útil", *erróneamente*.

Veamos un nuevo ejemplo para ayudarnos a entender que el valor P nos da la probabilidad de equivocarnos al rechazar la H_0. Un jugador de un casino sospecha que el dado usado para cierta apuesta está trucado en el sentido de que a la larga la cara "5" sale con frecuencia superior a 1/6. Para intentar aclarar ese punto hace una "investigación" que consiste en tirar el dado 3 veces. El resultado es que sale "5" en las tres tiradas.

- **Hipótesis Nula, H_0:** a la larga, la cara "5" sale 1/6 de las veces: $\pi 5 = 1/6$.
- **Valor Esperado bajo H_0:** en tres tiradas esperamos $3 \cdot 1/6 =$ **0.5**, es decir 0 o 1 veces el "5".
- **Valor Observado:** el "5" sale **3** veces.
- **Valor P del Test:** probabilidad de "5" las tres veces con un dado no trucado. Es $P = (1/6)3 =$ **0.004**.

Puesto que es $P < 0.01$ se califica este resultado como "muy significativo" y se considera notable evidencia contra la H_0 y, por tanto, a favor de que el dado está trucado. Es una conclusión razonable, siempre y cuando se tenga presente que un dado no trucado da ocasionalmente este resultado.

Concretamente, el valor P del test nos indica que si hacemos esa prueba con muchos dados, en 4 de cada 1000 no trucados saldrían los tres "5" y nos inclinaríamos –erróneamente– a considerarlos trucados.

6.12. ASPECTOS A RECORDAR

- La Inferencia Estadística, IE, intenta conocer acerca de los **parámetros** de la población investigada a partir del conocimiento de los **estadísticos** de la muestra analizada.
- Durante decenios se han infrautilizado los Intervalos de Confianza, que en muchas aplicaciones son la herramienta más útil de la IE. Son más fáciles de interpretar que los tests estadísticos y además informan sobre la **magnitud del efecto**, cosa que los tests no hacen.
- En un estudio destinado a ver si un nuevo fármaco "A" es útil para una enfermedad que sin tratamiento se cura en el 7% de los casos, la **Hipótesis Nula**, H0, dice que A cura el 7%.

© Todos los derechos reservados

- En los **Tests de Significación** se confronta la Hipótesis Nula con los datos de la muestra que se ha analizado. La H_0 se rechaza si los datos son incompatibles o difícilmente compatibles con ella. Y si son compatibles con la H_0 esta no se rechaza (lo que no implica afirmar que es cierta).
- En los Tests de Significación se confronta el **valor** del estadístico **esperado** bajo la H_0 con su **valor observado** en la muestra. La H_0 se rechaza si el valor observado se aleja mucho del esperado y no se rechaza si el valor observado es cercano al esperado. La evidencia contra la H_0 es mayor cuanto más se aleja el valor observado del valor esperado.
- El **valor P del test** dice cuan probable es obtener una muestra con valor observado tan alejado del esperado como el de nuestro estudio o más alejado.
- Cuanto más alejado está el valor observado del esperado menor es el valor P y mayor es la evidencia contra la H_0. Pero no hay un valor frontera. Los valores P= 0.06 o P= 0.04 nos dicen prácticamente lo mismo que P= 0.05 y ninguno de los tres garantiza que H_0 sea falsa.
- Si en un estudio destinado a ver si el fármaco "A" es útil para una enfermedad que sin tratamiento se cura en el 20% de los casos, en una muestra de N= 100 tratados con A se curan 32, es P = **0.0013**. Ese número dice que si A es inútil y se hacen millones de estudios como este, en la mayoría de ellos el número de curados será próximo a 20 y solamente en 13 de cada 10000 estudios aparecerán por azar 32 o más curados.
- Si un laboratorio que busca medicamentos útiles para una enfermedad prueba con muchos, para cada uno de ellos plantea como H_0 que no es útil y la rechaza si aparece $P \leq 0.05$, a la larga declarará como útiles 5 de cada 100 medicamentos inútiles.

© Todos los derechos reservados

Resumen

© Todos los derechos reservados

© Todos los derechos reservados

Distribución normal y teorema central del límite

7

Visión general

- Definición del intervalo de confianza de una proporción.
- Definición del valor “p” de una proporción.
- Definición del valor “p” del test con binomial.

Objetivo

- Conocer la inferencia estadística con una proporción.

© Todos los derechos reservados

7.1. DISTRIBUCIÓN DE VARIABLES CONTINUAS

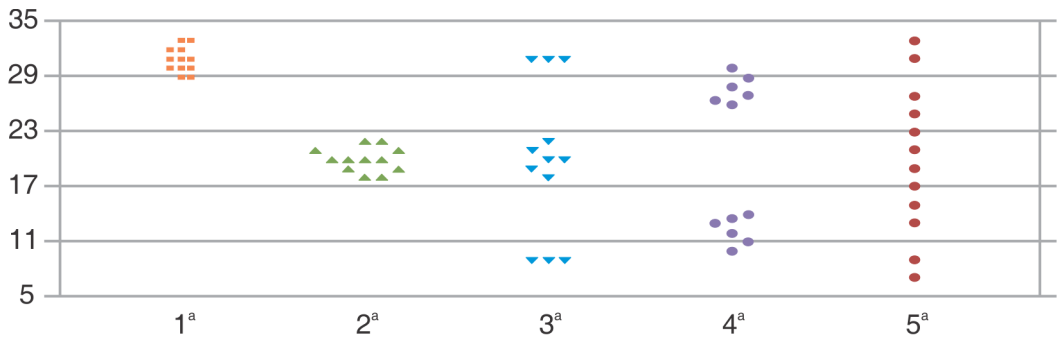
En el capítulo anterior veíamos que distintas distribuciones Binomiales con el mismo valor esperado, E, dan la misma distribución si el tamaño de la muestra, N, es grande y la FR de individuos con la característica dicotómica en la población, π, es pequeña; y que esas probabilidades se aproximan muy bien con la fórmula de Poisson. Esto es porque cuando N crece y π disminuye, de modo que $E = N \cdot \pi$ se mantiene constante, la fórmula Binomial tiende a la de Poisson. Calcular probabilidades con la fórmula de Poisson es más sencillo y además permite hacerlo cuando no se conoce N ni π, pero se conoce, o se puede estimar, $E = N \cdot \pi$.

También vimos que en Binomiales con N grande la FR de muestras con número de casos positivos comprendido en un intervalo depende solamente de los valores estandarizados de los extremos del intervalo. Las FR correspondientes a cada valor de Z están indicadas en las llamadas tablas de la distribución Normal.

Pero la distribución Normal, DN, –definida porque más allá de cada valor Z hay la FR de casos que indica la tabla– aparece en otras muchas circunstancias y resulta especialmente útil en muy distintas aplicaciones de Inferencia Estadística.

La ley de la distribución Normal, constituye una de las más sorprendentes y útiles de la Naturaleza. Es de aplicación tan general como la Ley de la Gravitación Universal y en belleza podría competir con la ley de autoinducción de Faraday y de Lenz.

En la gráfica siguiente se representan los puntos correspondientes a 5 muestras donde a cada uno de los 12 individuos de cada muestra se le mide una variable cuantitativa con valores entre 5 y 35.



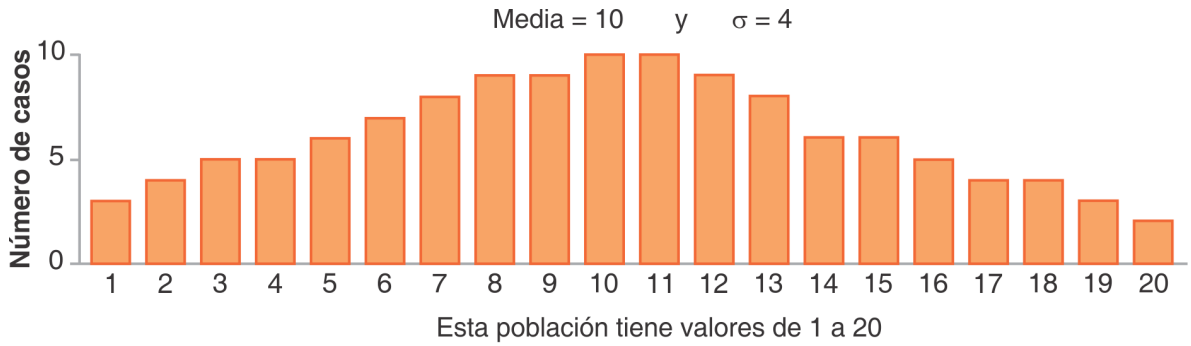
	1ª	2ª	3ª	4ª	5ª
Media	31	20	20	20	20
Desviación	1.4	1.3	8.2	8.2	8.2

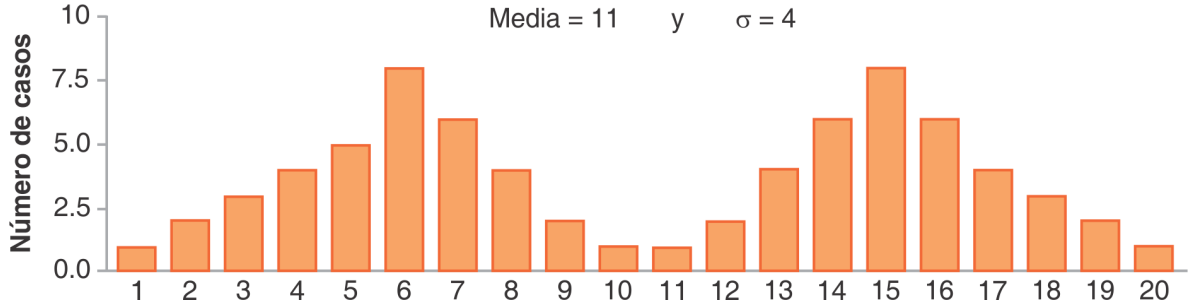
La primera muestra a la izquierda tiene, como puede verse en la escala vertical, todos sus valores entre 29 y 35, con media 31 y desviación estándar 1.4, como se indica en las dos últimas filas. La segunda muestra tiene todos los valores entre 17 y 23, la tercera tiene tres valores en el tramo superior, 6 en el intermedio y otros 3 en el inferior, etc.

Compare las medias y desviaciones de estas cinco muestras. Las tres últimas tienen igual media y desviación. ¿Quiere eso decir que los valores se distribuyen de idéntico modo? Ciertamente no. En la tercera muestra hay 6 valores próximos a la media, mientras que en la cuarta no hay ninguno en ese intervalo de 17 a 23. Y en la quinta los valores se reparten uniformemente en todo el rango entre 5 y 35.

En colectivos con muchos valores ocurre lo mismo. Dos colectivos pueden tener igual Media y Desviación y sin embargo los valores puede que se agrupen de diferente modo. En las gráficas siguientes la altura de cada barra indica el número de veces que se presenta cada uno de los posibles 20 valores.

© Todos los derechos reservados





Vemos que las dos distribuciones tienen igual media y dispersión, pero los valores se distribuyen de modo muy distinto. La primera tiene muchos individuos con valor en torno a 10 y el número de ellos va disminuyendo a medida que nos alejamos de 10 en ambos sentidos. La segunda tiene pocos individuos con valor en torno a 10 y muchos en torno a 6 y a 15.

© Todos los derechos reservados

7.2. LA DISTRIBUCIÓN NORMAL

Para exponerla consideremos estas tres variables que expresan magnitudes diferentes, que se han medido en poblaciones diferentes y de las que se han obtenido muy diferentes valores de medias y desviaciones:

1 → "G", Peso de la población de tigres de Malasia: μ = 500 y σ = 100.

2 → "X", Inteligencia de la población de españoles: μ = 160 y σ = 10.

3 → "D", Dureza de la población de miles de rocas lunares: μ = 20 y σ = 2.

En cada una de las poblaciones, tigres, españoles y rocas lunares, se calcula la proporción de individuos que tienen el valor de su variable en los intervalos que se especifican, y se obtienen estos resultados:

Tigres	μ= 500 σ = 100	P(G > 600) = **0.1587**	P(G > 700) = **0.0228**
Españoles	μ=160 σ = 10	P(X > 170) = **0.1587**	P(X > 180) = **0.0228**
Rocas	μ=20 σ = 2	P(D > 22) = **0.1587**	P(D > 24) = **0.0228**

Siendo individuos, variables y cantidades numéricas diferentes, ¿qué cosa es igual en las tres poblaciones? En las tres ocurre que el 15.87% de los individuos tiene valor superior a la media más una desviación y el 2.28% de los individuos tiene valor superior a la media más dos desviaciones, es decir:

$P(X > \mu + 1\sigma) = P(Z > 1) = 0.1587.$

$P(X > \mu + 2\sigma) = P(Z > 2) = 0.0228.$

Se cumple, además, la simétrica respecto a la media. Por ejemplo, en la población de tigres si el 15.87% pesa más de 600, otro 15.87% pesa menos de 400. Además, como el 50% pesa más que la media, los que tienen peso entre 500 y 600 son 0.50–0.1587 = 0.3413. Y también el 0.3413 de los individuos pesan entre 400 y 500, de modo que los que pesan entre 400 y 600 son 0.3413 · 2 = 0.6826.

La siguiente tabla nos recuerda la FR de individuos con Z mayor que ciertas cantidades. La tercera fila da la FR de individuos con Z entre cero y el valor de la cabecera de la columna, que es 0.50 menos el valor de la fila anterior. Las FR de la cuarta son el doble de las de la fila anterior (por la simetría antes citada).

Z →	0.10	0.20	0.25	0.50	0.75	1.00	1.50	2.00	2.50	3.00
P más allá de Z	.4602	.4207	.4013	.3085	.2266	.1587	.0668	.0228	.0062	.0013
P entre μ y Z	.0398	.0793	.0957	.1915	.2734	.3413	.4332	.4772	.4938	.4987
P entre -Z y Z	.0796	.1596	.1914	.3830	.5468	.6826	.8664	.9544	.9876	.9974

Por ejemplo, para Z= 0.5 la "P más allá de Z" es 0.3085. Es decir:

30.85% de los tigres pesan más de 550.

30.85% de los españoles tienen inteligencia superior a 165.

30.85% de las rocas tienen dureza superior a 21.

De cada 10000 rocas, 13 tienen dureza superior a 26, otras 13 tienen dureza menor de 14. Y 0.5000–0.0013 = 0.4987, es decir 4987 de cada 10.000, tienen dureza entre 20 y 26 (fila "P entre μ y Z" en la tabla). Y otras tantas están entre 14 y 20, de modo que con dureza entre 14 y 26 ($-3 < Z < 3$) hay 2 · 0.4987 = 0.9974.

La tabla de la distribución Normal, al final de la asignatura, da las FR para "más allá de Z", desde Z = 0.02 hasta Z = 4. Se dice que una variable se distribuye de modo Normal cuando se cumplen las FR recogidos en esta tabla. Y se suele indicar como N (500,100) para el peso de los tigres, N (160,10) para la inteligencia de los españoles y N (20,2) para la dureza de las rocas.

Por ejemplo, vemos que P (Z > 1.96) = 0.025 o 2.5%, lo que quiere decir:

a) 2.5% de los tigres pesan más de la media más 1.96 desviaciones, es decir, 500+100 • 1.96 = 500+196 = 696 (y otro 2.5% menos de 500–196 = 304).

© Todos los derechos reservados

b) 2.5% de españoles tienen inteligencia mayor de media más 1.96 desviaciones, esto es, 160+10 • 1.96 = 160+19.6 = 179.6 (y otro 2.5% menos de 160–19.6 = 140.4).

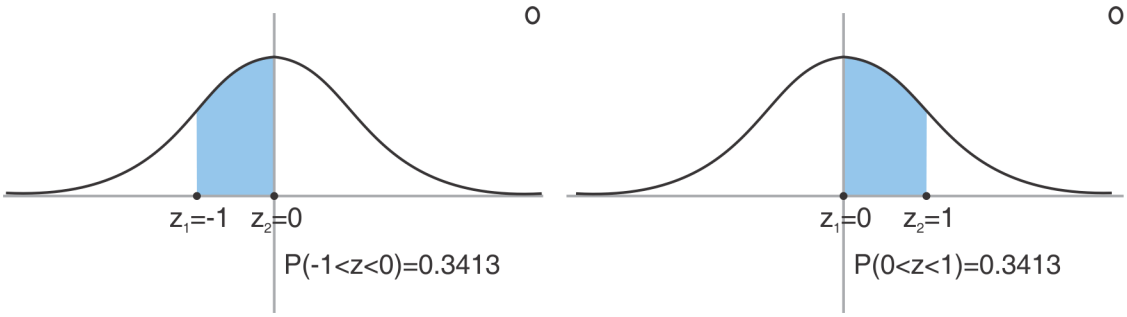
c) 2.5% de las rocas tienen dureza superior a la media más 1.96 desviaciones, es decir, 20+2 • 1.96 = 20+3.92= 23.92 (y otro 2.5% menos de 20–3.92 = 16.08).

Muchas variables de la naturaleza tienen distribución muy próxima a la Normal, lo que permite, si se conoce la media y la desviación, calcular la proporción de individuos con valor dentro de cualquier intervalo. En el próximo capítulo veremos la gran utilidad de ello en la investigación.

Esta distribución se representa mediante una curva en forma de campana que se conoce con el nombre de Campana de Gauss. En ella, la superficie encerrada entre la curva y cualquier intervalo del eje horizontal (de los valores de la variable) representa la FR de individuos con valor de la variable en dicho intervalo (es la superficie sombreada en las figuras).

Por ejemplo, puesto que es P (Z >1) = 0.1587 y por encima de la media están la mitad de los valores, es P (0 < Z < 1) = 0.5–0.1587 = 0.3413, es decir, en el 34.13% de los individuos el valor de la variable está entre la media y la media más una desviación, y en otros tantos el valor de la variable está entre la media y la media menos una desviación:

Prop ($\mu-\sigma < X < \mu$) = P ($\mu < X < \mu+\sigma$) = 0.50–0.1587 = 0.3413 o
P (–1 < Z < 0) = P (0 < Z < +1) = 34.13%.

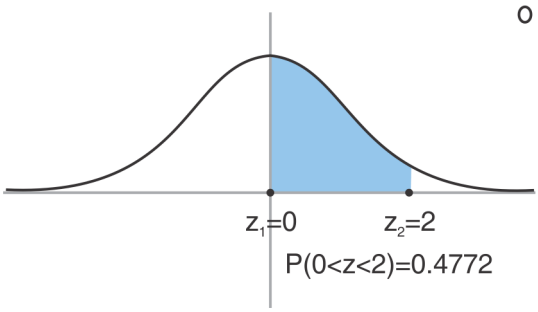


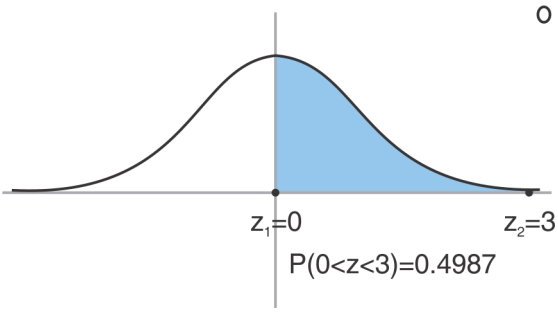
De la misma forma, con 2 y 3 desviaciones estándar por encima y por debajo de la media hay 47.72% y 49.87% de individuos respectivamente.

Prop($\mu-2\sigma < X < \mu$) = P ($\mu < X < \mu+2\sigma$) = 0.5–0.0228 = 0.4772 o
P (–2 < Z < 0) = P (0 < Z < +2) = 0.4772

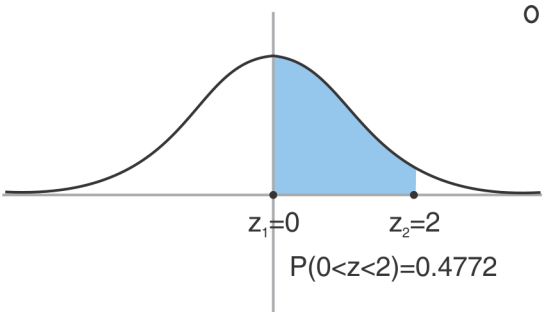
Prop($\mu - 3\sigma < X < \mu$) = P ($\mu < X < \mu + 3\sigma$) = 0.5 – 0.0013 = 0.4987 o
P (–3 < Z < 0) = P (0 < Z < +3) = 0.4987

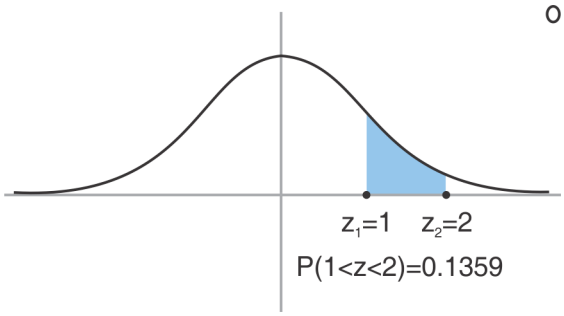
© Todos los derechos reservados





Otros ejemplos: P (0 < Z < 2) = 0.5– P (Z > 2) = 0.5–0.0228 y P (1 < Z < 2) = P (Z > 1)–P (Z > 2) = 0.1587–0.0228.



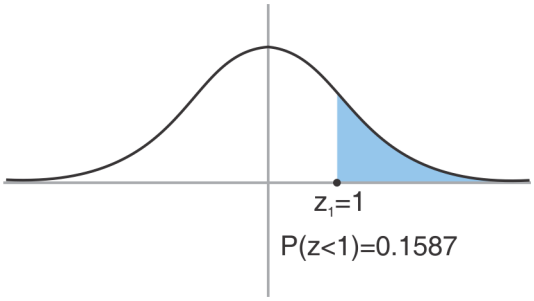


Ejemplo 7.1

Si la duración de la vida "T", en la población de palomas de Macondo tiene μ = 100 y σ = 5 y es DN.

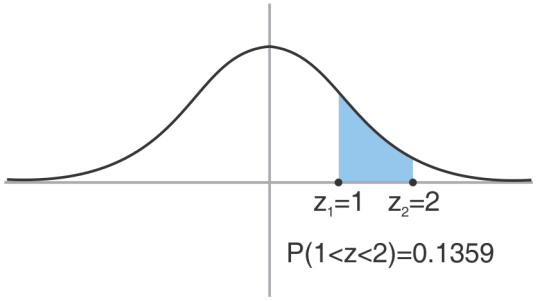
a) ¿Qué proporción de palomas viven entre 90 y 100? P (90 < X < 100) = P (–2 < Z < 0)= P (0 < Z < 2) = 0.5–P (Z > 2) = 0.5–0.0228 = 0.4772 (Ver figura anterior).

b) ¿Qué proporción de palomas viven más de 105? P (X > 105) = P (Z > 1)= 0.1587.

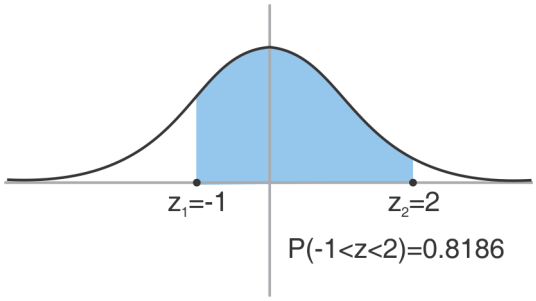


© Todos los derechos reservados

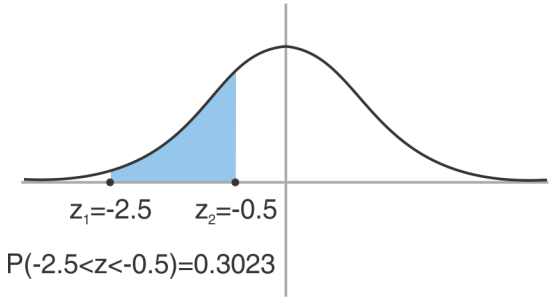
c) ¿Qué proporción viven entre 105 y 110? P (105 < X < 110)= P (1 < Z < 2) = P (Z > 1)–P (Z > 2) = 0.1587–0.0228 = 0.1359.



d) ¿Qué proporción viven entre 95 y 110? P (95 < X < 110) P (–1 < Z < 2) = P (–1 < Z < 0)+P (0 < Z < 2) = P (0 < Z < 1)+P (0 < Z < 2) [0.5–P (Z > 1)]+[0.5–P (Z > 2)] = 0.3413+0.4772 = 0.8185.



e) ¿Qué proporción viven entre 87.5 y 97.5? P (87.5 < X < 97.5) = P (–2.5 < Z < –0.5) = P (0.5 < Z < 2.5) = P (Z > 0.5)–P (Z > 2.5) = 0.3085–0.0062 = 0.3023.



En esta asignatura se anotará como "Z_α" el valor de Z que deja a su derecha probabilidad = α y la izquierda de su negativo otro tanto, de modo que entre ambas colas suman 2α de la probabilidad y por tanto, entre –Z y +Z queda el $1-2\alpha$ de la probabilidad de la Distribución Normal.

Por ejemplo:

© Todos los derechos reservados

$Z_{0.01} = 2.33$ porque $P(Z > 2.33) = 0.01 \rightarrow P(-2.33 < Z < 2.33) = 0.98$.

$Z_{0.05} = 1.65$ porque $P(Z > 1.65) = 0.05 \rightarrow P(-1.65 < Z < 1.65) = 0.90$.

$Z_{0.09} = 1.34$ porque $P(Z > 1.39) = 0.09 \rightarrow P(-1.34 < Z < 1.34) = 0.82$.

En otros libros se anota como $Z\alpha$ el valor de Z que deja a su derecha probabilidad= 2α. Ambos convenios son válidos si se recuerda lo que dicen.

7.3. DISTRIBUCIÓN DE LAS MEDIAS MUESTRALES EN EL MA. TEOREMA CENTRAL DEL LÍMITE

Vamos a estudiar cómo se distribuyen las medias muestrales, M, en el Muestreo Aleatorio, MA. La variable X en una población tiene Media y Desviación μ_x y σ_x. Hagamos MA y consideremos la población formada por las medias muestrales de muestras de N individuos:

$$M_1, M_2, M_3, \ldots\ldots, M_{1000}, \ldots\ldots, M_{1000\,000}, \ldots\ldots, M_{1000\,000\,000} \ldots$$

Tendremos tantas medias como muestras saquemos en el MA. Se trata de calcular, en esta larga lista de números, cuánto vale su media y su desviación. Cualquiera que sea la forma de la distribución de X y con cualquier N se cumple que la media y la desviación de las M son:

$$\mu_M = \mu_X \qquad \sigma_M = \frac{\sigma_X}{\sqrt{N}}$$

Ejemplo: Si X tiene en una población μ_X= 60 y σ_X= 6 y se hace MA de N= 4, μ_M= 60 y σ_M= 6/2= 3

Es decir, la media de todas las M´s es 60 (como la media de los individuos) y la σ entre las M´s es 3 (menor que la σ entre los individuos).

Pero conocer la μ_M y σ_M no permite saber la distribución, es decir, la proporción de muestras con media en cualquier intervalo que sea de interés. Esas proporciones se conocen si es Distribución Normal. Y la distribución de las medias muestrales es Normal si se da una de estas dos condiciones:

A) si X tiene DN en la población de individuos,

B) si el tamaño de las muestras es $N \geq 30$.

A) Si X tiene DN en la población de individuos, las M's también tienen DN

Fíjese que estamos hablando de dos distribuciones diferentes:

- La distribución de la variable X, con media μ_X y desviación σ_X, nos dice el porcentaje de individuos con X en cada intervalo.

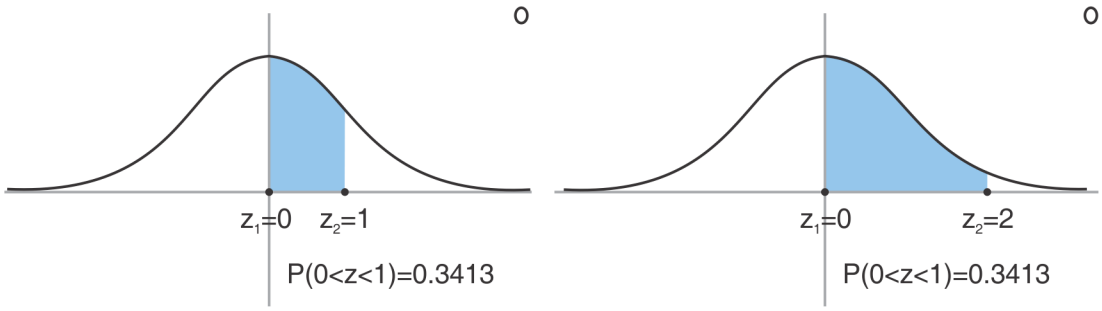
© Todos los derechos reservados

- La distribución de las medias de las muestras de N individuos, con media μ_M y desviación σ_M, nos dice el porcentaje de muestras con M en cada intervalo.

EJEMPLO 7.2

Si el peso de los individuos de una población tiene distribución N (100,24), la proporción de individuos con peso entre 100 y 112 es P(100 < Peso < 112) = P (0 < Z < 0.5) = 0.5–P (Z > 0.5) = 0.5–0.3085 = 0.1915.

- Si hacemos MA con N= 4, las medias muestrales, M, tienen μ_M = 100 y σ_M = 24/2 = 12. La proporción de muestras de 4 individuos con media entre 100 y 112 es P (100 < M < 112) = P (0 < Z < 1) = 0.5–P (Z > 1) = 0.5–0.1587 = 0.3413.
- Si hacemos MA con N = 9, las medias muestrales, M, tienen μ_M = 100 y σ_M= 24/3= 8. La proporción de muestras de 9 individuos con media entre 100 y 112 es P (100 < M < 112) = P (0< Z < 1.5) = 0.5–P (Z > 1.5) = 0.5–0.0668 = 0.4332.
- Si hacemos MA con N= 16, las medias muestrales, M, tienen μ_M = 100 y σ_M= 24/4 = 6. La proporción de muestras de 16 individuos con media entre 100 y 112 es P (100 < M < 112)= P (0< Z < 2) = 0.5–P (Z > 2) = 0.5–0.0228 = 0.4772.



B) Teorema Central del Límite, TCL: Si N ≥ 30, las medias muestrales, M, tienen, muy aproximadamente, DN, cualquiera que sea (Normal o no) la distribución de X en los individuos.

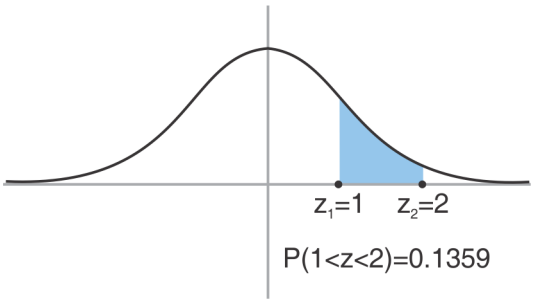
EJEMPLO 7.3

Si en otra población X no tiene DN y es μ_X= 100 y σ_X= 24 y hacemos MA de N= 64, las medias muestrales siguen DN (pues N ≥ 30) con μ_M= 100 y σ_M= 24/8= 3 y la proporción de muestras con media entre 97 y 100 es:

P (97 < M < 100)= P (–1< Z < 0) = 0.3413.

Y la proporción de muestras con media entre 103 y 106 es P (103 < M < 106)= P (1< Z < 2)= 0.1359.

© TODOS LOS DERECHOS RESERVADOS



El TCL constituye una herramienta fundamental en la Inferencia Estadística. Como acabamos de ver, dice que si N ≥ 30, en el MA las medias muestrales se distribuyen muy aproximadamente de modo Normal, cualquiera que sea la forma de la distribución en la población. Esto nos permite calcular el porcentaje de muestras con media comprendida en cierto intervalo.

	Distribución de las medias muestrales en el muestreo aleatorio	
	Siempre $\mu_M = \mu_X$ y $\sigma_M = \sigma_X/\sqrt{N}$	
	Distribución Nacional en la población	Distribución NO Normal en población
N < 30	Las M tienen distribución Normal	Las M tienen distribución NO Normal
N ≥ 30	LAs M tienen distribución Normal	Las M tienen distribución Normal **Teorema Central del Límite**

© Todos los derechos reservados

C) Si X no tiene DN en la población de individuos y es N < 30, las medias muestrales no tienen DN (pero se cumple que μM= μ y σM = $\sigma/\sqrt{N}$).

Ejemplo 7.4

Si H tiene distribución no Normal (100, 24), la proporción de individuos con H entre 100 y 112 es P (100< H <112)= P (0 < Z < 0.5) es NO CONOCIDA.

- Si hacemos MA con N= 4, es μ_M= 100 y σ_M= 24/2= 12. La proporción de muestras con M entre 100 y 112: P (100 < M < 112)= P (0< Z <1) es NO CONOCIDA.
- Si hacemos MA con N= 9, es μ_M= 100 y σ_M= 24/3= 8. La proporción de muestras con M entre 100 y 112 es P (100 < M < 112)= P (0< Z < 1.5) es NO CONOCIDA.
- Si hacemos MA con N= 36, es μ_M= 100 y σ_M= 24/6= 4 y tienen DN. La proporción de muestras con M entre 100 y 112 es P (100 < M < 112) = P (0< Z < 3)= 0.4987, pues puede aplicarse el TCL al ser N > 30.

- Si hacemos MA con N= 64, μ_M= 100 y σ_M= 24/8= 3 y tienen DN. La proporción de muestras con M entre 100 y 112 es P (0< Z < 4)= 0.49997, por el TCL.

7.4. ASPECTOS A RECORDAR

- Los valores estandarizados, Z, dicen cuantas desviaciones se aleja un cierto valor de la variable de la media del grupo. Se calcula $Z= (X-\mu)/\sigma$.
- Se dice que una variable tiene Distribución Normal, DN, cuando:
  - Por encima de la media más **0.02** desviaciones hay un **49.20%** de los valores.
  - Por encima de la media más **0.04** desviaciones hay un **48.40%** de los valores.
  - Por encima de la media más **0.06** desviaciones hay un **47.61%** de los valores.
  - Por encima de la media más **4** desviaciones hay **3** por **100.000** de los valores, y el resto de las proporciones que especifica la tabla correspondiente.
- Muchas variables cuantitativas de la naturaleza, tanto viva como inerte, se ajustan a esa Distribución Normal, es decir, los porcentajes de individuos con valor de la variable comprendido en cualquier intervalo son muy próximos a los que postula esa distribución.
- En el MA de tamaño N, las medias muestrales de variables numéricas, M, tienen media y desviación dadas por $\mu_M= \mu_X$ y $\sigma_M= \sigma_X/\sqrt{N}$, siendo μ_X y σ_X la media y desviación de la variable original.
- En el MA, las medias muestrales de variables numéricas, X, tienen distribución próxima a la DN –se puede calcular el porcentaje de muestras con medias en cualquier intervalo– si la variable es Normal (cualquiera que sea N) o si $N \geq 30$ (cualquiera que sea la distribución de X).
- En el Muestreo Aleatorio, MA, la proporción de individuos con cierta característica dicotómica tiene distribución próxima a la Normal si es $N \geq 30$. El número de individuos por muestra con la característica también tiene distribución próxima a la Normal.

© Todos los derechos reservados

© Todos los derechos reservados

Resumen

© Todos los derechos reservados

© Todos los derechos reservados

Inferencia estadística con una media

8

Visión general

- Definición del intervalo de confianza para una media.
- Definición de test de significancia para una media.
- Test de normalidad.
- Test no paramétricos.

Objetivo

- Conocer la inferencia estadística con una media.

© Todos los derechos reservados

8.1. INTERVALO DE CONFIANZA PARA UNA MEDIA

El comportamiento de una variable cuantitativa en una muestra se resume dando su Media Aritmética, M y su Desviación Estándar, S. La Inferencia Estadística, IE, intenta conocer acerca de la media poblacional, μ, a partir del conocimiento de la Media muestral. Se hace calculando el Intervalo de Confianza, IC, y el valor P del test para las hipótesis que puedan ser de interés.

El IC es más informativo y más fácil de interpretar que el valor P de los tests. Durante decenios los IC fueron infrautilizados y los tests fueron usados en demasía y muchas veces inapropiadamente. Esas distorsiones aún permanecen pero empiezan a corregirse en los últimos años.

El proceso lógico es el mismo que en el capítulo 7. El estudiante que necesite reforzar esos conceptos básicos aquí encontrará de nuevo su exposición detallada.

Si en una muestra de N individuos encontramos que una variable cuantitativa tiene media muestral M y desviación estándar muestral S, podemos calcular el IC para el valor de la media poblacional, μ. La fórmula utilizada es:

$$IC(\mu) \equiv M \pm Z \cdot S / \sqrt{N} = M \pm Z \cdot ES$$

En la que Z refleja la confianza con la que se quiere calcular el intervalo, como en el caso de los IC para la proporción poblacional.

- Cuando es N ≥ 30, Z es el valor estandarizado que en la distribución Normal, DN, deja entre él y su negativo una FR igual a la confianza que queremos para nuestro intervalo. Así, para IC al 95%, es Z = 1.96, porque por encima de Z = 1.96 está el 2.5% de los valores de la distribución Normal, por debajo de Z = –1.96 hay otro 2.5% (la DN es simétrica) y entre Z = –1.96 y Z = 1.96 están el 95% de ellos. Y para una confianza de 99% es Z = 2.58 (tabla al final de la asignatura). Los valores de Z para el cálculo del IC para una proporción poblacional (capítulo 7) se rigen por el mismo criterio.

  Por ejemplo, si en una muestra de N= 100 la variable X tiene M = 200 y S = 30, tenemos ES = 30 /10 = 3.

  $IC_{95\%}$ (μ) = 200±1.96 · 3 = 200±5.9 = 194.1 y 205.9

  $IC_{99\%}$ (μ) = 200±2.58 · 3 = 200±7.7 = 192.3 y 207.9

  Es decir, en la muestra la media de X es 200. No sabemos cuánto vale la media en la población, pero tenemos confianza 95% en que esté entre 194.1 y 205.9, y confianza 99% en que la media poblacional esté entre 192.3 y 207.9.

- Cuando es N < 30, el valor Z se obtiene de la tabla "t de Student" (al final de la asignatura) con N–1 grados de libertad. Por ejemplo, si en una muestra de N = 4 la variable "X" tiene M = 200 y S = 30, tenemos ES = 30/2= 15.

  $IC_{95\%}$ (μ) = 200±3.182 · 15 = 200±47.7 = 152.3 y 247.7

  $IC_{99\%}$ (μ) = 200±5.841 · 15 = 200±87.6 = 112.4 y 287.6.

  Siendo 3.182 el valor que se encuentra en la tabla "t" en la fila correspondiente a N–1 = 3 grados de libertad, gl. Dentro de la fila de 3 gl se toma el valor correspondiente a la columna que tiene "1 menos la confianza" como "bil" en la cabecera. En este caso, para confianza 0.95 se toma el valor de la columna que tiene 0.05 en la cabecera. Para confianza 0.99 se toma el valor de la columna que tiene 0.01 en la cabecera, que es 5.841. Tenemos confianza 95% en que la media de la población está entre152.3 y 247.7, y tenemos confianza 99% en que está entre 112.4 y 287.6.

8.2. TEST PARA UNA MEDIA CON σ CONOCIDA Y DN

Vemos en primer lugar la situación en que se conoce la σ de la variable en la población, lo cual es muy infrecuente en la práctica, pero ayuda a entender la situación habitual en la que no se conoce ese dato y se estima a partir de la muestra.

En todos los ejemplos que siguen vamos a asumir que en Alemania la variable "X" (número de horas pasadas ante el televisor por persona al año) tiene DN con media poblacional 500 y σ = 60.

© Todos los derechos reservados

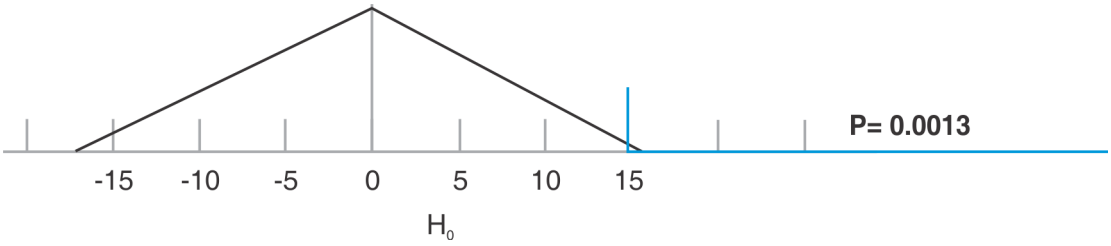
Sospechamos que en otros países la media poblacional es distinta de la alemana, es decir, no es 500. Para ver si es cierto, en cada país se toma una muestra aleatoria, se plantea la **Hipótesis Nula, H_0**, que dice que la media poblacional es 500 y se calcula el valor P del test, que es **la probabilidad de obtener una muestra con media como la de nuestro estudio o aun más alejada de 500**, si la media poblacional es 500. Cuanto menor sea el valor P más nos invita a rechazar la H_0 propuesta. Asumiremos que en las poblaciones que estamos investigando la σ es próxima a 60 y la variable tiene DN. Más adelante vemos qué hacer si no se dan estas condiciones.

a) Muestra de un solo individuo

1. Un individuo tomado al azar de la población belga tiene X = 530.

   Valor P del Test = $P(X \geq 530 | \mu_B = 500) = P(Z \geq 0.5) = 0.3085$.

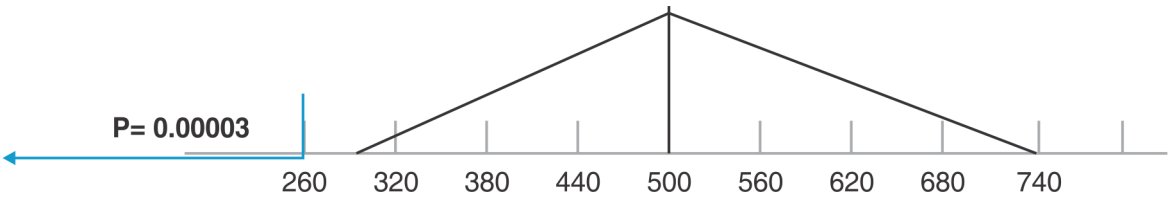
   Si la media poblacional belga es 500, $\mu_B = 500$, más de 30 de cada 100 individuos de la población tendrán valor 530 o mayor. El dato (X = 530) es compatible con la H_0 (media poblacional = 500) y por ello no invita a rechazarla.



2. Un individuo extraído al azar de la población corsa tiene X = 260.

   Valor P del Test = $P(X \leq 260 | \mu_C = 500) = P(Z \leq -4) = 0.00003$.

   Es muy difícil que en una población con media 500 se encuentre un individuo con X = 260 o menor. El dato (X = 260) es muy difícilmente compatible con la H_0 (media poblacional corsa = 500) y por ello invita a rechazarla. Creemos que es $\mu_C < 500$.



3. En una muestra de N = 100 franceses se encuentra media muestral, M = 479.

   → La Hipótesis Nula que planteamos en el test es H_0: $\mu_F = 500$.

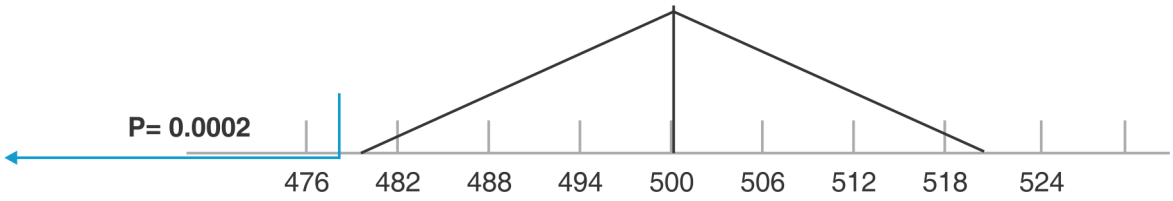
   → Si es cierta H_0, esperamos encontrar en la muestra una M próxima a 500.

   En el MA con N = 100 es $\sigma_{MEDIAS} = 60/\sqrt{100} = 6$.

   Si la variable tiene DN las medias del MA también la tienen. Calculamos el valor P del test, que lo llamaremos P_{UNI}.

   $Z = (479 - 500)/6 = -3.5 \rightarrow P_{UNI} = P(M_F \leq 479 | \mu_F = 500) = P(Z \leq -3.5) =$ **0.0002**.

© Todos los derechos reservados



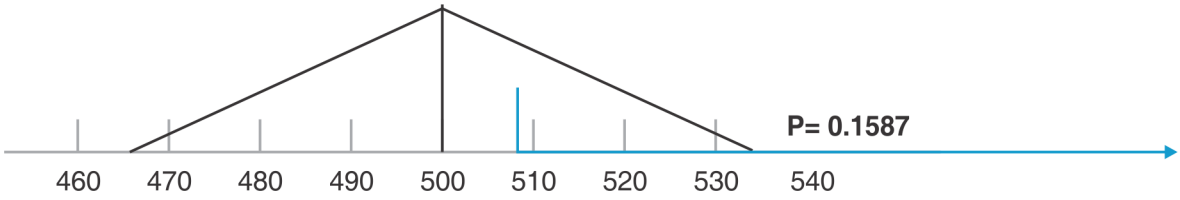
Si la media poblacional francesa fuera 500 es muy difícil que la muestra1 de N = 100 sea del orden de 479. Por eso rechazamos esa H_0 y creemos que es $\mu_F < 500$.

4. En una muestra de 36 georgianos se encuentra M = 508.

   En el MA con N = 36 es $\sigma_{MEDIA} = 60/\sqrt{36} = 10$.

   $Z = (508–500)/10 =$ **0.8** $\rightarrow P = P(M_G \geq 508 | \mu_G = 500) = P(Z \geq 0.8) =$ **0.2119**.

   Si la media poblacional georgiana fuera 500 es fácil que la muestra 1 de N= 36 sea del orden de 508. Por eso no rechazamos esa H_0 y creemos que puede ser $\mu_G = 500$.



Para la media muestra 1 de franceses, M = 479, se interpreta el valor P del test como la FR de estudios en los que la media muestra 1 es $M \leq 479$, mientras que para la media de los georgianos la P es la FR de estudios con $M \geq 508$. ¿Por qué unas veces se habla de media muestra 1 mayor y otras menor que la obtenida en la muestra? Depende de que la media muestra 1 sea menor o mayor que la propuesta en la H_0. El valor P del test es la FR de estudios con resultado como el obtenido o "más extremo", si H_0 es cierta, es decir si es $\mu = 500$. Para Francia con M = 479, resultados más extremos que el obtenido son aquellos con $M \leq 479$. Para Georgia, con M = 508, resultados más extremos son aquellos en que es $M \geq 508$.

Siguen otros ejemplos del valor P y la conclusión que se obtendría con algunos otros valores de la media muestra 1 M, si es N = 36 y $\sigma = 60$ ($\sigma_{MEDIA} = 10$) y si se plantea como H_0 que la media poblacional es $\mu = 500$:

Con M = 500 es Z = 0 $\rightarrow$ P = 0.50 $\rightarrow$ Puede ser $\mu = 500$.

Con M = 497 es Z = –0.3 $\rightarrow$ P = 0.3821 $\rightarrow$ Puede ser $\mu = 500$.

Con M = 490 es Z = –1 $\rightarrow$ P = 0.1587 $\rightarrow$ Puede ser $\mu = 500$.

Con M = 460 es Z = –4 $\rightarrow$ P = 0.00003 $\rightarrow$ Casi seguro que es $\mu < 500$.

Con M = 350 es Z = –15 $\rightarrow$ P = 10–40 $\rightarrow$ Seguro que es $\mu < 500$.

En estos ejemplos con N = 100 y N = 36 no se necesita que la variable tenga DN en población, pues el TCL nos dice que las medias del MA tienen DN, cualquiera que sea la distribución de la variable.

© Todos los derechos reservados

8.3. TEST CON UNA MEDIA NO CONOCIENDO σ

En todos los casos anteriores se calculó el número de desviaciones estándar que la media muestral (M) se aleja de la media que postula la H_0 (μ) con la siguiente expresión:

$$Z = \frac{\mu - M}{\sigma_M} = \frac{\mu - M}{\sigma / \sqrt{N}}$$

Donde σ_M es la desviación estándar entre las medias muestrales y σ es la desviación estándar entre los individuos de la población. Con el valor de Z así obtenido se calcula el valor P del test en las tablas de DN. Pero lo más habitual es que no se conozca la σ de la población y en esos casos se sustituye por la desviación estándar de la muestra, S, quedando la expresión:

$$Z = \frac{\mu - M}{S / \sqrt{N}} = \frac{\mu - M}{ES}$$

A la desviación estándar de la muestra dividida por la raíz cuadrada de N la llamamos **Error Estándar** o Típico de la muestra, ES. Cuando se introdujo el Error Estándar en el capítulo 2 solamente pudimos decir que era una cantidad importante en la Inferencia Estadística. Ahora ya sabemos su función: *es el mejor estimador de la desviación estándar entre las medias muestrales en el MA*.

En esta fórmula "Z" nos dice cuántos ES se aleja M de μ, cantidad que se parece mucho al número de σ_{MEDIA} que M se aleja de μ (con muestras "grandes", digamos mayores de 30, el ES se aproxima mucho a σ_{MEDIA}) y se mira en las tablas de DN para ver que probabilidad le corresponde.

Consideremos el ejemplo de una muestra de N = 81 individuos donde la variable "X" tiene **M = 60** y S = 27. El ES es 27/9 = 3.

1º. Para la Hipótesis Nula, H_0, que dice que la media poblacional es μ **= 62** es:

   $Z = (60-62)/3 = -0.67 \rightarrow P_{UNI} = P(Z \leq -0.67) =$ **0.25**.

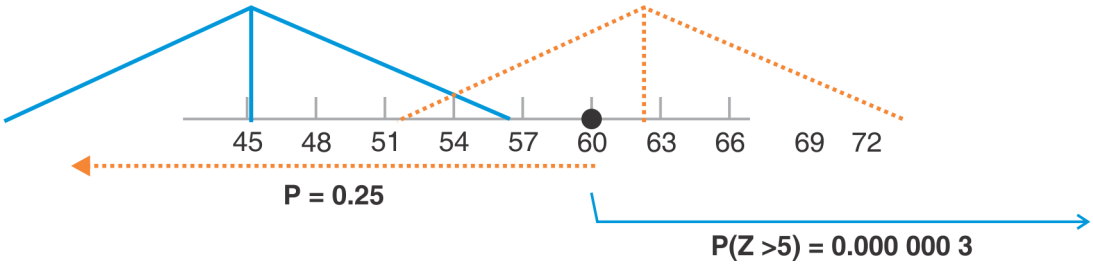
   Si fuera $\mu = 62$, y se repitieran muchos estudios como este, las M's tomarían valores en torno a 62, y en 25 de cada 100 muestras sería $M \leq 60$ (el valor de M obtenido en nuestra muestra). El dato es compatible con que en población sea $\mu = 62$.

2º. Para la H_0 que dice que la media poblacional es μ **= 45** es:

   $Z = (60-45)/3 = 5 \rightarrow P_{UNI} = P(Z \geq 5) =$ **0.000.000.3**.

   Si fuera $\mu = 45$, y se repitieran muchos estudios como este, las M's tomarían valores en torno a 45 y aparecería $M \geq 60$ solo en 3 de cada 10.000.000 estudios como este. El dato es muy difícilmente compatible con que en población sea $\mu = 45$.

© Todos los derechos reservados



8.4. TEST NO CONOCIENDO σ Y MUESTRAS PEQUEÑAS

Cuando es **N < 30** (muestra con menos de 30 individuos) hay que distinguir dos situaciones:

a) Si es razonable asumir que la variable tiene DN en la población de individuos se actúa como en el apartado anterior pero la probabilidad que corresponde al valor de "Z" obtenido se mira en la tabla t de Student con N–1 grados de libertad, gl, (al final de la asignatura) en lugar de usar la tabla de la Normal.

EJEMPLO 8.1

En una muestra de N = 4 se encuentra M = 50 y S = 8, es decir, ES = $8/\sqrt{4} = 4$. Si se plantea como H_0 que es $\mu = 45$, se calcula Z = (50–45)/4 =1.25.

Para buscar el valor de P en la tabla "t de Student" nos situamos en la fila de gl = 4–1 = 3 (la muestra es de N = 4) y en ella encontramos el valor 1.25 en la primera columna, que tiene como cabecera P_{UNI} = 0.15. Este es el valor P del test.

Si la media poblacional fuera 45, al hacer MA las medias muestrales tomarían valores próximos a 45. La media de las medias muestrales sería 45 y la desviación entre ellas sería próxima a 4. El valor 50 se aleja aproximadamente 1.25 desviaciones de la media y la probabilidad de que por azar aparezca una muestra con media 50 o mayor es 0.15. No se rechaza esa hipótesis.

$P(M \geq 50 | \mu = 45) = P(Z \geq 1.25) = 0.15$ en tablas t de Student con 3 gl.

EJEMPLO 8.2

Si con la muestra del ejemplo anterior, se plantea como H_0 que es $\mu = 100$, es Z = (100–50)/4 = 12.5.

En la tabla "t de Student", en la fila de gl = 3 encontramos el valor 12.94 en la última columna a la derecha, que es próximo a 12.5 y tiene como cabecera P_{UNI} = 0.001. Este es el

© TODOS LOS DERECHOS RESERVADOS

valor P del test. Es decir, si la media poblacional fuera 100, al hacer MA la media de las medias muestrales sería 100 y la desviación entre ellas sería próxima a 4. El valor 50 se aleja aproximadamente 12.5 desviaciones de la media y la probabilida de que por azar aparezca una muestra con media 50 o menor es aproximadamente 0.001. Se rechaza esa H_0. Será $\mu < 100$.

$P(M \leq 50 | \mu = 100) = P(Z \geq 12.5) \approx 0.001$.

Fíjese que en la tabla de t de Student, para cualquier fila, cualquier gl, a medida que aumentan los valores del interior de la tabla disminuyen los correspondientes valores de P_{UNI}. Por ejemplo, para 5gl, al valor Z = 1.156 le corresponde $P_{UNI} = 0.15$ y a Z = 1.476 le corresponde $P_{UNI} = 0.10$,... Si en un estudio con 6 individuos se obtiene, por ejemplo, Z = 1.3, le corresponderá $0.10 < P_{UNI} < 0.15$, y pondremos $P_{UNI} > 0.10$, que es lo relevante para no rechazar H_0.

Ejemplo 8.3

En una muestra de N = 9 se encuentra M = 200 y S = 45, es decir, $ES = 45/\sqrt{9} = 15$. Si se plantea como H_0 que es $\mu = 290$, es $Z = (290 - 200)/15 = 6$.

En la tabla "t de Student", en la fila de gl = 8 encontramos el valor 5.04 en la última columna a la derecha, que tiene como cabecera $P_{UNI} = 0.001$. Si hubiéramos obtenido Z = 5.04, sería P = 0.001, pero como es Z > 5.04, es P < 0.001. Si fuera $\mu = 290$, al hacer MA la media de las medias muestrales sería 290 y la desviación entre ellas sería próxima a 15. El valor 200 se aleja aproximadamente 5 desviaciones de la media y la probabilidad de que por azar aparezca una muestra con media 200 o menor es menor de 0.001.

$P(M \leq 200 | \mu = 290) < 0.001$. Se rechaza esa H_0. Será $\mu < 290$.

Ejemplo 8.4

Si con la muestra del ejemplo anterior se plantea como H_0 que es $\mu = 180$, es $Z = (180-200)/15 = 1.33$.

En la tabla "t de Student", en la fila de gl = 8 encontramos el valor 1.39 en la columna que tiene como cabecera $P_{UNI} = 0.10$. Si hubiéramos obtenido Z = 1.33, sería P = 0.10, pero como es Z < 1.33, es P > 0.10. Si fuera $\mu = 180$, al hacer MA la media de las medias muestrales sería 180 y la desviación entre ellas sería próxima a 15. El valor 200 se aleja aproximadamente 1.33 desviaciones de la media y la probabilidad de que por azar aparezca una muestra con media 200 o mayor es mayor de 0.10.

$P(M \geq 200 | \mu = 180) > 0.010$. No se rechaza esa H_0, puede que sea cierta.

© Todos los derechos reservados

b) En teoría, con muestras pequeñas (menores de 30) estos cálculos basados en la tabla t de Student, que llamaremos "test-t" solamente son válidos si la variable en cuestión sigue DN en la población. Pero eso no ocurre exactamente en ninguna población real y en la práctica es difícil saber cuánto se aparta de la Normalidad la variable implicada en nuestro estudio. Por ello lo habitual es asumir que no se aparta sensiblemente y usar este test basado en la tabla t de Student. En un apartado posterior se comenta este punto más detalladamente.

8.5. VALOR P A UNA COLA Y A DOS COLAS

Consideremos el ejemplo anterior de una muestra de N = 36 con M = 490. Si se sabe que la σ entre los individuos es 60, es $\sigma_{MEDIA} = 60/6 = 10$ y para la H_0 que dice que $\mu = 500$ encontramos $Z = -1$ y $P = 0.1587$, que es la probabilidad de encontrar una muestra con media 490 o menor, si es $\mu = 500$. Al valor P así obtenido se le llama $P_{unilateral}$ (abreviadamente P_{UNI}), porque da la probabilidad de muestras con media que se alejen de 500 tanto como la de nuestra muestra o más aún, ***en el mismo sentido***, es decir, en el mismo "lado" de la distribución.

A veces se utiliza la llamada $P_{bilateral}$ (abreviado P_{bil}) que es el doble del P_{UNI}, y en este caso indica la probabilidad de encontrar muestras con media 490 o menor o con media 510 o mayor, si es $\mu_E = 500$ (510 se aleja de 500 lo mismo que 490, pero en el otro sentido). Es decir:

$$P_{Unilateral} = P\,(M \leq 490 \mid \mu = 500) = P\,(Z \leq -1) = 0.1587.$$

$$P_{Bilateral} = P\,(M \leq 490 \text{ o } M \geq 510 \mid \mu_G = 500) = P\,(Z \leq -1) + P\,(Z \geq 1) = 0.3174.$$

En general, se llama valor P bilateral a la probabilidad de obtener una muestra en que el valor del estadístico se aleje de lo esperado bajo la H_0 tanto como el de nuestro estudio o aún más ***en cualquier sentido*** y es $P_{bil} = 2 \cdot P_{UNI}$.

Cuando hablamos del valor P del test podemos referirnos a la unilateral o a la bilateral. En esta asignatura la mayor parte de las veces manejamos la P unilateral. Si se desea la bilateral basta con multiplicar por dos la unilateral.

8.6. INTERVALO DE CONFIANZA VERSUS TESTS ESTADÍSTICOS

En la mayoría de los casos el Intervalo de Confianza, IC, es más informativo y de interpretación más directa que el valor P del test, pues además de orientarnos sobre si hay motivos fuertes o no para rechazar la H_0, nos informa sobre la magnitud del efecto que estamos estudiando.

Para ver la relación entre el valor P y los límites del IC supongamos que la media poblacional de "X", número de horas que cada persona dedica al año a hacer deporte, era 400 en la población. Se hizo una campaña de educación para la salud animando a los ciudadanos a hacer más deporte; esta es útil si tras ella la μ de X es mayor de 400. Para investigarlo medimos, tras la

© Todos los derechos reservados

campaña, X en una muestra de 49 individuos y planteamos como H_0 que la campaña no fue efectiva y sigue siendo $\mu X = 400$. En la tabla que sigue damos los valores que tomaría el valor P y los IC al 95% y al 99% ***para distintos valores de la media*** muestral, M, si en la muestra es S = 70 (ES = 10). Recuerde que es IC = M±Z · ES, con Z = 1.96 ó 2.58 dependiendo de la confianza, obtenido de las tablas de distribución Normal, pues N > 30.

M	2	P_{UNI}	P_{BIL}
410	1	0.1587	0.3174
416.5	1.64	0.0505	0.1010
419.6	1.96	0.0250	0.0500
423.3	2.33	0.01	0.02
425.8	2.58	0.005	0.01
440	4	0.00003	0.00006

$IC_{95\%}$ = M ± 19.6	$IC_{99\%}$ = M ± 25.8
390.4 ↔ 429.6	384.2 ↔ 435.8
396.8 ↔ 436.0	390.6 ↔ 442.2
400 ↔ 439.2	393.8 ↔ 445.4
403.7 ↔ 442.9	397.5 ↔ 449.1
406.2 ↔ 445.4	**400** ↔ 451.6
420.4 ↔ 459.6	414.2 ↔ 465.8

Comentamos algunos hechos que son de validez general y que pueden observarse en los ejemplos anteriores:

1. Cuando el IC95% incluye el valor que propone la H_0, es $P_{UNI} > 0.025$ para esa H_0.
2. Cuando un extremo del IC95% coincide con el valor que propone la H_0, la P bilateral es 0.05 y la unilateral es 0.025.
3. Cuando un extremo del IC99% coincide con el valor que propone la H_0, la P_{BIL} es 0.01 y la unilateral es 0.005.
4. En general, cuando un extremo del IC$\propto$% coincide con el valor que propone la H_0, la P_{BIL} es $1-\propto$ y la $P_{unilateral}$ es $(1-\propto)/2$.
5. Cuando el IC99% no incluye el valor que propone la H_0, es $P_{UNI} < 0.005$. Para ilustrar cómo el IC nos informa sobre la magnitud del efecto estudiado y el valor P no lo hace, veamos los datos de una muestra tomada en Oviedo (con M = 410) y otra tomada en Cartagena (con M = 460), ambas de N = 1225, con S = 70 y por ello ES = 70/√1225 = 70/35 = 2. Para ambas se hace test para H_0: $\mu = 400$.

	N	M	S	ES	Z	P_{UNI}
Oviedo	1225	**410**	70	2	5	$3 \cdot 10^{-7}$
Cartagena	1225	**460**	70	2	30	10^{-100}

$IC_{99\%}$ = M ± 5.2
404.8 ↔ 415.2
454.8 ↔ 465.2

En ambas el valor P es muy pequeño y nos da la seguridad de que la media poblacional tras la campaña no es 400 sino mayor, es decir, que la campaña ha sido efectiva. Pero solamente los IC nos permiten saber que en Oviedo el efecto producido por la campaña es moderado, pues el IC indica que la media poblacional está entre 404.8 y 415.2, es decir, casi entre 5 y 15 horas más que antes de la campaña, mientras que en Cartagena el efecto fue mucho mayor: el aumento del número medio de horas está entre 54.8 y 65.2, aproximadamente.

© Todos los derechos reservados

Con muestras muy grandes un pequeño efecto puede ser muy significativo y el investigador debe distinguir claramente entre la magnitud del efecto y su significación estadística (valor P del test). Y a ello ayuda decisivamente el IC.

8.7. MUESTRAS MUY PEQUEÑAS

Es frecuente que se ponga en duda la validez de una conclusión alegando que el estudio contenía un número demasiado pequeño de individuos. Para rebatir esa objeción hay que darse cuenta de que el valor P del test se calcula teniendo en cuenta el tamaño de la muestra, de modo que indica la probabilidad de encontrar ese resultado o uno más extremo en una muestra de ese tamaño, si la H_0 es cierta.

Los valores de P muy pequeños constituyen evidencia contra la Hipótesis aunque en la muestra haya muy pocos individuos.

8.8. AUSENCIA DE NORMALIDAD Y TESTS NO PARAMÉTRICOS

En teoría, con muestras pequeñas (menores de 30) estos cálculos basados en la distribución t de Student, que llamaremos "test–t" solamente son válidos si la variable en cuestión sigue DN en la población. Si no se cumple esa condición el valor P del test–t, que llamaremos "P–t", no es exactamente igual al "P–verdadero", que es la probabilidad de obtener muestras con media que se aleje de la poblacional propuesta por la H_0 tanto como la de nuestro estudio o más aun.

Muchos libros dicen (o en su defecto sus lectores interpretan) que hay tests, llamados de Normalidad, que a partir de los datos de la muestra permiten saber si la variable tiene o no DN en la población. De ese modo, si la variable sigue DN se puede aplicar el test–t para una media y si no sigue DN hay que aplicar otros tests llamados "no paramétricos", que no exigen DN. Pero esa frase es una simplificación equívoca si se toma como receta ciega que sustituya la comprensión del tema. En la práctica esa condición no debe tomarse como restricción muy estricta y hay que tener en cuenta los siguientes matices importantes:

1. Ninguna variable de la naturaleza se ajusta exactamente a la DN. La DN es un modelo matemático al que muchas distribuciones reales de la Naturaleza se aproximan lo suficiente como para poder asumir que siguen esa distribución, aunque ninguna de ellas la seguirá con total exactitud (como ningún ángulo físico será exactamente recto, aunque muchos lo son tan aproximadamente que se les puede tratar como tales y utilizar con ellos el teorema de Pitágoras con provecho práctico).
2. El test–t es apropiado cuando la variable tiene distribución muy próxima a la Normal. Pero ningún test de normalidad puede confirmar, con los tamaños habituales en las investigaciones, que tiene en población distribución muy próxima a la Normal. Solamente pueden –si aparece valor P muy pequeño– confirmar que no hay normalidad.

© Todos los derechos reservados

3. Con tamaños de muestras menores de 30, que es en las que nos preocupa la no normalidad en población (pues en ellas no se puede aplicar el TCL), esos tests tienen poca potencia estadística para detectarla y muchas veces se les escapa. Es decir, aunque una variable tenga en la población distribución muy alejada de la normal, los tests de normalidad lo detectan pocas veces.
4. Aunque la variable tenga en población distribución moderadamente alejada de la Normal, el "P–t" es muy próximo al "P–verdadero". A esta característica se le llama "***Robustez***" frente a la no normalidad.
5. Recordemos que si al test de comparación de medias le correspondiera, por ejemplo, P–verdadero = 0.043 y el P–t es 0.064, la diferencia no es relevante. Ambas cifras dicen que la evidencia contra la H_0 es moderada. Que uno esté un poco por debajo del 0.05 y otro un poco por encima de él no implica que la conclusión sea esencialmente diferente.
6. Incluso diferencias considerablemente grandes entre el P–verdadero y el P–t son irrelevantes si ambos valores llevan claramente a la misma conclusión. Por ejemplo, si fuera P–verdadero = 0.000080 y P–t = 0.000002, el primero es 40 veces mayor que el segundo, pero ambos nos llevan claramente a rechazar H_0. Y si fuera, por ejemplo, "P–verdadero = 0.33 y P–t = 0.11, ambos nos llevan a no rechazar la H_0.
7. Como consecuencia de los apartados 4, 5 y 6 lo más frecuente es que el test no paramétrico lleve a la misma conclusión que el test–t, bien porque ambos dan valor P grande (se acepta como posible la H_0) o ambos lo dan pequeño (se rechaza la H_0) o ambos lo dan intermedio y no cabe pronunciarse. Por todo ello, en la mayoría de los casos se aplica el test–t, asumiendo que las conclusiones a que lleva no se alejarán sensiblemente de las que se habrían alcanzado con el P-verdadero.

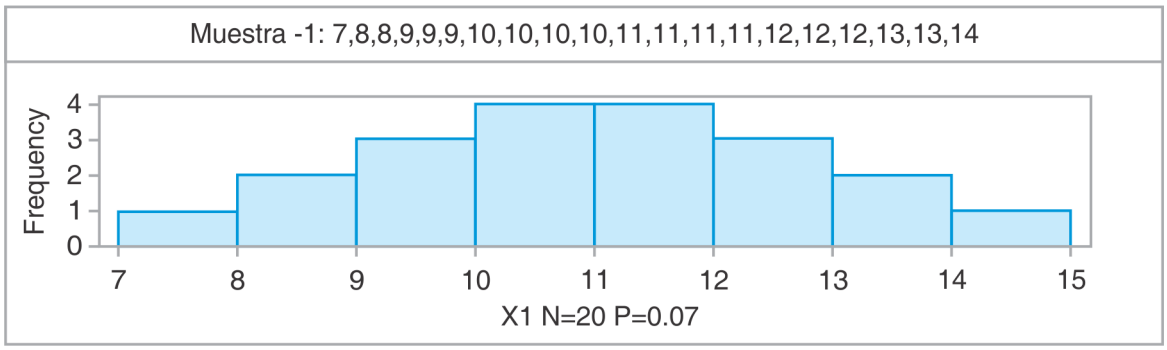
© Todos los derechos reservados

8.9. TEST DE NORMALIDAD

Los ejemplos que siguen ayudan a ver que el test de normalidad da menor valor P cuanto más se apartan los datos muestrales de la normalidad. Y también muestran que incluso con muestras que se apartan notablemente de la normalidad, el test no lo detecta, es decir no da un valor P muy pequeño que invite claramente a rechazar la hipótesis de normalidad en la población. Los gráficos representan distribuciones de frecuencias de los 20 valores de cada muestra de modo que serán distribuciones tanto más parecidas a la Normal cuando más se parezcan a la curva Normal, esto es, barras altas en torno a la media y menores a medida que nos alejamos de ella.

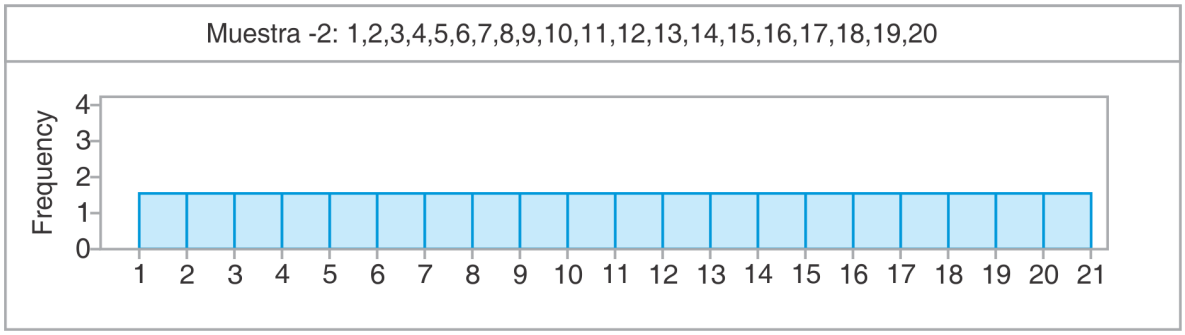
Ejemplo 8.5

El gráfico que sigue corresponde a una muestra de 20 valores (aparecen encima del gráfico) con distribución muestral parecida a la DN (cúmulo de datos en torno a la media y escasez de ellos en las colas). La altura de cada barra es proporcional al número de veces que se repite cada valor. El test de normalidad plantea que en la población la variable es Normal y da P = 0.87, un valor P grande, como cabía esperar.



Ejemplo 8.6

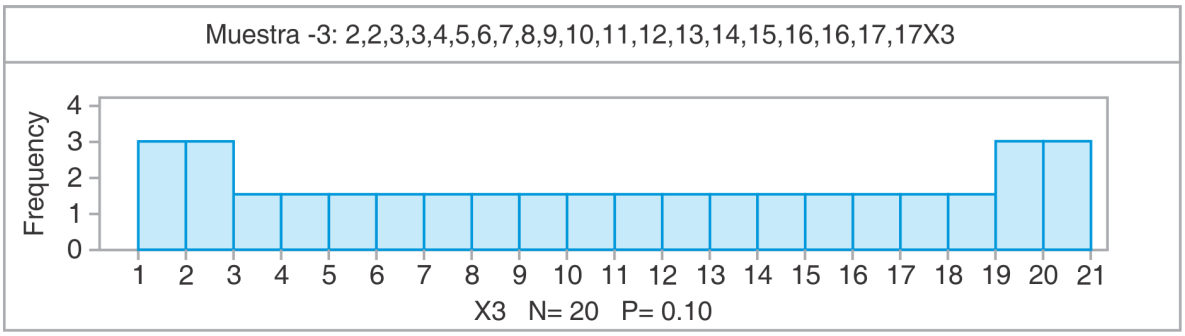
El gráfico siguiente corresponde a una muestra de 20 valores con distribución muestral plana, bastante alejada de DN. A pesar de ello el test de normalidad da P = 0.55, lo que no invita a rechazar la hipótesis de normalidad.



Ejemplo 8.7

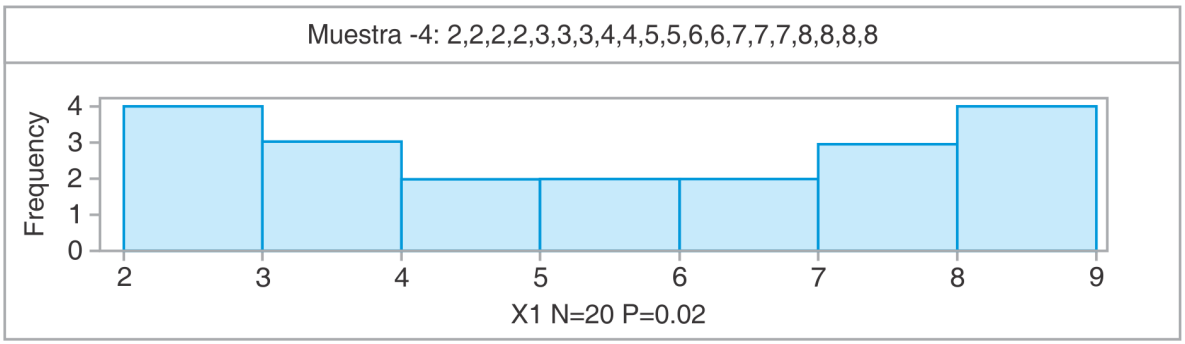
El gráfico que sigue corresponde a una muestra de 20 valores con distribución muestral que presenta más acumulación de valores en los extremos, al revés que la DN. El test de normalidad da P = 0.10, que tampoco invita a rechazar la hipótesis de normalidad.

© Todos los derechos reservados



EJEMPLO 8.8

El gráfico que sigue corresponde a una muestra de 20 valores con distribución muestral que presenta más acumulación de valores en los extremos que en la muestra anterior, es decir, más alejada de la normalidad. El test de normalidad da P = 0.02, que invita a rechazar la hipótesis de normalidad en la población, pero no con evidencia tan fuerte como podría pensarse al observar la muestra.



Estos ejemplos muestran que la mayoría de las muestras pequeñas, cualquiera que sea su composición, son compatibles con que la población tenga DN y por ello no constituyen argumento fuerte en contra del uso del test de Student.

8.10. TESTS NO PARAMÉTRICOS

Para el caso que nos ocupa en que la H_0 atribuye un valor a la media poblacional de una variable, un posible test no paramétrico sería el llamado Test de la Mediana que plantea como H_0 que la mediana de esa variable en la población es ese valor, y calcula el valor P del test con la Distribución Binomial.

© TODOS LOS DERECHOS RESERVADOS

Cuando se aplica este test no paramétrico es frecuente que lleve a la misma conclusión que el paramétrico, como en este ejemplo de 20 valores:

2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 7, 8, 8, 8, 8 → Media = 5 y ES = 0.51.

Ejemplo 8.9

Para la H_0: media poblacional = 2.5 el test–t da Z = 4.9 con 19 gl y P = 0.00005. Podemos plantear como H_0 equivalente a la anterior (aunque no exactamente igual): Mediana poblacional = 2.5 y hacer un test basado en la Binomial. Si es cierta esta H_0 la mitad de los valores de la población son mayores que 2.5. Por ello esperamos que en la muestra sean mayores que 2.5 la mitad de los valores, es decir, 10. En la muestra vemos 16 valores mayores de 2.5. El valor P del test se calcula con la Binomial de N = 20 (tamaño de la muestra estudiada) y π = 0.5 (esperamos la mitad de los individuos con valor mayor de 2.5). Si en población es π = 0.5, la probabilidad de obtener muestras con 16 o más valores mayores de 2.5, en una muestra de N = 20 es P = 0.005.

Este último test infrautiliza la información, pues ignora el valor propio de cada individuo y se limita a considerar si es o no mayor de 2.5. Aunque la variable no tenga DN en población cabe pensar que el valor P = 0.00005 refleja mejor la realidad que el P = 0.005.

Ejemplo 8.10

Consideremos ahora la H0: media poblacional = 4.5 que con el test–t da Z = 0.98 con 19 gl y P = 0.17. Para la H_0: Mediana poblacional = 4.5, esperamos en la muestra 10 valores mayores de 4.5, pero hay 11. Con Binomial de N = 20 y π = 0.5, la probabilidad de obtener muestras con 11 o más casos mayores de 4.5 es P = 0.41. Ambos tests (Student y de la mediana) coinciden en considerar que los datos no son evidencia contra la H_0.

Ejemplo 8.11

Consideremos ahora la H_0: media poblacional = 3.5 que da Z = 2.94 con 19 gl y P = 0.004. Si la H_0: Mediana poblacional = 3.5 es cierta, esperamos en la muestra 10 valores mayores que 3.5, pero hay 13. Con Binomial de N = 20 y π = 0.5 la probabilidad de obtener muestras con 13 o más valores mayores de 3.5 es P = 0.130. También en este caso el test de la mediana infrautiliza la información, pues ignora el valor propio de cada individuo y se limita a considerar si es o no mayor de 3.5. El hecho de que el valor P no invite a rechazar la H_0 no sería una referencia vinculante. Aun en el caso de que la variable no tenga DN en la población el valor P = 0.004 reflejaría mejor la realidad que el valor P = 0.130.

© Todos los derechos reservados

8.11. ESTIMADORES SESGADOS Y ESTIMADORES INSESGADOS. VARIANZA MUESTRAL = SC/(N-1)

Si hacemos Muestreo Aleatorio de una población y en cada muestra se calcula un estadístico tendremos tantos valores de él como muestras se han tomado.

- Un estadístico se dice que es un estimador insesgado de cierto parámetro si la media de los valores del estadístico evaluada sobre todas las muestras del MA coincide con el valor del parámetro.
- Un estadístico se dice que es un estimador sesgado de cierto parámetro si la media de los valores del estadístico no coincide con el valor del parámetro.

La media muestral es estimador insesgado de la media poblacional

Si en cada muestra del MA se calcula la media, algunas de las medias muestrales tienen valor mayor que la media poblacional, otras medias muestrales tienen menor valor que la media poblacional y se puede demostrar que la media de todas las medias muestrales coincide con el valor de la media poblacional.

La varianza muestral es estimador sesgado de la varianza poblacional

Si en cada muestra se calcula la varianza, SC/N, algunas de esas varianzas muestrales tendrán valor mayor que la σ^2, otras tendrán menor valor que la σ^2 y se demuestra que la media de todas ellas no coincide con el valor de la σ^2.

La cuasivarianza muestral es estimador insesgado de la varianza poblacional

Si en cada muestra se calcula la cuasivarianza, S^2 = SC/(N–1), algunas de esas S^2 tendrán valor mayor que la σ^2, otras S^2 tendrán menor valor que la σ^2 y se demuestra que la media de todas las S^2 coincide con el valor de la σ^2.

Por ejemplo, si de una población con σ^2 = 10 se toma millones de muestras de N = 4, la media de SC/N –1 = SC/3 es 10, y la media de SC/N = SC/4 es 7.5, como se muestra en la siguiente tabla.

Se toman millones de muestras de N = 4 de una población con σ^2 = 10			
	SC	SC/N = SC/4	SC/N–1 = SC/3
Muestra 1ª	32	8.00	10.67
Muestra 2ª	29	7.25	9.67
Muestra 3ª	27	6.75	9.00
Muestra 4ª			
Muestra 5ª			

© Todos los derechos reservados

Se toman millones de muestras de N = 4 de una población con $\sigma^2 = 10$			
...			
Muestra 1 000 000^{ésima}	35	8.75	11.67
		↓ Media = 7.5	↓ Media = 10

8.12. ASPECTOS A RECORDAR

- A partir del valor de la media muestral hacemos Inferencia para la media poblacional calculando el IC y el valor P de los tests.
- Los IC son más informativos y fácilmente interpretables por el investigador que los tests: cuando el valor de la media poblacional nos informa sobre el efecto que hace cierto factor, el test únicamente nos dice, si es el caso, que en la población el efecto no es cero, pero no nos informa sobre la magnitud de ese efecto. El IC informa sobre la magnitud del efecto que hace el factor.
- El valor P es la proporción de valores que en la DN son iguales o mayores que $Z = (\mu - M)/ES$
- Con muestras con menos de 30 individuos, para encontrar el valor P de los tests y los límites del IC se utilizan las tablas t de Student.
- El valor P del test que utiliza la distribución t de Student indica exactamente lo que esperamos, si la variable tiene en la población DN, pero nunca podemos tener la seguridad de que se cumple esa condición.
- Aunque la variable tenga en población distribución considerablemente alejada de la Normal el "P–verdadero" no es muy lejano al "P–t". A esta característica se le llama "Robustez" frente a la no normalidad.
- Pueden aplicarse otros tests que no exigen Normalidad y por ello su resultado no se afecta si no se da esa condición. Se les llama "tests no paramétricos" y son menos potentes, es decir, cuando la H_0 no se cumple tienen menor capacidad para detectarlo.
- Cuando se aplica el test no paramétrico lo más frecuente es que lleve a la misma conclusión que el paramétrico.
- Por todo lo anterior, es razonable usar el test de Student en prácticamente todos los casos.
- El valor P del test no es la probabilidad de que la H_0 sea cierta o sea falsa, es la probabilidad de obtener el tipo de resultado encontrado en nuestro estudio si fuera cierta la H_0.
- Si se conoce la probabilidad inicial de que H_0 sea cierta, se puede calcular la probabilidad de que lo sea teniendo en cuenta el resultado del estudio.
- La Inferencia Bayesiana, puede enriquecer el análisis de la información en muchos casos. Su presencia en las publicaciones científicas es aún escasa, pero irá aumentando progresivamente en el futuro inmediato.

© Todos los derechos reservados

Resumen

© Todos los derechos reservados

© Todos los derechos reservados

La inferencia estadística con dos medias

9

Visión general

- Definición del intervalo de confianza para la diferencias de las medias muestrales.
- Definición del test de significancia para la diferencia de medias muestrales.
- Test no paramétricos.

Objetivo

- Conocer la inferencia estadística con dos medias.

© Todos los derechos reservados

9.1. IC PARA DIFERENCIA DE DOS MEDIAS

Muchos profesores de idiomas sospechan que la facilidad para aprender una segunda lengua es mayor en mujeres que en varones. Para ver que la media de la población de mujeres es mayor que la media de la población de varones se hizo un estudio con 49 mujeres y 64 varones, evaluando esa facilidad mediante pruebas que daban resultado entre 0 y 200. El estudio se hizo en países con idiomas muy distintos: China, España y Rusia.

En China dio resultados que no apoyaban claramente esa hipótesis:

China	N	Media, M	S	ES	M_{MUJ} - M_{VAR}
Mujeres	49	**123**	21	3	**3**
Varones	64	**120**	32	4	

Pero en España dio resultados muy contundentes a favor de ellas:

España	N	Media, M	S	ES	M_{MUJ} - M_{VAR}
Mujeres	49	**140**	21	3	**50**
Varones	64	**90**	32	4	

Finalmente, en Rusia los resultados fueron intermedios:

Rusia	N	Media, M	S	ES	M_{MUJ} - M_{VAR}
Mujeres	49	**125**	21	3	**15**
Varones	64	**110**	32	4	

En este capítulo vemos cómo los cálculos estadísticos nos pueden ayudar a interpretar mejor los resultados en situaciones como la de Rusia, donde los datos no permiten en principio decantarse claramente por una u otra postura. La diferencia de medias muestrales es 15 pero no conocemos el valor de la diferencia entre las medias poblacionales. Haremos la Inferencia Estadística calculando el **Intervalo de Confianza** para la diferencia de medias poblacionales y el **valor P del test** que propone un determinado valor para esa diferencia. Todos los paquetes de programas estadísticos dan esos resultados. El investigador necesita interpretarlos correctamente y para ello no es necesario conocer las fórmulas usadas, pero se dan aquí porque no son muy complejas y ayudarán al estudiante interesado en entender los fundamentos lógicos de la Inferencia.

A partir de los datos muestrales se calcula el intervalo dentro del cual tenemos confianza en que se encuentre la diferencia de las medias poblacionales.

$$IC(\mu_1 - \mu_2) = (M_1 - M_2) \pm Z \cdot E$$

Donde E se calcula de diferente forma dependiendo del tamaño de las muestras:

$$E = \sqrt{ES^2{}_1 + ES^2{}_2}$$

E = (muestras grandes, N1 ≥ 30 y N2 ≥ 30).

$$E = \sqrt{\frac{SC_1 + SC_2}{N_1 + N_2 - 2}} \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

E = (muestras pequeñas, N1 < 30 y/o N2 < 30).

© Todos los derechos reservados

Y Z indica la confianza que queremos para el intervalo, igual que en el IC para una media o una proporción, en los capítulos anteriores. Para **muestras grandes** se usa la tabla de la DN en la que Z = 1.96 para confianza 95% y Z = 2.58 para confianza 99%.

En el ejemplo es $E = (3^2+4^2)^{½} = 5$ y $(M_1–M_2) = 15$.

$IC_{95\%}$ (μ1–μ2) = 15±1.96 · 5 = 15±9.8 = **5.2** y **24.8**.

$IC_{99\%}$ (μ1–μ2) = 15±2.58 · 5 = 15±12.9 = **2.1** y **27.9**.

Es decir, tenemos confianza 95% en que la media poblacional de esta variable sea en las mujeres entre 5.2 y 24.8 unidades mayor que en los varones.

Con **muestras pequeñas** el valor de Z se toma de las tablas de la distribución "t de Student" con grados de libertad "$N_1+N_2 –2$".

Por ejemplo, si los tamaños de las muestras del ejemplo anterior hubieran sido 16 mujeres y 9 varones, y los resultados:

N_1 = **16** mujeres: M_1 = **79**, SC_1 = 6615, S_1 = 21, ES_1 = 5.25.

N_2 = **9** varones: M_2 = **73**, SC_2 = 8192, S_2 = 32, ES_2 = 10.67.

$$E = \sqrt{\frac{6615+8192}{16+9-2}}\sqrt{\frac{1}{16}+\frac{1}{9}} = 10.57$$

En la tabla "t de Student", en el cruce de la fila de gl = 16+9–2 = 23 y la columna encabezada por "BIL" igual a "uno menos la confianza", vemos que para el IC95% es Z = 2.069 y para el IC99% es Z = 2.807.

$IC_{95\%}$ (μ1–μ2) = 6 ± 2.069 · 10.57 = 15 ± 21.1 = – **6.1 y 36.1**.

Para interpretar correctamente los límites con signo negativo debemos tener en cuenta el orden en el que hemos restado. Así, tenemos 95% de confianza en que $\mu_{MUJ}–\mu_{VAR}$ esté entre –6.1 y 36.1, es decir, puede que la media poblacional de mujeres supere a la de varones en 36.1, o que esté 6.1 por debajo de ella, o que se diferencien cualquier cantidad entre las dos anteriores.

9.2. TEST PARA LA DIFERENCIA DE DOS MEDIAS

Si planteamos como H_0 que en la población **rusa** no hay diferencias entre sexos, es decir, que las medias poblacionales de ambos sexos son iguales, H_0: $\mu_1 = \mu_2$ o $\mu_1–\mu_2 = 0$, se calcula:

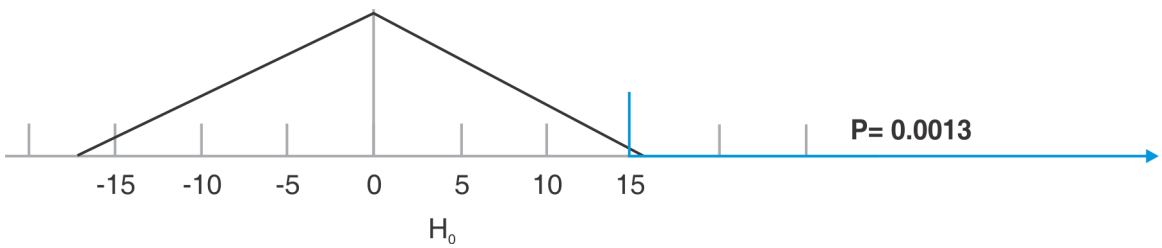
© Todos los derechos reservados

$$Z = \frac{|M_1 - M_2|}{E} = \frac{15}{5} = 3$$

La cantidad "E" del denominador es la estimación de la Desviación Estándar de la diferencia de medias muestrales si se repite millones de veces el estudio. Es decir, si las medias poblacionales son iguales y se repite millones de veces el estudio y en cada repetición se calcula la diferencia $M_{MUJ}–M_{VAR}$, en la mayoría de las repeticiones la M de mujeres será próxima a la de varones, $M_{MUJ}–M_{VAR}$ será próxima a cero. La media de esas diferencias será cero y la desviación estándar entre ellas estará en torno a 5. Por ello Z = 15/5 = 3 nos dice que la diferencia obtenida en nuestra muestra se aleja 3 desviaciones estándar de la media de las diferencias.

En las tablas de DN vemos que P_{UNI} = P (Z > 3) = 0.0013. Es decir, P ($M_{MUJ}–M_{VAR}$ >15|μ_{MUJ} = μ_{VAR}) = 0.0013. Solamente en 13 de cada diez mil repeticiones la M de mujeres aventajará, por azar, a la de varones en 15 o más unidades.

Distribución de $M_{MUJ}–M_{VAR}$ si es $\mu_{MUJ} = \mu_{VAR}$ y se repite el estudio millones de veces:



Y en otros trece estudios de cada diez mil, la media de mujeres estará 15 o más unidades por debajo de la de varones. Por tanto, la diferencia de medias muestrales será 15 o más, a favor de alguno de los grupos, en 26 estudios cada diez mil.

A esa cantidad la llamamos **Probabilidad Bilateral** y escribimos P_{BIL} = 0.0026.

Para calcular la P_{BIL} se tiene en cuenta la "otra cola", es decir, la situación simétrica a la obtenida en nuestro experimento respecto de la H_0 planteada. Si en nuestro ejemplo se obtuvo diferencia $M_{MUJ}–M_{VAR}$ = 15 (ventaja de 15 a favor de las mujeres), respecto de una ventaja poblacional de 0 (lo planteado en la H_0), la situación simétrica es $M_{MUJ}– M_{VAR}$ = –15 (ventaja de 15 a favor de los varones).

Tan pequeño valor P es evidencia notable en contra de la H_0 y a favor de que la media poblacional de Rusia es mayor en mujeres que en varones.

Los siguientes ejemplos ilustran que, en general, se obtiene menor valor P del test cuanto:

a) Mayor es la diferencia de las medias muestrales.

b) Menor es la variabilidad de la variable, menores desviaciones muestrales.

© Todos los derechos reservados

c) Mayor es el tamaño de las muestras.

Ejemplo 1

	N	M	S	ES
Mujeres	49	**79**	10.5	1.5
Varones	64	**73**	5.6	0.7

$M_1 - M_2$	E	Z	P_{UNIL}	$IC_{95\%}(M_1 - M_2)$
6	1.65	3.64	0.0001	2.77 y 9.23

Ejemplo 2

Mujeres	100	**79**	10.5	1.05
Varones	144	**73**	5.6	0.47

$M_1 - M_2$	E	Z	P_{UNIL}	$IC_{95\%}(M_1 - M_2)$
6	1.13	5.28	0.000002	3.8 y 8.2

Ejemplo 3

Mujeres	49	**79**	10.5	1.5
Varones	64	**77**	5.6	0.7

$M_1 - M_2$	E	Z	P_{UNIL}	$IC_{95\%}(M_1 - M_2)$
2	1.65	1.22	0.11	–1.23 y 5.23

Ejemplo 4

Mujeres	49	**79**	3.5	0.5
Varones	64	**77**	2.4	0.3

$M_1 - M_2$	E	Z	P_{UNIL}	$IC_{95\%}(M_1 - M_2)$
2	0.58	3.43	0.0003	0.86 y 3.14

Ejemplo 5

Mujeres	49	**79**	27	4
Varones	64	**77**	24	3

$M_1 - M_2$	E	Z	P_{UNIL}	$IC_{95\%}(M_1 - M_2)$
6	5	1.2	0.12	–3.8 y 15.8

Ejemplo 1 vs Ejemplo 2: mayor evidencia contra la hipótesis de igualdad de medias poblacionales e IC más estrecho en "2", pues se estudian muestras más grandes, aunque con igual diferencia de medias muestrales e igual desviación.

Ejemplo 3 vs Ejemplo 4: menor valor P e IC más estrecho en "4", pues con tamaños de muestra y diferencia de medias muestrales igual que en el "3", en el "4" la dispersión es menor en las muestras. Lo mismo ocurre en "1" y "5".

Ejemplo 1 vs Ejemplo 3: menor valor P en "1" pues hay mayor diferencia de medias muestrales, siendo muestras de igual tamaño y con la misma dispersión.

También se puede plantear como H_0 que la diferencia de medias poblacionales tiene cierto valor distinto de cero. Entonces se calcula:

$$Z = \frac{\left|(\mu_1 - \mu_2) - (M_1 - M_2)\right|}{E}$$

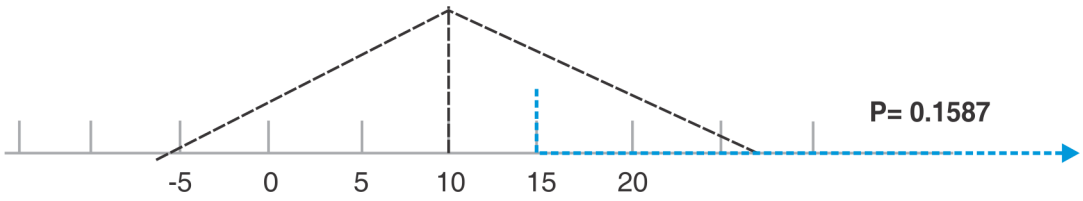
© Todos los derechos reservados

La fórmula para el test de igualdad de medias poblacionales es un caso particular de esta, con $(\mu 1-\mu 2) = 0$. El numerador es la "distancia" entre la diferencia de medias poblacionales que propone la Hipótesis $(\mu 1-\mu 2)$ y la diferencia de medias muestrales obtenida en el estudio (M_1-M_2).

Por ejemplo, si un especialista cree que la μ_{MUJ} aventaja a la de varones en 10 unidades, para la H_0: $\mu_{MUJ}-\mu_{VAR} = 10$ y los datos obtenidos en el estudio de Rusia, es:

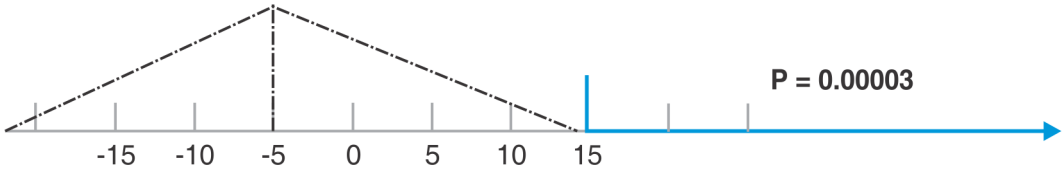
$Z = (15-10)/5 = 5/5 = 1$ y en la DN es $P(Z > 1) = 0.1587$.

Si es $\mu_{MUJ}-\mu_{VAR} = 10$ y se repitiese este estudio millones de veces, en la mayoría de ellas $M_{MUJ}-M_{VAR}$ tomará valores próximos a 10, pero en más de 15 cada cien repeticiones la M de mujeres aventajará a la de varones en 15 o más unidades.



Si se plantea como H_0 que la μ_{MUJ} es 5 unidades menor que la de varones, H_0: $\mu_{MUJ} - \mu_{VAR} = -5$ y $Z = (15-(-5))/5 = 20/5 = 4$ y en la DN es $P(Z > 4) = 0.00003$.

Si es $\mu_{MUJ}-\mu_{VAR} = -5$ y se repitiese este estudio millones de veces, en la mayoría $M_{MUJ} - M_{VAR}$ tomará valores próximos a -5, y solo en 3 de cada cien mil repeticiones la M de mujeres aventajará, por azar, a la de varones en 15 o más unidades.

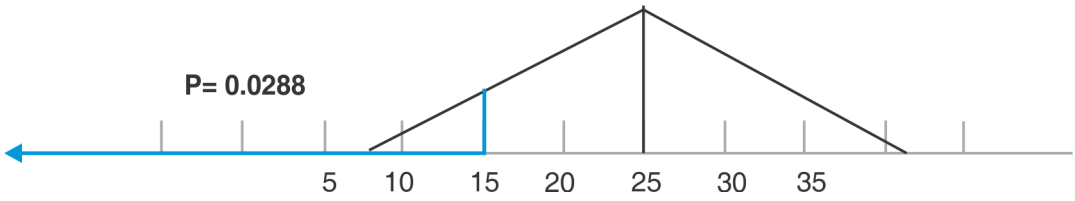


Si se plantea como H_0 que la media poblacional de mujeres es 25 unidades mayor que la de varones, para H_0: $\mu_{MUJ}-\mu_{VAR} = 25$, es:

$Z = (15-25)/5 = 10/5 = 2$ y en la DN es $P(Z > 2) = 0.0228$.

Si es $\mu_{MUJ}-\mu_{VAR} = 25$ y se repitiese este estudio millones de veces, en la mayoría de ellas $M_{MUJ}-M_{VAR}$ tomará valores próximos a 25, y solamente en 2 o 3 de cada cien repeticiones la M de mujeres aventajará a la de varones en 15 o menos unidades.

© Todos los derechos reservados



Fíjese que ahora interpretamos el valor P_{UNI} como la FR de estudios con diferencia de medias muestrales menor que 15, y en los ejemplos anteriores hablábamos de diferencia de medias muestrales mayor de 15, siendo 15 la diferencia obtenida en nuestros datos. ¿Por qué unas veces hablamos de diferencia mayor que la obtenida y otras veces menor? Depende de la H_0 planteada. El valor P se refiere a la proporción de estudios con resultado como el nuestro o **más "extremo"** si H_0 es cierta. En los ejemplos anteriores con H_0: $\mu_1-\mu_2 = 0$ y H_0: $\mu_1-\mu_2 = 10$ y H_0: $\mu_1-\mu_2 = -5$, resultados más extremos son aquellos con $M_1-M_2 > 15$, mientras que con H_0: $\mu_1-\mu_2 = 25$, resultados más extremos son los que tienen $M_1-M_2 < 15$.

Si alguna muestra tiene N menor de 30, el valor P del test se busca en las tablas t de Student (con grados de libertad "N_1+N_2-2"), en vez de las de DN.

© Todos los derechos reservados

9.3. IC PARA DIFERENCIA DE DOS MEDIAS CON DATOS APAREADOS

En algunas ocasiones se trata de estudiar una misma variable en un mismo grupo de individuos en dos situaciones diferentes, por ejemplo, antes y después de hacer ejercicio, o de seguir un determinado tratamiento; y se pretende saber si el ejercicio o el tratamiento modifican la media de la población. En este caso hablamos de *datos* apareados. Hay clara diferencia entre las situaciones anteriores de este capítulo y la que ahora abordamos. En ambos casos comparamos dos medias muestrales, pero en las situaciones anteriores las dos medias muestrales están calculadas sobre *dos grupos* de individuos diferentes, y en este caso las medias muestrales están calculadas sobre un *único grupo* de individuos, a cada uno de los cuales se le mide la misma variable en dos ocasiones.

Por ejemplo, para ver si tomar medio litro de cerveza aumenta el tiempo de respuesta, X, se mide lo que cada sujeto tarda en responder a cierto estímulo en condiciones estándar (X_{Antes}) y tras beber cerveza ($X_{Después}$). Si para cada individuo se calcula la diferencia $D = X_{Después}-X_{Antes}$, la media de las D es igual a $M_{Después}-M_{Antes}$ (la media de las diferencias es igual a la diferencia de las medias). En una muestra de 5 individuos se obtiene:

Antes:	13	14	11	5	7	→	M_{Antes}=	**10**	SC_{Antes}=	60
Después:	15	19	17	9	10	→	$M_{Después}$=	**14**	$SC_{Después}$=	67
Diferencias **D**:	2	5	6	4	3	→	$M_{Diferencias}$=	**4**		
	SC(D)= 10						ES(D)= 0.71			

Tenemos una sola muestra y en cada individuo se valora el aumento de X tras ingerir cerveza, de forma que el primero aumentó su tiempo de respuesta en 2, el segundo en 5,..., y se calcula la SC y el ES de estas diferencias.

$$\text{Intervalo de confianza al 95\%} \quad IC(\mu_{Diferencia}) \equiv M_{Diferencia} \pm Z \cdot ES(D)$$

Donde Z se mira en la tabla t de Student (en la fila con N–1 grados de libertad y en la columna que tiene como cabecera 5% bilateral) si es N < 30; y en la de la DN si es N = 30. Con 4 gl es Z = 2.776.

$IC_{95\%}(\mu_D) = 4 \pm 2.776 \cdot 0.71 = 4 \pm 1.97 =$ **2.03** y **5.97**

El aumento medio es 4 unidades en la muestra y tenemos confianza 95% en que beber medio litro de cerveza aumenta la media poblacional entre 2.03 y 5.97.

9.4. TESTS PARA DIFERENCIA DE MEDIAS CON DATOS APAREADOS

9.4.1. TEST PARA HIPÓTESIS DE "NO EFECTO POBLACIONAL" → H_0: $\mu D = 0$ Ó $\mu_{DESPUÉS} = \mu_{ANTES}$

Hipótesis Nula H_0: medio litro de cerveza **No** incrementa el tiempo de respuesta.

$$Z = \frac{|M_{Diferencia}|}{ES(D)}$$

En el ejemplo: $M_{Diferencia} = 4$ y ES (D) = 0.71 → Z = 4/0.71 = 5.66.

En las tablas t de Student (dado que N < 30) con 4 gl da $P_{UNIL} < 0.005$.

Si la cerveza no aumenta X y se hiciera este estudio millones de veces, la M de las diferencias tomaría valores en torno a 0, y solamente en menos de 5 cada mil estudios sería 4 (lo encontrado en las muestras) o más.

9.4.2. TEST QUE PROPONE CIERTO EFECTO POBLACIONAL → H_0: $\mu_D = \mu_{DESPUÉS} - \mu_{ANTES} = \Delta$

$$\text{Para } H_0: \mu_D = \mu_{Después} - \mu_{Antes} = \Delta \text{ se calcula: } Z = \frac{|\Delta - M_{Diferencia}|}{ES(D)}$$

© Todos los derechos reservados

Con Δ el aumento que propone H_0, y $M_{Diferencia}$ y ES (D) obtenidos en la muestra.

Por ejemplo, estudios previos muestran que medio litro de vino incrementa la media poblacional de X en 11 unidades y se quiere ver si los datos muestrales son compatibles con que la cerveza haga igual efecto. Plateamos la H_0 que dice que medio litro de cerveza incrementa la media poblacional en 11.

$\rightarrow H_0$: $\mu_{Diferencia}$ =11 $\rightarrow$ Z = (11–4)/0.71 = 9.8 $\rightarrow$ **P_{UNIL} < 0.0005** (tablas t con 4 gl).

$\rightarrow$ Se rechaza H_0 $\rightarrow$ Concluimos que $\mu_{Después}-\mu_{Antes}$ es < 11.

Si la cerveza aumenta la media poblacional de X en 11 y se hicieran millones de estudios como este, la media muestral de las diferencias tomará valores en torno a 11 y en menos de 5 cada diez mil estudios la media muestral de las diferencias sería 4 o menor. La misma proporción de veces (menos de 5 cada 10000 estudios) la media muestral de las diferencias será 18 o mayor yo que un aumento de 18 unidades se distancia del aumento propuesto en la H_0 (11 unidades) lo mismo que 4, pero en el otro sentido.

Cuando los datos son "apareados" deben ser analizados como tal, pues hacerlo como "independientes" puede conducir a conclusiones erróneas. Como ejemplo haremos el test para la H_0 que dice que $\mu_{Antes} = \mu_{Después}$ tratándolo como si fueran dos medias de dos grupos.

Sería E = $(60/20+76/20)^{½}$ = 2.6 y por tanto t_8 = 4/2.6 = 1.53 $\rightarrow$ P > 0.05, que no invita a rechazar H_0, es decir, no es evidencia suficiente a favor de que la ingesta de cerveza alarga el tiempo de respuesta.

© Todos los derechos reservados

9.5. ASUNCIONES DEL TEST "T" PARA COMPARAR DOS MEDIAS

Llamaremos "test-t" al procedimiento descrito en los apartados anteriores, "**P-t**" al valor P del test que da ese procedimiento y "**P-verdadero**" a la verdadera probabilidad de obtener por azar (cuando son iguales las medias poblacionales) medias muestrales tan distintas como las de nuestro estudio o más distintas. Este valor mide el grado de evidencia en contra de la hipótesis nula. En teoría, con muestras pequeñas, digamos menores de 30, el P–t es igual al P–verdadero si se dan dos circunstancias:

1ª- Normalidad: la variable estudiada debe tener DN en las dos poblaciones implicadas. Aquí se aplican todas las observaciones del apartado 9.8 y siguientes.

2ª- Homogeneidad de Varianzas: la variable en estudio tiene la misma varianza en las dos poblaciones muestreadas. En este sentido hay que tener en cuenta los siguientes matices:

1. Existen tests –llamados de Homogeneidad de Varianzas– que plantean como H_0 que las varianzas de las dos poblaciones son iguales, pero nunca pueden confirmar que esa H_0 es cierta, solo pueden, si el valor P obtenido es muy pequeño, confirmar que es falsa.

2. La diferencia entre el "**P-t**" y el "**P-verdadero**" no es muy grande aunque las dos varianzas poblacionales sean notablemente diferentes. A esto se le llama "**Robustez**" frente a la no homogeneidad de varianzas.

Por ello en la mayoría de los casos puede aplicarse el test–t con la confianza de que el valor P encontrado con esta técnica se aproxima lo suficiente al P–verdadero como para hacer válidas las conclusiones alcanzadas.

9.6. TESTS NO PARAMÉTRICOS

En las pocas ocasiones en que hay evidencia clara contra la Homogeneidad de Varianzas o contra la Normalidad, pueden aplicarse los llamados **Tests No Paramétricos**, que no requieren esas condiciones. Suelen ser un poco menos potentes, es decir, cuando H_0 no es cierta lo detectan menos veces que el test-t. Y suelen dar valores P próximos a los que da el test–t y en cualquier caso de rango similar. El cálculo del valor P de los tests no paramétricos puede ser tedioso, pero los ordenadores lo hacen en milisegundos. Aquí no vamos a explicar esos cálculos, sino a comparar el valor P que se obtiene con ellos con el obtenido con el test–t.

En la tabla que sigue se muestran tres ejemplos con 7 valores en cada una de las dos muestras. Se dan las medias y varianzas en cada una de las muestras, los valores P del test de Normalidad, del test de Homogeneidad de Varianzas, del test-t y del no paramétrico.

Ejemplo 1º							MED	S^2	P Nor.	P Hom-var.	P-t	P No - Para.
2	4	7	8	9	12	14	**8**	17.6	0.95	0.08	.060	.075
1	4	4	6	5	8	7	**5**	5.28	0.85			

Ejemplo 2º							MED	S^2	P Nor.	P Hom-var.	P-t	P No - Para.
17	4	7	8	9	12	14	**10.1**	19.8	0.95	0.06	.009	.010
1	4	4	6	5	8	7	**5**	5.28	0.85			

Ejemplo 3º							MED	S^2	P Nor.	P Hom-var.	P-t	P No - Para.
17	20	7	8	9	12	14	**12.4**	23.6	0.63	0.045	.0017	.0015
							5	5.28	0.85			

Con el test de Normalidad se obtiene en todos los ejemplos valores altos indicando que no hay evidencia en contra de que la variable tenga DN en las respectivas poblaciones. Los tests que plantean varianzas iguales en las dos poblaciones dan valores P en torno a 5%, que podría hacernos dudar de que haya igualdad de varianzas poblacionales, pero no es evidencia muy fuerte contra esa hipótesis. Ante la duda de que haya homogeneidad de varianzas hacemos el test no paramétrico Los valores P–t no son muy diferentes de los respectivos valores P del no paramétrico. Así, para el primer ejemplo el test–t da P = 0.06 y el test no-paramétrico da P =

© Todos los derechos reservados

0.075, por lo que ambos aportan solamente una modesta evidencia contra la hipótesis de igualdad de medias poblacionales. Esta proximidad entre los valores P del test–t y del no–paramétrico se da también en los otros dos ejemplos. No siempre ambos valores P son tan próximos, pero en pocas ocasiones la divergencia entre ellos es relevante.

9.7. ASPECTOS A RECORDAR

- Hablamos de comparación de "medias independientes" cuando se mide una variable en dos muestras. A partir del valor de las dos medias muestrales haremos Inferencia acerca de las medias poblacionales, calculando el Intervalo de Confianza para la diferencia de medias poblacionales y haciendo Tests.
- Si una o ambas muestras son "pequeñas", digamos menores de 30, para el cálculo de los IC y del valor P de los tests se usa la distribución t de Student.
- Si se plantea como H_0: $\mu_1 = \mu_2$, es decir, $\mu_1 - \mu_2 = 0$, el valor P dice la proporción de veces que se encontraría diferencia de medias muestrales como la de nuestro estudio o mayor, si fueran iguales las medias poblacionales y el estudio se repitiera millones de veces.
- Si se plantea como H_0 que es $\mu_1 - \mu_2 = K$, el valor P dice la proporción de veces que se encontraría diferencia de medias muestrales tan alejada de K como la de nuestro estudio o aun más alejada, si es $\mu_1 - \mu_2 = K$.
- Hablamos de comparación de "medias apareadas" cuando en cada individuo de una muestra se mide la misma variable dos veces, por ejemplo, antes y después de aplicar un tratamiento. El análisis se hace calculando la diferencia en cada individuo, la media muestral de esas diferencias, M_{DIF}, y su error estándar, ES_{DIF}. La media de las diferencias es igual a la diferencia de las medias de los dos grupos.
- Típicamente se plantea como H_0 que el tratamiento no modifica realmente la variable estudiada, es decir, que μ_{DIF} es cero en población (la media poblacional de esa variable es igual antes y después del tratamiento). El valor P del test dice la proporción de veces que se encontraría diferencia de medias muestrales como la de nuestro estudio o mayor, si fueran iguales las medias poblacionales y el estudio se repitiera millones de veces.
- Si se plantea como H_0 que el tratamiento aumenta la media poblacional de esa variable en K unidades, el valor P dice la proporción de veces que se encontraría MDIF como la de nuestro estudio o más alejada de K, si fuera cierta H_0 y el estudio se repitiera millones de veces.
- Aun en el caso de que la variable estudiada tenga en la población distribución moderadamente alejada de la Normal y/o de que las varianzas sean moderadamente diferentes en las poblaciones, el valor P obtenido con este test indica con buena aproximación lo que de él se espera. Se dice que este test es robusto frente a la no normalidad y la no homogeneidad de varianzas.

© Todos los derechos reservados

© Todos los derechos reservados

Resumen

© Todos los derechos reservados

© Todos los derechos reservados

La inferencia estadística con dos proporciones

10

Visión general

- Definición del intervalo de confianza para la diferencia de proporciones poblacionales.
- Definición del y test de significancia para la diferencia de proporciones poblacionales.
- Definición de Test para la igualdad de proporciones poblacionales.
- Comparación de dos proporciones con datos pareados e independientes.

Objetivo

- Conocer la inferencia estadística con dos proporciones.

© Todos los derechos reservados

10.1. INTERVALO DE CONFIANZA PARA π_2 - π_1

Hacia 1875 Joseph Lister publicaba en Lancet datos mostrando que cuando las amputaciones de miembros se hacían por cirujanos que se habían lavado previamente las manos esmerada y meticulosamente, aumentaba la proporción de pacientes que sobrevivían. Y se lamentaba de no poder saber si la diferencia ocurrida en su muestra podría haber ocurrido por azar, sin que esas abluciones tuvieran realmente efecto beneficioso. Sus datos eran aproximadamente estos:

- Entre 40 operados sin abluciones curaron 10: $p_1 = 10/40 = 0.25$.
- Entre 30 operados con abluciones curaron 12: $p_2 = 12/30 = 0.40$.

Hoy día la Estadística tiene herramientas que ayudan a formarse opinión en ese sentido, pero no siempre permiten llegar a una conclusión definitiva. Veremos cómo se calculan e interpretan los Intervalos de Confianza para la diferencia de proporciones poblacionales y el valor P del test que plantea como H_0 que las dos proporciones poblacionales son iguales. Para este ejemplo la H_0 dice que los lavados en cuestión no son efectivos, es decir que la proporción poblacional de curaciones es igual con y sin abluciones.

Usaremos esta notación para indicar cada una de las cantidades implicadas, llamando "tratamiento 1" al hecho de no lavarse y "2" al hecho de lavarse:

	Curan	**No curan**	Total		
Trat. 1	$X_1 = 10$	$\mathbf{R_1} = 30$	$N_1 = 40$	$p_1 = X_1/N_1 = \mathbf{0.25}$	$q_1 = R_1/N_1 = 0.75$
Trat. 2	$X_2 = 12$	$\mathbf{R_2} = 18$	$N_2 = 30$	$p_2 = X_2/N_2 = \mathbf{0.40}$	$q_2 = R_2/N_2 = 0.60$
Total	$X = 22$	$\mathbf{R} = 48$	$N = 70$	$p = X/N = \mathbf{0.3143}$	$q = R/N = 0.6856$

El intervalo de confianza para la diferencia de proporciones poblacionales se calcula con la siguiente fórmula, donde Z depende de la confianza que queramos tener en que esa diferencia se encuentre dentro del intervalo. De la misma forma que para el IC de una proporción, Z se obtiene de las tablas de distribución Normal, siendo Z = 1.96 para confianza 95% y Z = 2.58 para confianza 99%.

Intervalo de Confianza

$$IC(\pi_2 - \pi_1) \equiv (p_2 - p_1) \pm Z\sqrt{p_1 \cdot q_1 / N_1 + p_2 \cdot q_2 / N_2}$$

En este ejemplo es IC95% $(\pi_2 - \pi_1) \equiv 0.15 \pm 1.96\ (0.25 \cdot 0.75/40 + 0.40 \cdot 0.60/30)^{1/2} = 0.15 \pm 0.22 = -$ **0.07** y **0.37**.

En las muestras, la FR de supervivencia con "2" es 15 (0.40–0.25 = 0.15) puntos mayor que con "1"; en la población no sabemos cuánto es esa diferencia, pero tenemos confianza 95% en que la FR poblacional de supervivencia con "2" sea desde 37 puntos mayor que con "1", hasta 7 puntos menor (el signo negativo indica ventaja a favor del "1").

10.2. TEST PARA LA IGUALDAD DE DOS PROPORCIONES POBLACIONALES

La H_0 es que "1" y "2" son igual de eficaces, es decir, $\pi_1 = \pi_2 \Leftrightarrow \pi_1 - \pi_2 = 0$.

En ese caso si se hicieran millones de estudios como este, la diferencia $(p_2 - p_1)$ tomaría valores en torno a 0 en la mayoría de los estudios. La media de esa población de valores $(p_2 - p_1)$ sería cero y los matemáticos demuestran que la σ entre esos valores se estima por $\sigma\ (p_2 - p_1) \approx [p\ q\ (1/N_1 + 1/N_2)]^{1/2}$

Se trata de ver cuántas desviaciones se aleja el valor encontrado en nuestro estudio de la media, es decir, de cero: $Z = (p_2 - p_1)/\sigma\ (p_2 - p_1)$.

© Todos los derechos reservados

Es decir, $Z = 0.15/[0.3143 \cdot 0.6856 \cdot (1/40+1/30)]^{1/2} = 0.15/0.11 =$ **1.36**.

En las tabla de DN vemos que $P(Z \geq 1.36) \approx$ **0.09**.

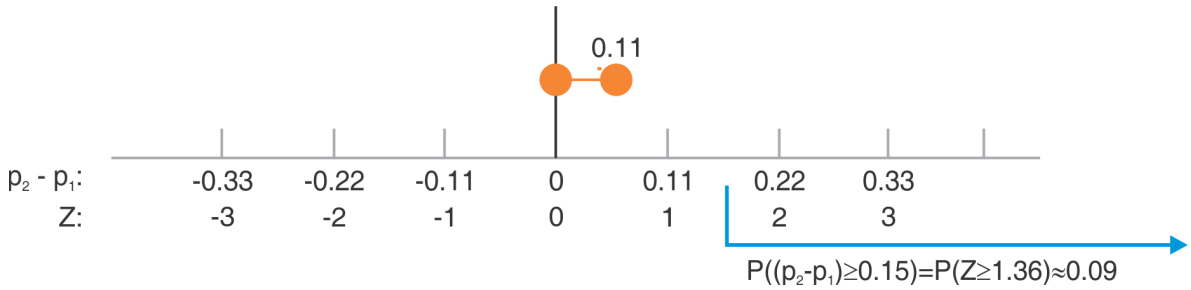
Es decir $P[(p_2-p_1) \geq 0.15 _ \pi_1-\pi_2 = 0] = P(Z \geq 1.36) =$ **0.09**.

Como en todos los tests de significación, el valor P indica la probabilidad de obtener un resultado como el de nuestro estudio o aún más alejado de lo esperado bajo la H_0, si dicha hipótesis es cierta. Si los lavados son inútiles, en 9 de cada 100 estudios, saldría $(p_2-p_1) \geq 0.15$. En otros 9 estudios de cada 100 encontraríamos que p_2-p es -0.15 o menos. Es decir, en 18 de cada 100 estudios las diferencias de FR muestrales serían 0.15 o más a favor de uno u otro: $P_{Bilateral}$ = **0.18**.

Hemos calculado el valor de Z con la primera de estas dos expresiones:

$$Z = \frac{|p_2 - p_1|}{\sqrt{p \cdot q \cdot \left(1/N_1 + 1/N_2 \right)}} = \sqrt{\frac{(X_1 \cdot R_2 - X_2 \cdot R_1)^2 \cdot N}{N_1 \cdot N_2 \cdot X \cdot R}}$$

La segunda da exactamente el mismo resultado y puede resultar más cómoda porque usa frecuencias absolutas.



En los ejemplos siguientes observe cómo el valor P y el IC varían en función de la diferencia muestral p_2-p_1 y del tamaño de las dos muestras. Es necesario distinguir claramente entre **magnitud del efecto** encontrado en la muestra (diferencia entre las proporciones muestrales) y **significación estadística** de ese efecto (valor P del test; diremos que un resultado es muy significativo si el valor de P es muy pequeño y es poco significativo si el valor de P es grande).

Ejemplo 10.1

Si los datos son ("+" son curados y "–" no curados):

© Todos los derechos reservados

	+	–	Total			$H_0: \pi_1 = \pi_2$
Trat. 1	3	9	12	$p_1 = 0.25$		**$P_{BIL} = 0.10$**
Trat. 2	6	4	10	$p_2 = 0.60$	$p_2 - p_1 = 0.35$	
Total	9	13	22	$p = .409$		$IC_{95\%} = -0.04$ y 0.74

Efecto grande en la muestra: "2" es bastante mejor que "1", **$p_2 - p_1 = 0.35$**.

Pequeña significación estadística, valor de P grande, no hay evidencia contra la hipótesis que propone igualdad de proporciones poblacionales, es decir, el resultado no demuestra que "2" sea realmente mejor que "1".

Ejemplo 10.2

Si los datos son:

	+	–	Total			$H_0: \pi_1 = \pi_2$
Trat. 1	12	36	48	$p_1 = 0.25$		**$P_{BIL} = 0.0009$**
Trat. 2	24	16	40	$p_2 = 0.60$	$p_2 - p_1 = 0.35$	
Total	36	52	88	$p = .409$		$IC_{95\%} = 0.15$ y 0.55

Igual efecto que en el anterior, $p_2 - p_1 = 0.35$, y tamaños de muestra mayores.

Esto hace que se obtenga menor valor P (estadísticamente muy significativo), gran evidencia contra la hipótesis de igualdad de proporciones poblacionales, es decir, el resultado demuestra que "2" es realmente mejor que "1". El IC nos dice que el tratamiento "2" cura entre 15 y 55 puntos más que el "1". Además se observa que el IC no contiene el 0 (proporciones poblacionales iguales).

Ejemplo 10.3

Si los datos son:

	+	–	Total			$H_0: \pi_1 = \pi_2$
Trat. 1	1000	3000	4000	$p_1 = 0.25$		**$P_{BIL} = 0.000002$**
Trat. 2	2900	7100	10000	$p_2 = 0.29$	$p_2 - p_1 = 0.04$	
Total	3900	10100	14000	$p = .2786$		$IC_{95\%} = 0.024$ y 0.056

© Todos los derechos reservados

Efecto pequeño, en la muestra el tratamiento "2" es solamente un poco mejor que el "1", $p_2 - p_1 = 0.04$. Significación estadística grande, valor P muy pequeño, clarísima evidencia contra la igualdad de proporciones poblacionales; se rechaza la H_0 concluyendo que "2" cura más que "1". Sin embargo, la ventaja del "2" sobre el "1" es modesta: como se aprecia en el IC, "2" cura entre 2.4 y 5.6 puntos más que el "1".

El procedimiento que acabamos de ver es una aproximación que resulta razonablemente buena cuando los "**valores Esperados**" bajo la hipótesis de igualdad de FR poblacionales no son inferiores a 4. Los valores Esperados son los que tendría que haber en cada celda para que con los mismos tamaños de muestras las FR de curaciones fueran igual en ambos tratamientos. Se calculan multiplicando, para cada una de las 4 celdas de la tabla 2 x 2, el total de la fila por el total de la columna y dividiendo por el total de la tabla, como se muestra a continuación:

Datos	**+**	–	
Trat. 1	3	9	12
Trat. 2	6	4	10
	9	13	22

→

Valores Esperados	**+**	–	
Trat. 1	9·12/22 = **4.9**	13·12/22 = **7.1**	12
Trat. 2	9·10/22 = **4.1**	13·10/22 = **5.9**	10
	9	13	22

Cuanto menores son los valores Esperados más puede alejarse el valor P obtenido por este procedimiento del verdadero. Cuando algunos Esperados son menores de 5 ese alejamiento puede ser considerable. En esos casos el valor P verdadero se calcula con el llamado método Exacto de Fisher, que es más laborioso, pero no supone problema para el ordenador. En el ejemplo de esta última tabla, el valor P obtenido arriba fue P_{BIL}= **0.10** y el método Exacto de Fisher da P_{BIL} = **0.19**. Vemos que la concordancia no es muy buena, pero la P aproximada nos lleva, como la exacta, a no rechazar la hipótesis de igualdad.

10.3. COMPARANDO DOS PROPORCIONES CON DATOS APAREADOS

Hemos visto cómo comparar la FR de individuos con una característica en dos grupos, por ejemplo, FR de enfermos con "E" entre los jóvenes y entre los viejos. Ahora estudiaremos cómo comparar la FR de dos características en un solo grupo de individuos. Así también ocurría en el análisis de "datos apareados" con variables cuantitativas, cuando estudiábamos medias. Como ejemplo compararemos la FR de personas con sordera frente a la FR de ellas con vértigo en una muestra de 80 que sufrió traumatismo de oído. Para ello tenemos que evaluar dos variables dicotómicas - la presencia de sordera y de vértigo - en cada individuo.

Los datos son: de 80 jóvenes, 16 tenían ambas enfermedades (S+ y V+), 4 tenían solo vértigo (S– y V+), 54 solo sordera (S+ y V–), y 6 ninguna (S– y V–).

© Todos los derechos reservados

	S+	S–	
V+	16	4	20
V–	54	6	60
	70	10	80

p_S = 70/80 = **0.875** p_V = 20/80 = **0.25**

$p_S > p_V$: $p_S - p_V$ = 0.875 – 0.25 = **0.625**

¿Hay diferencia en la población y de qué orden es?

Por tanto, 70 tenían sordera (con o sin vértigo), 87.5% del total y 20 tenían vértigo (con o sin sordera), 25% del total. La sordera es 62.5 puntos (0.875–0.25) más frecuente que el vértigo en esta muestra de 80 individuos.

Ahora no se trata de evaluar la FR de una característica en dos grupos, sino la FR de dos características en un solo grupo. Los cálculos para el IC y para el test son distintos, pero la interpretación es equivalente.

10.3.1. Intervalo de confianza para (π_S - π_V) con datos apareados

Con la notación de la tabla siguiente, la fórmula para calcular el IC es:

$$IC \equiv (p_S - p_V) \pm Z\frac{\sqrt{b+c}}{N}$$

	S+	S–
V+	a	b
V–	c	d

$IC_{95\%}$ (π_S–π_V) ≡ (0,875–0,25) ± 1,96 $(58)^{1/2}$/80 = 0,625±0,186 = 0,44 y 0,81.

En la muestra estudiada la sordera es 62.5 puntos más frecuente que el vértigo.

En la población no sabemos cuánto es esa diferencia, pero tenemos confianza del 95% en que la sordera es, en población, entre 44 y 81 puntos más frecuente que el vértigo.

10.3.2. Tests para hipótesis para $\pi_S = \pi_V$ con datos apareados

Para la H_0: Ambas enfermedades son igual de frecuentes en población: $\pi_S = \pi_V$.

$$Z = \frac{|b-c|-1}{\sqrt{b+c}} \qquad Z = \frac{|54-4|-1}{\sqrt{54+4}} = 6.43 \qquad P_{UNI} < 10^{-11}$$

© Todos los derechos reservados

En la muestra la sordera es 62.5 puntos más frecuente que el vértigo ($p_S - p_V = 0.625$), ¿puede ser que en la población de la que procede esta muestra no haya diferencia entre la frecuencia de ambas enfermedades? El valor P del test se obtiene mirando la cantidad Z encontrada en las tablas de la DN. Si ambas enfermedades son igual de frecuentes en población, es decir $\pi_S - \pi_V = 0$, y repetimos este estudio millones de veces, la mayoría de ellas será $p_S - p_V$ próximo a 0, y menos de 1 veces cada cien mil millones será, por azar, $(p_S - p_V) \geq 0.625$ (que es la diferencia encontrada en nuestro estudio). La evidencia contra la H_0 es muy fuerte. La rechazamos y concluimos que la sordera es, en población, más frecuente que el vértigo.

Si se calcula el valor P para igualdad de FR poblacionales considerando, *erróneamente*, las dos proporciones muestrales como datos independientes se obtiene $P_{UNI} < 10-15$. En este ejemplo la conclusión sería igual a la obtenida con el cálculo correcto, pero en otros puede llevar a conclusiones erróneas.

10.4. DOS PROPORCIONES: DATOS APAREADOS VS. DATOS INDEPENDIENTES

© Todos los derechos reservados

En el ejemplo usado antes para comparar proporciones con datos apareados podemos también practicar la comparación de proporciones con datos independientes. El mismo conjunto de datos nos permite hacer Inferencia con proporciones independientes y apareadas.

1º	S+	S–		Datos Apareados	Datos Independientes
V+	16	4	20	FR (S) 70/80 = **0.875**	FR (S | V+) = 16/20 = **0.80**
V–	54	6	60	FR (V) = 20/80 = **0.25**	FR (S | V–) = 54/60 = **0.90**
	70	10	80	FR (S)–FR (V) = **0.625**	FR (S | V+) < FR (S | V–) FR (S | V–) – FR (S | V+) = **0.10**

Una cosa es estudiar cual de las dos enfermedades es más frecuente (datos apareados) y otra ver si las enfermedades están asociadas, es decir, si la presencia de una aumenta o disminuye la presencia de la otra (datos independientes). En esta muestra se ve:

a) La sordera afecta a más personas que el vértigo (el 87.5% tiene sordera frente al 25% que tiene vértigo). Se comparan dos proporciones con datos apareados; hay un solo grupo de 80 individuos. Para valorar si en la población de la que procede la muestra una enfermedad es más frecuente que la otra, se utiliza el test visto en el apartado 11.3.

b) La sordera es menos frecuente cuando hay vértigo que cuando no hay vértigo: p(S|V+) = 16/20= **80%** frente a p(S|V–) = 54/60 = **90%**, es decir, tienen sordera un 80% de los que tienen vértigo y un 90% de los que no lo tienen; esto es, la sordera no es igual de frecuente en presencia que en ausencia de vértigo. Por tanto, decimos que en la muestra estas dos enfermedades están asociadas, no son independientes, pues la presencia de una afecta a la presencia de la otra. Los denominadores de estos dos porcentajes son diferentes, pues se comparan dos grupos de individuos distintos (los que tienen vértigo que son 20, y los

que no tienen vértigo, que son 60). Para valorar si en las poblaciones de las que proceden las muestras *sordera* y vértigo son *independientes*, se utiliza el test visto en 11.2.

Los tres ejemplos que siguen ayudan a entender la diferencia entre esos dos conceptos. En los tres supuestos el vértigo es 30 puntos más frecuente que la sordera, pero puede verse que en el 1º la sordera es igual de frecuente en los V+ que en los V–, es decir, no hay asociación entre las dos enfermedades; en el 2º la sordera es más frecuente en los V+ que en los V– (sordera y vértigo no son independientes; hay asociación positiva); y en el 3º la sordera es menos frecuente en los V+ que en los V– (sordera y vértigo no son independientes; hay asociación negativa).

1º	S+	S–		**FR(V) vs. FR(S)**	**FR(S | V+) vs. FR (S | V-)**
V+	20	20	40	FR (V) = **0.8**	FR(S | V+) = **0.5**
V–	5	5	10	FR (S) = **0.5**	FR(S | V–) = **0.5**
	25	25	50	FR (V)–FR (S) = **0.3**	FR(S | V+) = FR (S | V–)

2º	S+	S–		**FR(V) vs. FR(S)**	**FR(S | V+) vs. FR (S | V-)**
V+	24	16	40	FR (V) = **0.8**	FR(S | V+) = **0.6**
V–	1	9	10	FR (S) = **0.5**	FR(S | V–) = **0.1**
	25	25	50	FR (V)–FR (S) = **0.3**	FR S | V+) > FR (S | V–) FR(S | V+) - FR (S | V–) = **0.5**

3º	S+	S–		**FR(V) vs. FR(S)**	**FR(S | V+) vs. FR (S | V-)**
V+	16	24	40	FR (V) = **0.8**	FR(S | V+) = **0.4**
V–	9	1	10	FR (S) = **0.5**	FR(S | V–) = **0.9**
	25	25	50	FR (V)–FR (S) = **0.3**	FR(S | V+) > FR (S | V–) FR(S | V+) – FR (S | V–) = **–0.5**

10.5. ASPECTOS A RECORDAR

- Hablamos de dos proporciones con muestras independientes si se mide la proporción de una característica en dos grupos distintos de individuos. Hablamos de dos proporciones con datos apareados si se mide la proporción de individuos con cada una de dos características en un mismo grupo.
- A partir de la diferencia de las proporciones muestrales se calcula un intervalo de confianza para la diferencia de las proporciones poblacionales. Se utilizan fórmulas diferentes dependiendo de si son muestras independientes o apareadas.
- En un estudio destinado a comparar la eficacia de dos fármacos, la Hipótesis Nula, H_0, dice que no hay tal ventaja. El valor P del test nos dice la proporción de estudios en los que se encontraría una diferencia de curaciones como la encontrada en nuestro estudio o aun mayor, si los dos fueran igual de eficaces en la población. Son datos independientes.

© Todos los derechos reservados

- En un estudio para ver si una enfermedad, o un síntoma (o cualquier otra característica dicotómica) es más frecuente que otra, la Hipótesis Nula dice que son igual de frecuentes. El valor P del test nos da la proporción de muestras en las que se obtendría una ventaja en la frecuencia de un carácter sobre el otro como el de nuestro estudio o aun mayor, si los dos caracteres fueran igual de frecuentes en población. Son datos apareados.
- En la primera tabla la enfermedad "S" es menos frecuente en los V+ (40%) que en los V– (90%). En la segunda "S" es más frecuente en los V+ (60%) que en los V– (10%). En ambas, la enfermedad "V" afecta al 80% y la "S" al 50%.

	S+	S–		**FR(V) vs. FR(S)**	**FR(S | V+) vs. FR (S | V-)**
V+	16	16	40	FR (V) = 40/50 = **0.8**	FR(S | V+) = 16/40 = **0.4**
V–	9	1	10	FR (S) = 25/50 = **0.5**	FR(S | V–) = 9/10 = **0.9**
	25	25	50		

	S+	S–		**FR(V) vs. FR(S)**	**FR(S | V+) vs. FR (S | V-)**
V+	24	16	40	FR (V) = **0.8**	FR(S | V+) = 24/40 = **0.6**
V–	1	9	10	FR (S) = **0.5**	FR(S | V–) = 1/10 = **0.1**
	25	25	50		

© Todos los derechos reservados

© Todos los derechos reservados

Resumen

© Todos los derechos reservados

© Todos los derechos reservados

Tamaño de muestra para estimación 11

Visión general

- Definición de tamaño de muestra para estimar una proporción.
- Definición de tamaño de muestra para estimar una media.
- Definición de tamaño de muestra para estimar la diferencia entre dos proporciones.
- Definición de tamaño de muestra para estimar la diferencia entre dos medias.

Objetivo

- Calcular el tamaño de muestra para una estimación.

© Todos los derechos reservados

11.1. UNA ANÉCDOTA PARA LLAMAR LA ATENCIÓN SOBRE UN ERROR

Si estamos interesados en conocer la media de cierta variable en una población, el único modo de calcularla exactamente es medir esa variable en todos los individuos de esa población. Pero en la práctica tomamos una muestra y calculamos la **media muestral**, que no tiene, en general, exactamente el mismo valor que la poblacional. Esperamos que tenga un valor "parecido", pero ¿cuán buena será la aproximación de la media muestral a la verdadera media poblacional? La respuesta es:

En general, cuanto mayor sea la muestra mejor será esa aproximación.

Pero el investigador intenta tomar una muestra lo suficientemente grande como para que la media muestral se aproxime mucho a la poblacional, y a la vez, lo más pequeña posible, para ahorrar tiempo y dinero.

Es creencia general que hay un tamaño idóneo que es el mínimo, y por ello el más barato, dentro de los suficientemente grandes como para dar una media muestral muy aproximada a la media

poblacional. Los profesionales que están diseñando una investigación se preguntan cuál es ese tamaño de muestra óptimo en el caso de su investigación. Pero la realidad es que en la mayoría de las situaciones *no hay tal tamaño idóneo* y la decisión sobre cuántos individuos incluir en nuestra investigación no depende de un cálculo estadístico.

Cuando se le explica al investigador que, en contra de su creencia, no hay un tamaño que sea el "adecuado" para la investigación que proyecta, suele responder que en los libros de estadística aparecen fórmulas para calcular esos tamaños y él cree que aplicando una de ellas aparecería el tamaño idóneo para su proyecto actual. Es cierto que existen esas fórmulas, pero en este capítulo veremos que pueden dar muchas respuestas distintas, todas ellas válidas, para el tamaño de la muestra buscado. Por ello en la decisión final del tamaño de muestra a usar en una investigación estas fórmulas tienen un peso muy reducido.

En este capítulo se comentarán las fórmulas para calcular el tamaño de muestra para la estimación de una media y una proporción poblacional, así como para la diferencia de dos proporciones y dos medias poblacionales. El tamaño de muestra en relación con tests estadísticos tiene mayor grado de dificultad, no matemática sino lógica.

La doctora Amelia Mendoza, jefa del Servicio de Neurología del Hospital de Macondo, decide hacer un estudio para estimar la FR de personas que consumen hipnóticos, CH, en la población constituida por todos los habitantes del Caribe. Para determinar el número de personas que debe encuestar visita a un bioestadístico, que tras comentar con ella el caso le sugiere entrevistar a 27.250 personas. Pero en el camino de vuelta coincide en el tren con otro bioestadístico, que tras comentar con ella el caso le sugiere entrevistar a 50 personas. Sospechando que había cometido algún error al dar los datos que le pedían los expertos o que alguno se equivocó en sus cálculos, vuelve a hablar con cada uno de ellos, pero, para su sorpresa, ambos se ratificaron en su cantidad.

Para resolver el conflicto convoca un comité de los mejores expertos europeos. Los convocados acuerdan por unanimidad que ambos bioestadísticos habían actuado correctamente, y añaden que 374 sería un tamaño adecuado.

Amelia no sale de su asombro, perplejidad y enojo, y decide convocar un comité de los mejores expertos mundiales. Tras escucharla, los nuevos convocados acuerdan que los tres tamaños propuestos anteriormente eran adecuados y añaden que 1 040 sería un tamaño adecuado. Finalmente Amelia renuncia al cálculo "científico" del "tamaño adecuado" de la muestra y decide actuar de acuerdo con los recursos humanos y económicos de que dispone. Encuesta a una muestra aleatoria de 1000 personas y encuentra que 200 de ellas son CH. Publica un trabajo diciendo que en la muestra estudiada fueron CH el 20% ($p = 0.20$) y el Intervalo de Confianza para el porcentaje poblacional, π, fue:

$$IC_{95\%}(\pi) = p \pm 1.96 \cdot [p(1-p) / N]^{1/2} = 17.5\% \text{ y } 22.5\%$$

Pero la autoridad sanitaria reprocha a la autora que el tamaño de la muestra fue decidido sin "criterios científicos" adecuados. ¿Está justificada esa crítica? A lo largo del capítulo veremos que no.

© Todos los derechos reservados

11.2. TAMAÑO DE MUESTRA PARA ESTIMAR UNA PROPORCIÓN

Es obvio que ningún tamaño muestral permite conocer exactamente el dato poblacional. A partir de la FR muestral, p, el investigador calcula un Intervalo de Confianza, IC, dentro del cual creemos que se encuentra la FR poblacional. En general, cuanto mayor sea la muestra más precisa será la estimación, es decir el IC para el valor poblacional, IC (π), será más estrecho. Veamos estos ejemplos:

Con N = **20** y **2** CHp = 10% $\rightarrow$ $IC_{95\%}$ (π) = 0.1% y 23% Ancho IC: **22.9**.

Con N = **400** y **40** CHp = 10% $\rightarrow$ $IC_{95\%}$ (π) = 7% y 13% Ancho IC: **6**.

Con N = **1600** y **160** CHp = 10% $\rightarrow$ $IC_{95\%}$ (π) = 8.5% y 11.5% Ancho IC: **3**.

No hay un tamaño "adecuado" sino que el IC es más estrecho cuanto mayor es el tamaño de la muestra.

Pero los libros dan esta fórmula para calcular el tamaño, N, de la muestra:

$$N = \frac{p'(1-p')Z^2}{d^2}$$

Vemos que N depende de tres cantidades, p', Z y d, cuyo significado es:

p´ $\rightarrow$ Indica el orden de magnitud de la FR poblacional que intentamos conocer con el estudio que estamos proyectando. Esa FR no se conoce, por eso se va a hacer un estudio para averiguar cuánto vale. La cantidad p' a poner en la fórmula debe ser una que suponemos estará cercana a la FR que intentamos averiguar. Si no hay ninguna información al respecto, ponemos p' = 0.5.

d $\rightarrow$ Queremos error de estimación menor de "d", es decir, que la FR que aparezca en la muestra no se aleje de la verdadera FR poblacional en más de "d": $|\pi - p| < d$.

Esta cantidad debe decidirla el investigador y no hay criterios "matemáticos" ni "científicos" para hacerlo. Según el valor que se ponga para "d", la fórmula da un N.

Z $\rightarrow$ Una vez que el investigador ha decido el valor "d" que quiere usar, hay que decirle que no hay ningún N que garantice totalmente el cumplimiento de esa condición, es decir, que vaya a ser $|\pi - p| < d$. La confianza que el investigador tiene en que ocurra eso es mayor cuanto mayor es la muestra. El valor Z depende de cuán grande queremos que sea esa confianza. Por ejemplo, para 90% de confianza, Z es el valor que deja entre él y su negativo el 90% de los valores de la Distribución Normal, es decir, Z = 1.65. Para 95% de confianza es Z = 1.96, etc.

Vemos, pues, que la fórmula **no** es para determinar el "tamaño adecuado", sino para calcular las "prestaciones" que proporciona cada tamaño, y para poder utilizarla tenemos que decidir:

© Todos los derechos reservados

a) Una estimación de la proporción poblacional que queremos estimar, p´.

b) La diferencia máxima que no querríamos sobrepasar entre la FR muestral que obtengamos con la muestra que vamos a estudiar, p, y la poblacional, π.

c) La confianza que queremos en que ocurra lo anterior.

Apliquemos la fórmula para tener 90% de confianza (Z = 1.65) en que la p que aparezca en la muestra no se aleje más de 7 puntos porcentuales (d = 0.07) de la verdadera proporción poblacional, π, y suponiendo que la FR poblacional sea del orden de 10% (p´= 0.10).

$$N = (0.1 \times 0.9)\ 1.65^2 / 0.07^2 = \mathbf{50}.$$

Los valores de p', d y confianza elegidos son razonables, como también lo son otros muchos otros posibles y no hay razón objetiva que haga más válidas esas cantidades que otras parecidas. Pero con otros valores de confianza, error máximo deseable y p', el resultado puede ser muy distinto. Otras posibles elecciones son:

Confianza	d	p'		N
90%	.07	.10	$1.65^2(0.1 \times 0.9)/0.07^2=$	**50**
99%	.04	.10	$2.58^2(0.1 \times 0.9)/0.04^2=$	**374**
99%	.04	.50	$2.58^2(0.5 \times 0.5)/0.04^2=$	**10.40**
99.9%	.01	.50	$3.3^2(0.5 \times 0.5)/0.01^2=$	**27.225**

¿Qué nos garantizan cada uno de estos tamaños? Si tomamos una muestra de N = 50 tenemos probabilidad 0.90 de que la FR encontrada en la muestra no se aleje de la poblacional, π, en más de 0.07, si π es del orden del 0.10. ¿Qué quiere decir esa probabilidad 0.90? Que en 90 de cada 100 veces que repitamos la toma de una muestra de ese tamaño, la p muestral no se aleja de la π poblacional más de 0.07.

Y si tomamos una muestra de N = 374 individuos tenemos probabilidad 99% de que la FR que obtendremos en esa muestra no se alejará más de 0.04 de la poblacional que queremos estimar, si ésta es del orden de 0.10.

11.3. TAMAÑO DE MUESTRA PARA ESTIMAR DIFERENCIA DE DOS PROPORCIONES

Muchos estudios están enfocados a valorar la diferencia de dos FR poblacionales. Un ejemplo típico es comparar la eficacia de un nuevo fármaco B con la del antiguo A, cuando no conocemos con precisión la FR de curaciones realmente obtenidas con A, π_A (si conocemos π_A es la situación del apartado anterior).

© Todos los derechos reservados

Llamamos "efecto de B" en población al aumento de curaciones en relación con el A, es decir, efecto real = $\pi_B - \pi_A$ y el efecto de B en las muestras es la diferencia de las FR de curados en ellas, $p_A - p_B$. Trataremos N enfermos con cada fármaco y a partir del efecto observado en las muestras estimamos el efecto real, $\pi_B - \pi_A$. ¿Qué tamaño deben tener las muestras? No hay un tamaño "adecuado". Cuanto mayores sean la muestras más estrecho será el intervalo de confianza para $\pi_B - \pi_A$.

La fórmula para calcular el tamaño, N, de cada muestra es:

$$N = \frac{\left[p'_A\left(1-p'_A\right)+p'_B\left(1-p'_B\right)\right]\cdot Z^2}{d^2}$$

pA' y **pB'** → Indican de qué orden son las FR de curaciones con cada fármaco, es decir, son los estimadores de π_A y de π_B. Si, como es lo habitual, se sospecha que B cura más que A pero no se sabe cuánto, se puede poner $p_A' = p_B' = 0.5$, con lo cual se obtiene tamaño *igual o mayor* que el necesario para la confianza y error máximo elegido. Al poner en la fórmula $p_A' = p_B' = 0.5$ no se está indicando que ambos curan la mitad de los enfermos, sino que la FR de curaciones no es muy lejano al 50%, aunque suponemos que B curará más que A, como ocurriría, por ejemplo, si A curara el 45% y B curara el 55%.

d → Queremos error de estimación menor de "d", es decir, que el efecto muestral, $P_B - P_A$, no se aleje del efecto poblacional en más de "d": $|\pi_B - \pi_A| - |p_B - p_A| < d$. Esta cantidad debe decidirla el investigador y no hay criterios "científicos" para hacerlo. Según el valor que se ponga para "d" en la fórmula va a salir un tamaño u otro y el tamaño que sale con cada valor de "d" que se ponga es el adecuado para ese error.

Z → Refleja la confianza que queremos tener en que sea $|\pi_B - \pi_A| - |p_B - p_A| < d$. No hay ningún tamaño que garantice totalmente el cumplimiento de esa condición. La confianza en que ocurra es mayor cuanto mayor son las muestras. Por ejemplo, para tener 90% de confianza en que la $|p_B - p_A|$ que aparezca en la muestra no se aleje más de 10 puntos porcentuales (d = 0.10) de $|\pi_B - \pi_A|$ y suponiendo que A cura en torno al 10% y B cura en torno al 60%, es decir, $p_A' = 0.10$, $p_B' = 0.60$, d = 0.10 y Z = 1.65:

$$N = [(0.1 \cdot 0.9)+(0.6 \cdot 0.4)]\ 1.65^2/0.1^2 = \mathbf{89.8 \approx 90}.$$

Los valores de p', d y confianza elegidos son razonables, como también lo son otros muchos posibles y no hay razón objetiva que haga más válidas esas cantidades que otras parecidas. Con otros valores de confianza, error máximo deseable y p', el resultado puede ser muy distinto. Véalo:

CONF.	p_A'	p_B'	d	$[p_A'(1 - p_A') + p_B'(1 - p_B')]\ Z^2/d^2$	N
0.90	0.10	0.60	0.10	0.33 · 2.72 / 0.01	**90**
0.90	0.50	0.50	0.02	0.50 · 2.72 / 0.0004	**3.400**
0.99	0.10	0.60	0.10	0.33 · 6.65 / 0.01	**219**
0.99	0.20	0.30	0.02	0.37 · 6.65 / 0.0004	**6.144**

© Todos los derechos reservados

Es decir, si A cura en torno al 10% y B cura en torno al 60%, tomando N = 90 en cada grupo, tenemos probabilidad 0.90 de que el efecto encontrado en la muestra no se aleje del poblacional en más de 0.10, (90 de cada 100 veces que repitamos el estudio con esos tamaños será $|\pi_B - \pi_A| - |p_B - p_A| < 0.10$).

Y si A cura en torno al 20% y B cura en torno al 30%, tomando N = 6.144 en cada grupo, tenemos probabilidad 0.99 de que el efecto encontrado en la muestra no se aleje del poblacional en más de 0.02.

11.4. TAMAÑO DE MUESTRA PARA ESTIMAR UNA MEDIA

El tamaño "adecuado" de una muestra para estimar la media poblacional, μ, de una variable cuantitativa se calcula, presuntamente, con esta fórmula:

$$N = \frac{Z^2\sigma^2}{d^2}$$

$\sigma 2$ $\rightarrow$ Es una estimación de la varianza poblacional, $\sigma 2$. Para la estimación de una FR poblacional si no teníamos noticia alguna sobre la magnitud de la π se ponía p' = 0.5. En este caso no hay un valor al que podamos recurrir si no tenemos información al respecto. Y hay que decidirse por alguna cantidad para que la fórmula opere.

d $\rightarrow$ Queremos que la media de la muestra no se aleje de la media poblacional en más de "d" unidades, es decir, queremos $|\mu - M| < d$.

Z $\rightarrow$ Depende de la confianza que queremos tener en que sea $|\mu - M| < d$. Es el valor de Z que en la Distribución Normal deja entre él y su negativo una FR de valores igual a esa confianza. Para 95% de confianza es Z = 1.96, para 99% de confianza es Z = 2.58, etcétera.

En función de los valores que el investigador decida para las tres cantidades anteriores, la fórmula dará valores diferentes de N. Veamos diferentes posibilidades:

Opción 1ª:

Confianza = 0.95, d = 2, $\sigma = 8 \rightarrow N = 1.96^2 \cdot 8^2 / 2^2 = 61.46 \approx 62$.

Con una muestra de N = 62 tenemos probabilidad 0.95 de que la media encontrada en la muestra no se aleje de la poblacional en más de 2, si la σ es del orden de 8.

Opción 2ª:

Confianza = 0.99, d = 1, $\sigma = 12 \rightarrow N = 2.58^2 \cdot 12^2 / 1^2 = 958.5 \approx 959$.

Con una muestra de N = 959 tenemos probabilidad 0.99 de que la media encontrada en la muestra no se aleje de la poblacional en más de 1, si la σ es del orden de 12.

© Todos los derechos reservados

Es decir, no existe el tamaño "adecuado" en general, sino el tamaño "adecuado" para tener cierta confianza en que la Media Muestral se alejará de la poblacional menos de "d" unidades.

11.5. ELEMENTOS DE SUBJETIVIDAD E IMPRECISIÓN AL APLICAR LA FÓRMULA PARA ESTIMAR UNA MEDIA

1º Para aplicar la fórmula que permita calcular el tamaño de muestra es necesario conocer el valor de la σ de esa variable en la población. Si proyectamos ese estudio porque no conocemos la media poblacional, es poco probable que conozcamos la desviación estándar.

2º ¿Cuánta diferencia podríamos aceptar entre la media muestral y la poblacional? Si queremos, por ejemplo, estimar la media poblacional del dinero que cada persona gasta en cosméticos cada año, unos pueden pensar que es suficiente que la media muestral no se aleje de la poblacional en más de US50, otros prefieren que no se aleje en más de US20 y otros aceptarían el error de estimación hasta US80.

3º Ningún tamaño de muestra garantiza totalmente que la media muestral y la poblacional se diferencien entre sí menos de cierta magnitud. Cuanto mayor sea la muestra más probable es que la diferencia sea pequeña. El investigador tiene que decidir cuánto quiere que sea esa probabilidad.

4º La mayoría de los estudios implican interés en más de una variable. La fórmula dará diferente resultado según la variable a la que se refiera.

5º Muchas comparaciones de interés afectarán a subgrupos de la muestra, cuyos tamaños no se pueden predecir de antemano.

© Todos los derechos reservados

11.6. TAMAÑO DE MUESTRA PARA ESTIMAR DIFERENCIA DE DOS MEDIAS

Se quiere estimar el posible efecto del ejercicio físico moderado, EFM, sobre la variable "X", concentración de glucosa en sangre en personas de 60 años. El efecto real del EFM es la diferencia entre las medias poblacionales de los sedentarios y de los que hacen EFM, $\mu_{SED}-\mu_{EFM}$. Estimaremos el efecto real por el efecto observado en las muestras, $M_{SED}-M_{EFM}$. ¿Cuántos individuos debe tener cada una de las dos muestras? La fórmula para calcular el tamaño, N, de cada muestra es:

$$N = \frac{\left(\sigma^2{}_{SED} + \sigma^2{}_{EFM}\right) Z_\alpha{}^2}{d^2}$$

σ_{SED} y σ_{EFM} → Son estimadores de las desviaciones estándar poblacionales. No tienen que ser iguales, pero en la mayoría de los casos no se conocen motivos para que sean diferentes y se estiman por la misma cantidad.

Z y d → Queremos tener cierta confianza en que el efecto encontrado en la muestra no se aleje del poblacional más de "d" unidades: $|(M_{SED}-M_{EFM}) - (\mu_{SED}-\mu_{EFM})| < d$. Por ejemplo, para tener confianza 90% en que el efecto encontrado en la muestra, $M_{SED}-M_{EFM}$, no se aleje más de 3 unidades (d = 3) del efecto real de la población, $\mu_{SED}-\mu_{EFM}$, y si σ_{SED} y σ_{EFM} están ambas en torno a 8.

Es $N = [64+64]\ 1.65^2/3^2 = 38.72 \approx 39$.

Con otros valores de confianza, error máximo deseable y σ, el resultado puede ser muy distinto. Estas son otras posibles elecciones:

CONF.	σ'_{SED}	σ'_{EFM}	d	$(\sigma^2_{SED}+\sigma^2_{EFM})\ Z_\alpha{}^2/d^2$	N
0.90	8	8	3	128 · 2.72 / 9	**39**
0.90	12	12	3	288 · 2.72 / 9	**87**
0.90	12	12	1	288 · 2.72 / 1	**784**
0.99	12	12	1	288 · 6.66 / 1	**1917**

Si la σ es del orden de 8 en ambas poblaciones, tomando N = 39 individuos en cada grupo tenemos probabilidad 0.90 de que el efecto encontrado en la muestra no se aleje del poblacional en más de 3 unidades.

Y si la σ es del orden de 12 en ambas poblaciones, tomando N = 1 917 en cada grupo, tenemos probabilidad 0.99 de que el efecto encontrado en la muestra no se aleje del poblacional en más de 1 unidad.

11.7. MITOS Y REALIDADES EN EL TAMAÑO DE LA MUESTRA

Los investigadores creen que la estadística proporciona fórmulas que dan la cantidad de individuos que deben ser estudiados. Eso no es cierto, pero muchos miembros de los comités encargados de aprobar y financiar los proyectos de investigación participan de esa errónea creencia y exigen que en los proyectos el tamaño de las muestras se justifique con esas formulas. La mayoría de los investigadores no saben cómo usar esas fórmulas ni cómo se interpreta su resultado, pero al verse obligados a manejarlas, en los proyectos de investigación ponen frases que no entienden, copiadas de otros proyectos, cuyos autores tampoco las entendían y las habían, a su vez, copiado de otros. Esto da lugar a una cadena de simulacros que lleva a una situación esperpéntica. Los pasos de esa secuencia son:

© Todos los derechos reservados

1. En el proceso de copia "ciega" se van acumulando errores que los copistas no detectan porque no entienden lo que están copiando.
2. La mayoría de los evaluadores arriba citados no entiende esas frases y las dan por buenas, incluyendo las equivocadas.
3. Los proyectos aprobados caen en manos de otros autores de nuevos proyectos, que repiten el proceso de copia añadiendo nuevas errores.
4. Tras varios decenios este ritual nos ha llevado a que en muchos proyectos se lean frases totalmente absurdas pero siempre aprobadas por los evaluadores.
5. En algunas ocasiones este simulacro se hace también de viva voz, en la exposición de los trabajos de Máster y Tesis Doctoral. En relación al tamaño de la muestra el investigador dice con aplomo una frase ininteligible y el tribunal asiente con solemnidad. Podría hablarse de un acuerdo "doble ciego", donde ninguna de las dos partes entiende nada, pero ambas asienten.

Como ejemplo de frases inadecuadas que se van deteriorando en las sucesivas copias a ciegas, podría tomarse que el rendimiento de los médicos de hospitales públicos se evalúa según una escala que va de 0 a 200. Para ver si el rendimiento en su trabajo hospitalario matinal es menor en los que tienen consulta privada por las tardes se evaluará dos muestras de N médicos.

¿Cuántos médicos "debe" haber en cada grupo? Ninguna fórmula dice qué tamaño "debemos" tomar, sino las prestaciones producidas por los diversos tamaños. Si estimamos que la σ está entre 4 y 7, en el proyecto en que se pide dinero para esta investigación lo correcto sería dar una tabla de este tipo.

	$\sigma_{SIN} = 4$	$\sigma_{CON} = 4$	$\sigma_{SIN} = 7$	$\sigma_{CON} = 7$
	Conf. = 90%	**Conf. = 99%**	**Conf. = 90%**	**Conf. = 99%**
d = 4	6	14	17	41
d = 2	22	54	66	163

Pero el investigador que no conoce este tema le pide a un estadístico que aplique la fórmula y acaba eligiendo, por ejemplo, $\sigma_{SIN} = \sigma_{CON} = 4$, confianza 99% y d = 2, con lo que sale **N = 54**. Por supuesto, podría acordarse otra confianza y otro error máximo, por lo que no hay ningún motivo para elegir N = 54 en vez de otro tamaño.

Pero el investigador no entiende las explicaciones del estadístico y acaba tomando nota de una frase que intenta transcribir al pie de la letra, pero en la que es fácil que se introduzca algún error. La frase que el estadístico le dicta podría ser ***Si las desviaciones son del orden de 4, con 54 individuos en cada grupo, tenemos confianza 99% en que el efecto muestral no difiera del poblacional en más de 2 unidades***.

El investigador acaba escribiendo en el proyecto:

> "El tamaño adecuado es N = 54, teniendo confianza 99% y orden de 4 en que el efecto muestral no difiera en más de 2 unidades de las desviaciones".

© Todos los derechos reservados

El siguiente investigador la copia con alguna pequeña variación.

> "El tamaño adecuado es N = 54, teniendo confianza ordenada 99% y 2 unidades de 4 desviaciones en que el efecto muestral no difiera significativamente".

Y cuando un tercer investigador copia esta frase la versión final es:

> "El tamaño adecuado es N = 54, teniendo 4 unidades y dos desviaciones de confianza ordenada significativamente en 99%".

Pero muy probablemente los evaluadores del proyecto la darán por buena. La situación recuerda la celebrada en todos los institutos de enseñanza media de España como una de las más jocosas. En la Roma Imperial, antes de empezar su encarnizada lucha, los gladiadores debían decir al Cesar: "*Ave, Caesar, morituri te salutan*". Un alumno de último año de bachiller la tradujo del latín como "*Las aves del Cesar murieron por falta de salud*". La diferencia entre ese episodio y el de las frases escritas y aceptadas sobre el tamaño de las muestras, es que todos los profesores y alumnos saben que esa traducción es un despropósito total, mientras que la mayoría de los que escriben y evalúan proyectos de investigación creen que las frases referidas al tamaño de las muestras son correctas. Cada uno de ellos no las entiende, pero da por supuesto que los otros las entienden y avalan. Salir de ese embrollo no requiere que los investigadores y los evaluadores se hagan expertos en Estadística avanzada. Solo es necesario tener claras algunas ideas básicas del análisis estadístico.

11.8. TAMAÑOS DE MUESTRA INADECUADOS

Aunque no hay un tamaño de muestra que sea "el adecuado", pues muy diferentes tamaños son válidos, ciertos tamaños son claramente "inadecuados", por ser excesivamente grandes o excesivamente pequeños. Supongamos que en una población se quiere estimar el sueldo medio con error no superior a 50€ y se sabe que la desviación es en torno a 300€. Si nos proponen tomar una muestra de N = 10, ¿qué podemos esperar de ella?

En la fórmula $N = Z^2\,\sigma^2/d^2$ despejamos $Z^2 = N \cdot d^2/\sigma^2 = 10 \cdot 502/3002 = 0.278$. Por tanto, Z = 0.53, que en las tablas de la Normal, nos da $P(-0.53 < Z < +0.53) = 2 \cdot (0.5 - 0.30) = 2 \cdot 0.20 = 0.40$. Es decir, con N = 10 tenemos confianza 40% en que la media muestral no se aleje de la poblacional en más de 50€. Por tanto, tenemos riesgo 60% de que la media muestral se aleje de la poblacional más de 50€. Si al investigador le parece riesgo demasiado alto y no tiene recursos para más tamaño, deberá posponer la investigación hasta conseguirlo.

Calculemos el tamaño necesario para que la confianza sea mayor, por ejemplo 90% (Z = 1.96) con $\sigma = 300$ y d = 50.

$$N = 1.65^2 \bullet 300^2/50^2 = 98$$

Si nuestros recursos no alcanzan para estudiar una muestra de 98 individuos, pero nos permiten llegar a N = 50, podemos calcular la confianza que ella proporciona:

© Todos los derechos reservados

$Z^2 = N \cdot d^2/\sigma^2 = 50 \cdot 50^2/300^2 = 1.389 \rightarrow Z = 1.18 \rightarrow P\ (-1.18 < Z < +1.18) = 0.76$

Con N = 50 tenemos confianza 76% en que la media muestral se aleje de la poblacional menos de 50€. El investigador podría considerarlo demasiado alto y optar por esperar a conseguir más recursos. Es una decisión que solo él puede tomar.

La Estadística no toma la decisión por él, pero le puede ayudar permitiéndole calcular los niveles de confianza para distintos errores de estimación.

Para el supuesto anterior otro ejemplo de tamaño no adecuado sería N = 5 000, porque si establecemos la confianza en 99%, es Z = 2.58. En $N = Z^2\ \sigma^2/d^2$ despejamos $d^2 = Z^2\ \sigma^2/N = 2.582 \cdot 3002/5.000 = 120$, es decir, d = 11. Vemos que con N = 5.000 tendríamos confianza 99% en que el error de estimación no supere los 11€, precisión innecesaria para el objetivo del estudio.

11.9. ASPECTOS A RECORDAR

- En todo estudio hay que decidir cuántos individuos se incluyen y muchos investigadores creen –erróneamente– que esa decisión se toma aplicando una fórmula.
- Las fórmulas relativas a este tema no dicen el tamaño que debe tener la muestra de una investigación, solamente ayudan a calcular las prestaciones que tendrá la muestra según sea su tamaño.
- Algunos investigadores no saben cómo usar esas fórmulas ni cómo interpretar su resultado, pero al verse obligados a manejarlas en los proyectos de investigación ponen frases que no entienden, copiadas de otros proyectos, cuyos autores tampoco las entendían y las habían, a su vez, copiado de otros.
- La realidad es que no hay un tamaño "más adecuado" que optimice la relación costo/beneficio. Casi todos los tamaños de muestra, excluidos los extremadamente pequeños o grandes, son válidos.
- Cuanto mayor es la muestra, más información nos da acerca de la población. Este incremento de la información al aumentar el tamaño de la muestra es gradual y no hay un punto de corte que separe tamaños suficientes de tamaños insuficientes.
- Las fórmulas no nos dicen cuál es el tamaño que debe usarse, sino cómo depende del tamaño la confianza en que el error de estimación no supere cierta cantidad. Esta confianza y la magnitud de ese error máximo debe elegirlas el investigador según su buen criterio. No hay argumentos matemáticos ni biológicos para ayudarle en esa elección.
- En el tamaño arrojado por la fórmula influye también la desviación estándar poblacional de las variables cuantitativas y la proporción poblacional de las variables dicotómicas. Como el investigador no conoce ese dato debe suponerlo, lo que introduce otro factor de imprecisión.
- Aunque no cabe hablar de tamaños adecuados, sí cabe hablar de tamaños inadecuados, bien porque son extremadamente pequeños y dan lugar a errores de estimación con margen muy amplio, o bien porque son extremadamente grandes y dan lugar a errores de estimación innecesariamente pequeños.

© Todos los derechos reservados

© Todos los derechos reservados

Resumen

© Todos los derechos reservados

© Todos los derechos reservados

Tablas de contingencia 12

Visión general

- Tablas con "R" muestras con "C" categorías.
- Comparación de varias proporciones.
- Muestreo para realizar una tabla de contingencia.
- Cálculo de Test con tablas de contingencia.

Objetivo

- Construir tablas de contingencia.

© Todos los derechos reservados

12.1. TABLAS RXC: R MUESTRAS CON C CATEGORÍAS

En una muestra de 300 personas se clasifican según el grupo de edad al que pertenecen y según la cantidad de deporte que hacen.

	Adolescente AD	Joven JO	Maduro MA	Viejo VI	Total
Mucho deporte = D2	16	10	2	2	30
Deporte moderado = D1	12	8	2	4	26
No deporte = D0	52	42	16	34	144
Total	80	60	20	40	200

Para estudiar la relación entre estas dos variables, haremos como en el Capítulo 3, la DF de Deporte condicionado a edad.

	AD	JO	MA	VI	Total
Mucho deporte = D2	0.20	0.17	0.10	0.05	0.15
Deporte moderado = D1	0.15	0.13	0.10	0.10	0.13
No deporte = D0	0.65	0.70	0.80	0.85	0.72
Total	1	1	1	1	1

Entre los adolescentes hacen mucho deporte el 20% (= 16/80), deporte moderado el 15%, y no hacen el 65%. En esta muestra la distribución de deporte no es igual en las 4 edades. Y nos preguntamos si también en la población habrá diferencias. La muestra nunca permite asegurar que en población sean iguales las distribuciones (aunque en las muestrales fueran muy parecidas o incluso iguales), pero puede permitir sospechar que en población no lo son, si en la muestra son muy diferentes. Planteamos como H_0 que las poblacionales son iguales y se rechaza si es muy improbable que las muestrales sean tan diferentes como las de nuestra muestra o aún más diferentes. El valor P del test nos da esa probabilidad.

En este capítulo vemos la Inferencia con estas Tablas de Contingencia que se forman al relacionar variables cualitativas o cuantitativas agrupadas en intervalos.

En general, hablamos de tablas de contingencia "RxC" siendo "R" el número de filas (de *row* en inglés) y C el número de columnas. Veamos un ejemplo de tabla 3x3. En una muestra de N = 200 individuos se estudia la relación entre raza ("A", "B" y "C") y resistencia a las infecciones (baja, media, alta). La tabla a la izquierda da el número de individuos en cada raza y resistencia.

Datos				
	Baja	Media	Alta	Total
A	2	22	16	40
B	12	30	18	60
C	26	48	26	100
Total	40	100	60	200

FR de resistencia en cada raza				
	Baja	Media	Alta	Total
A	**.05**	**.55**	**.40**	**1**
B	**.20**	**.50**	**.30**	**1**
C	**.26**	**.48**	**.26**	**1**
Total	.20	.50	.30	

Vemos en los marginales que hay 40 de la raza A, 60 de B y 100 de C y que hay 40 con resistencia baja, 100 con media y 60 con alta. La tabla de la derecha da las FR de individuos de cada tipo de resistencia condicionada a cada raza. Vemos que la resistencia tiene las cifras más favorables en la raza A (40% alta y 5% baja) y las menos favorables en C (26% alta y 26% baja). Se plantea como H_0 que raza y resistencia son independientes en población, es decir, la DF de la resistencia es igual en las tres razas. Si en la muestra la DF de la resistencia es parecida en las tres Razas, no rechazaremos H_0, y si la DF muestral de la Resistencia es muy diferente entre las razas, rechazaremos H_0. El cálculo del valor P del test puede ser pesado pero todos los paquetes informáticos lo dan. Aquí resulta ser 0.08. Si fuera cierta la H_0, diferencias muestrales como estas o mayores aparecerían por azar en 8 de cada 100 veces que repitiéramos el estudio. No hay evidencia fuerte contra H_0, no queda probado que la evolución sea diferente en alguna de las razas.

© Todos los derechos reservados

Cuanto mayor son las diferencias entre las distribuciones muestrales menor es el valor P. La siguiente tabla tiene datos muy parecidos a la anterior pero algunas diferencias aumentadas (en la raza C ahora hay 42 individuos con resistencia baja y 10 con alta).

Datos	Baja	Media	Alta	Total
A	2	22	16	40
B	12	30	18	60
C	42	48	10	100
Total	56	100	44	200

FR de resistencia en cada raza	Baja	Media	Alta	Total
A	.05	.55	.40	1
B	.20	.50	.30	1
C	**.42**	.500	**.10**	1
Total	.23	.50	.22	

Ahora es P = **0.00001**

Y unas mismas diferencias es más fácil que aparezcan por azar cuando los tamaños son menores. En la siguiente tabla las FR de resistencia son iguales que en la anterior pero el tamaño de las muestras es la mitad:

Datos	Baja	Media	Alta	Total
A	1	11	8	20
B	6	15	9	30
C	**21**	24	5	50
Total	23	50	22	100

FR de resistencia en cada raza	Baja	Media	Alta	Total
A	.05	.55	.40	1
B	.20	.50	.30	1
C	**.42**	.500	**.10**	1
Total	.23	.50	.22	

Ahora es P = **0.00400**

Veamos ahora un ejemplo de **tabla 2 x 4**. En la variable "Evolución" de los pacientes con neumonía se consideran C = 4 categorías: mal, regular, bien y muy bien. Se compara la evolución en una muestra de **40** casos causados por bacterias frente a la evolución de otra muestra de **60** casos causados por virus.

Valores absolutos	Mal	Reg.	Bien	M-B	Total
Bacteria	16	12	8	4	40
Virus	6	15	15	24	60

Proporciones	Mal	Reg.	Bien	M-B
	.40	**.30**	**.20**	**.10**
	.10	**.25**	**.25**	**.40**

En la muestra de 40 casos bacterianos el 40% (16/40) evolucionan mal, el 30% (12/40) regular, etcétera. En las *muestras* observamos que la evolución no es igual en las dos etiologías. Es más favorable en la vírica. Pero podría ser que las FR *poblacionales* fueran iguales en ambas etiologías. Planteamos esa H_0 y se encuentra **P = 0.0004**. Nos da la probabilidad de que por azar aparezcan diferencias entre las DF muestrales como estas o aún mayores, si no hubiera diferencias entre las

© Todos los derechos reservados

DF poblacionales. Invita a pensar que la evolución no es igual en las dos poblaciones. En el último apartado se explica el cálculo del valor P, pero lo importante para el investigador no es saber hacer ese cálculo, sino interpretarlo correctamente cuando lo calcula el ordenador o lo lee en las publicaciones científicas.

Es obvio que el valor P es menor cuanto más se diferencian entre sí las distribuciones muestrales y/o cuanto mayor son las muestras. En la siguiente tabla las diferencias entre las distribuciones muestrales son menores que en la anterior.

	Mal	Reg.	Bien	M-B	Total
Bacteria	16	12	8	4	40
Virus	12	15	15	18	60

Mal	Reg.	Bien	M-B
.40	**.30**	**.20**	**.10**
.20	**.25**	**.25**	**.30**

Aquí es **P = 0.04**. Pero esas mismas diferencias entre las dos distribuciones muestrales es más difícil que aparezcan por la VRM con muestras mayores. En la tabla siguiente las FR son iguales, pero el tamaño de las muestras es el doble.

	Mal	Reg.	Bien	M-B	Total
Bacteria	32	24	16	8	80
Virus	24	30	30	36	120

Mal	Reg.	Bien	M-B
.40	**.30**	**.20**	**.10**
.20	**.25**	**.25**	**.30**

Ahora es **P = 0.0009**. Es más difícil que aparezcan estas diferencias.

© Todos los derechos reservados

12.2. COMPARACIÓN DE VARIAS PROPORCIONES

Retomemos la tabla de contingencia 3 x 4 de la introducción:

	Adolescente AD	Joven JO	Maduro MA	Viejo VI	Total
Mucho deporte = D2	16 (0.20)	10 (0.17)	2 (0.10)	2 (0.05)	30 (0.15)
Deporte moderado = D1	12 (0.15)	8 (0.13)	2 (0.10)	4 (0.10)	26 (0.13)
No deporte = D0	52 (0.65)	42 (0.70)	16 (0.80)	34 (0.85)	144 (0.72)
Total	80	60	20	40	200

Cada celda da la frecuencia absoluta y la FR respecto al total de esa edad. Aquí la H_0 es que en población la DF de deporte es la misma en todas las edades. El test da P = 0.34, por lo que no la rechazamos. Los datos no son evidencia fuerte a favor de que la práctica del deporte varíe con la edad. Pero podemos continuar el análisis juntando ambos grupos de deportistas en uno solo. Así la tabla de 2x4 es:

	AD	JO	MA	VI	Total
Deporte, D	28	18	4	6	56
No deporte	52	42	16	34	144
Total	80	60	20	40	200
% **D+**	**0.35**	**0.30**	**0.20**	**0.15**	0.28

De los 80 adolescentes, 28 son deportistas, un 35%; de los 60 jóvenes, 18 son deportistas, un 30%, etc. Como la DF de deporte para cada edad tiene solamente dos posibles valores (deportistas y no deportistas) y como las FR de ambas suman la unidad, la situación queda descrita dando solamente una de ellas. Para destacarlas más se registran en la última fila. Ahora se trata de comparar estas cuatro proporciones y la H_0 es que en población las 4 son iguales. Se quiere ver si las diferencias observadas en las muestras pueden deberse al azar o son tan grandes que llevan a rechazar H_0. El test da P = 0.11. Las diferencias muestrales no son suficiente evidencia a favor de que haya diferencias en población.

Si las distintas muestras se pueden ordenar *a priori* por algún criterio, además del test anterior puede estudiarse si la proporción de casos "positivos" crece o decrece progresivamente con la variación de dicho criterio. En este ejemplo las cuatro situaciones se han ordenado de menor a mayor edad. El test de **Tendencia Lineal** propone como H_0 que no hay crecimiento ni decrecimiento progresivo al aumentar la edad. Los cálculos para obtener el valor P son tediosos, pero los hace el ordenador. Se obtiene P = **0.01**, es decir, si en población la proporción de deporte no crece ni decrece progresivamente con la edad, en 1 de cada cien veces que repitiéramos este estudio encontraríamos una tendencia decreciente como la de nuestra muestra o aún mayor. Es una notable evidencia contra la H_0 y a favor de que la proporción de deporte disminuye con la edad en la población.

Para ver las diferencias entre el test general, que detecta cualquier tipo de diferencias entre las FR de los 4 grupos, y el test de Tendencia Lineal, que detecta específicamente *crecimiento* o *decrecimiento progresivo*, consideremos estos datos, muy parecidos a los de la tabla anterior, pero están intercambiadas las proporciones de deporte en Jóvenes y Viejos. Aquí la proporción de deporte es 0.15 en los Jóvenes y 0.30 en los Viejos.

	AD	JO	MA	VI	Total
Deporte, D+	28	6	4	18	56
No deporte	52	34	16	42	144
Total	80	40	20	60	200
% D+	**0.35**	**0.15**	**0.20**	**0.30**	0.28

El test general da P = 0.11, como antes, puesto que son los mismos 4 grupos. Pero el de Tendencia Lineal ahora da P = 0.55, puesto los grupos están ordenados de distinta forma y ahora ya no se ve una tendencia decreciente.

© Todos los derechos reservados

Aún podemos considerar otra variación de los datos. La siguiente tabla es muy parecida a la anterior pero la proporción de deporte es mayor que antes en adolescentes y en viejos. Al ser mayores las diferencias entre las 4 proporciones muestrales esperamos un valor P menor en el test general.

	AD	JO	MA	VI	Total
Deporte, D+	44	6	4	30	56
No deporte	36	34	16	30	144
Total	80	40	20	60	200
% D+	**0.55**	**0.15**	**0.20**	**0.50**	0.28

Obtenemos P = 0.0002 para el test general que plantea la hipótesis de igualdad de FR poblacionales, pero sigue sin haber tendencia al crecimiento o decrecimiento al aumentar la edad, por lo que para el test de Tendencia Lineal esperamos un valor de grande. En efecto, los cálculos dan P = 0.56. Con estos datos concluiríamos que la FR poblacional de deporte no es igual en las cuatro edades, pero no concluiríamos que esas FR aumenten o decrezcan progresivamente al aumentar la edad.

12.3. COLAPSANDO TABLAS

En algunos casos puede ayudar al análisis fundir dos o más categorías en una sola, si la naturaleza del problema lo permite. En el apartado anterior se colapsaron las celdas de mucho y moderado deporte en una sola, generando la tabla:

	AD	JO	MA	VI	Total
Deporte, D+	28	18	4	6	56
No deporte	52	42	16	34	144
Total	80	60	20	40	200
% **D+**	**0.35**	**0.30**	**0.20**	**0.15**	0.28

Para la H_0: en población la FR de deporte es igual en las 4 edades, el valor P = 0.11 no es evidencia contra ella. Como en la muestra la FR de deporte es menor en las edades superiores, se analizó específicamente la Tendencia Lineal al decrecimiento con el aumento de la edad y se encontró P = 0.01, que es un notable aval a la sospecha de que hay esa tendencia en la población.

Otro enfoque podría ser considerar que debido a la presión de la vida laboral y a las costumbres imperantes la bajada en la FR real (es decir, en población) de deporte se produce sobre todo al pasar de joven a maduro. Podemos plantear como H_0 que la FR es en población igual en adolescentes y jóvenes, mientras que también es muy parecida en maduros y viejos, pero menor que en las edades juveniles.

© Todos los derechos reservados

Para analizar esa H_0 colapsamos las dos primeras columnas y las dos últimas, teniendo entonces solo dos grupos de comparación: adolescentes y jóvenes frente a maduros y viejos.

	AD + JO	MA + VI	Total
Deporte, D+	46	10	56
No deporte	94	50	144
Total	140	60	200
% D+	**0.33**	**0.17**	0.28

En la muestra la diferencia en la FR de deporte entre esos dos grupos etáreos es 0.33–0.17 = **0.16** y el test que plantea como H_0 que en población esa diferencia es cero, da P = **0.0116**, es decir, notable evidencia de que la FR de deportistas es en población mayor en el primer grupo. El hecho de tener solamente dos grupos permite expresar las diferencias de proporciones de deporte mediante un número, en este caso **0.16**, y calcular un IC para el valor poblacional de esa diferencia. En este caso el IC al 95% sería **0.04–0.28**. La FR de deporte en la población es entre 4 y 28 puntos mayor en el primer grupo.

Como segundo ejemplo de colapso veamos de nuevo los datos de la tabla donde se comparaba la resistencia a la infecciones en tres razas:

Datos originales				
	Baja	Media	Alta	Total
A	2	22	16	**40**
B	12	30	18	**60**
C	26	48	26	**100**
Total	40	100	60	200

FR de cada nivel de resistencia en cada raza				
	Baja	Media	Alta	
A	**.05**	**.55**	**.40**	1
B	**.2**	**.50**	**.30**	1
C	**.26**	**.48**	**.26**	1
Total	.20	.50	.30	

En la muestra hay diferencias moderadas entre las tres razas, pero el valor P era 0.08, solamente una modesta evidencia contra la H_0. Si el investigador tuviera motivos previos a la obtención de los datos para pensar que quizá la resistencia es semejante en B y C, podría:

a) Hacer test comparando B contra C para confirmar que las diferencias muestrales entre ellos son compatibles con la igualdad en población.

Datos				
	Baja	Media	Alta	Total
B	12	30	18	**60**
C	26	48	26	**100**

FR de resistencia en cada raza				
	Baja	Media	Alta	
B	**.20**	**.50**	**.30**	1
C	**.26**	**.48**	**.26**	1

En el test da **P = 0.66**. Las divergencias muestrales bien pueden ser por azar.

© Todos los derechos reservados

b) Si el anterior test no da un valor P muy pequeño, se puede hacer un test comparando A contra B+C.

	Baja	Media	Alta	Total
A	2	22	16	40
B+C	38	78	44	160
Total	40	100	60	200

	Baja	Media	Alta	
A	**.050**	**.550**	**.400**	1
B+C	**.237**	**.468**	**.275**	1
Total	.20	.50	.30	

Ahora es **P = 0.023** que constituye cierta evidencia a favor de que hay diferencias entre estos dos grupos.

Podría ahora considerarse, si fuera lógico hacerlo independientemente de la observación de los datos, que quizá la FR con evolución media es en población muy similar en los dos grupos étnicos ahora considerados, pero puede haber diferencia entre baja y alta. Para comparar alta contra baja se forma esta tabla excluyendo a los de la clase media.

	Baja	Alta	Total
Blanca	2	16	18
Neg+Orie	38	44	82
Total	40	60	100

	Baja	Alta	
A	**.11**	**.89**	1
B+C	**.46**	**.54**	1
Total	.20	.30	

Se plantea como H_0 que la FR poblacional de alta es igual en ambos grupos. Ahora es P = **0.003**. Al tratarse de la comparación de dos FR, lo más adecuado es dar su diferencia en las muestras y el IC para el valor de esa diferencia en la población. En este ejemplo diríamos que una vez excluidos los de clase media, la FR con resistencia alta es en A 0.35 mayor que en los B+C (0.89 – 0.54 = 0.35) y que el IC al 95% es 0.17 y 0.53.

Las posibilidades de colapsar tablas son muchas, pero deben basarse en razones lógicas independientemente de la observación de los datos.

12.4. EL TIPO DE MUESTREO EN LAS TABLAS DE CONTINGENCIA

A los datos de una Tabla de Contingencia se puede llegar a través de diferentes tipos de muestreo. Dependiendo de cómo hayan sido muestreados los individuos de la población, podremos estimar unas u otras FR poblacionales.

© TODOS LOS DERECHOS RESERVADOS

Muestreo 1

Reconsideremos el ejemplo en que se toma una muestra de cada una de tres razas y una vez tomada se cuenta el número de personas con cada nivel de resistencia. El investigador decidió estudiar 40 de raza A (2 resultaron tener baja resistencia, 22 media y 16 alta), 60 de B y 100 de C.

Estos son los datos:

Datos originales				
	Baja	Media	Alta	Total
A	2	22	16	**40**
B	12	30	18	**60**
C	26	48	26	**100**
Total	40	100	60	200

FR de cada nivel de resistencia en cada raza				
	Baja	Media	Alta	
A	**.05**	**.55**	**.40**	1
B	**.20**	**.50**	**.30**	1
C	**.26**	**.48**	**.26**	1
Total	.20	.50	.30	

La FR muestral de personas con cada nivel de resistencia en cada raza estima la correspondiente FR poblacional y a partir del valor muestral se puede calcular el IC para el valor poblacional. Por ejemplo, en los A la FR poblacional con resistencia baja se estima con 2/40 = 0.05 y la FR poblacional de C con resistencia media se estima con 48/100 = 0.48.

Pero con este muestreo no se puede estimar la FR poblacional de cada nivel de resistencia ignorando la raza. Aunque en la población muestreada no hubiera otras razas aparte de A, B y C, la FR de personas con resistencia alta en la población general **no** se puede estimar con 60/200 = 0.30, porque ese valor depende de cuantos individuos se haya decidido tomar de cada raza. Y de igual forma, 40/200 = 0.20 no estima la FR poblacional con resistencia media, ni 60/200 = 0.30 estima la FR de individuos con resistencia alta en la población.

Y, por ejemplo, con las muestras tomadas de esa forma la FR marginal de A en esa tabla (40/200 = 0.20) **no** refleja la FR de A en la población general, al igual que 100/200 = 0.5 no estima la FR de C en la población. Y de la misma forma ocurre con las distribuciones de raza condicionadas a resistencia: la FR de B condicionada a resistencia media en esa tabla (30/100 = 0.30) no refleja la FR de B en la población de personas con resistencia media, y 16/60 = 0.27 no estima la FR de A en la población de personas con alta resistencia.

Muestreo 2

Los datos de la tabla que estamos comentando podrían haberse obtenido tras tomar: una muestra de 40 sujetos con resistencia baja:

- Una muestra de 100 sujetos con resistencia media.
- Una muestra de 60 sujetos con resistencia alta.

Y contar en cada una de ellas los que había de cada raza.

© Todos los derechos reservados

Datos originales				
	Baja	Media	Alta	Total
A	2	22	16	40
B	12	30	18	60
C	26	48	26	100
Total	40	100	60	200

FR de cada nivel de resistencia en cada raza				
	Baja	Media	Alta	
A	.05	.22	.27	.2
B	.30	.30	.30	.3
C	.65	.48	.43	.5
Total	1	1	1	

En ese caso la FR de A en la población de personas de resistencia baja sí se estima por 2/40 = 0.05 Y la FR de B condicionada a resistencia media en esa tabla (30/100 = 0.30) sí refleja la FR de B en la población con resistencia media. Pero no se puede estimar la FR de A en la población general, pues el número de A depende de cuántos individuos de cada resistencia se decidió estudiar en las muestras.

Muestreo 3

Otro modo de muestrear habría sido tomar una muestra aleatoria de 200 individuos de la población general y contar el número de ellos de cada raza y cada resistencia. En este caso tanto las distribuciones marginales de la tabla como las condicionadas en ambos sentidos estiman las correspondiente FR poblacionales, siempre teniendo en cuenta la VRM.

De la misma forma que los distintos tipos de muestreos permiten estimar unas FR y otras no, también permiten plantear unos tests y otros no. Si es de interés, puede hacerse un test que plantee, por ejemplo, como H_0 cierta distribución de razas en la población (este test no tendría sentido hacerlo con el muestreo 1, en el que el investigador decidió tomar esas cantidades de individuos de cada raza).

Con el tipo 3 de muestreo podemos plantear, por ejemplo, la H_0 que dice que en la población hay 30% de A, 30% de B y 40% de C. En el último apartado de este capítulo vemos cómo hacer este test. El test para independencia entre las dos variables, raza y resistencia, se realiza e interpreta del mismo modo que cuando se toma una muestra de cada raza.

12.5. CÁLCULO PARA TEST CON TABLAS DE CONTINGENCIA

Para hacer el test hay que calcular en primer lugar los Valores Esperados bajo la H_0. El valor esperado de cada celda se calcula multiplicando el total de la fila por el total de la columna y dividiendo ese producto por el total de la tabla. En el ejemplo de las neumonías era:

© Todos los derechos reservados

Valores absolutos					
	Mal	Reg.	Bien	M-B	Total
Bacterias	**16**	**12**	**8**	**4**	40
Virus	**6**	**15**	**15**	**24**	60
	22	27	23	28	100

Proporciones			
Mal	Reg.	Bien	M-B
.40	**.30**	**.20**	**.10**
.10	**.25**	**.25**	**.40**

Valor esperado de la casilla Mal-Bacterias: E = 40 x 22/100 = **8.8**

Valor esperado de la casilla Mal-Vírico: E = 60 x 22/100 = **13.2**

Del mismo modo se hace para el resto de las celdas.

La tabla siguiente da todos los Valores Esperados, los totales de cada fila y columna sumando estos valores esperados. Puede verse que esos totales son los mismos que con los observados. También se ve que las FR condicionadas a la etiología en esta tabla de valores esperados son iguales en las dos causas. Esto ocurre siempre con los valores esperados. Es decir, los valores *Esperados* son el número de pacientes que tendría que haber de cada etiología y cada evolución (cada casilla de la tabla) para que, con los mismos tamaños de muestra, las dos distribuciones condicionadas fueran iguales.

Valores absolutos					
	Mal	Reg.	Bien	M-B	Total
Bacterias	**8.8**	**10.8**	**9.2**	**11.6**	40
Virus	**13.2**	**16.2**	**13.8**	**17.4**	60
Total	21	27	23	29	100

Proporciones con los valores esperados			
Mal	Reg.	Bien	M-B
.21	.27	.23	.29
.21	.27	.23	.29

Cuanto mayor es la divergencia entre las dos distribuciones muestrales, la condicionada a bacterias y la condicionada a virus, mayor es la distancia entre los valores Observados y los Esperados. Esa divergencia se resume mediante un estadístico con dos sumatorios (uno para recorrer todas las filas y el otro todas las columnas), que tendrá tantos sumando como celdas haya en la tabla de contingencia:

$$X^2_{C-1} = \sum_{1}^{2} \sum_{1}^{C} \frac{(E_i - O_i)^2}{E_i}$$

En nuestro ejemplo el sumatorio abarca a las 8 casillas de la tabla.

$X^2 = (16-8.8)^2/8.4 + (6-13.2)2/13.2 + ... = 18$

© Todos los derechos reservados

Cuando las dos distribuciones muestrales son iguales, los O_i (observados) son iguales a los E_i (esperados) y X^2 es cero. Y cuanto más divergen las dos distribuciones muestrales más se alejan los O_i de los E_i y mayor es X^2. Pequeños valores de X^2, pueden deberse al azar y no son evidencia contra la H_0. Cuanto mayor sea el valor X^2 de la muestra, más evidencia es contra la H_0.

Para encontrar el valor P habría que construir una población donde se cumple la H_0, hacer MA, es decir, sacar millones de parejas de muestras aleatorias y ver en cuántas de ellas el valor X^2 es igual o mayor que en la de nuestro estudio.

Los programas informáticos de estadística calculan el valor X^2 de cada estudio y el valor P que le corresponde. También se puede encontrar el valor P que corresponde a cada X^2 en la tabla Chi-cuadrado, al final de la asignatura. En este ejemplo se buscaría el valor P en la fila correspondiente a 3 "grados de libertad", gl.

En general, el número de gl es el número de filas menos uno multiplicado por el número de columnas menos uno. Con esas tablas no se puede calcular el valor P detalladamente, pero se puede saber si es mayor o menor que ciertas cantidades. Con 3 gl al valor $X^2 = 11.24$ le corresponde $P = 0.01$ y a $X^2 = 12.84$ le corresponde $P = 0.005$. Como a medida que aumenta X^2 disminuye P, a $X^2 = 18$ (el obtenido con nuestros datos) le corresponde $P < 0.005$. Con esta tabla no se puede ser más precisos. El valor $P = 0.0004$ que se dio inicialmente lo proporciona el paquete informático.

12.6. UNA MUESTRA CON C CATEGORÍAS

Un caso muy particular –y menos frecuente– de tablas de contingencia es el de "tablas" de dimensión 1 x C, que se generan cuando tenemos una sola muestra y sus individuos se clasifican en C categorías posibles. Consideremos una enfermedad cuya distribución estacional (por estaciones del año) poblacional en 1950 era prácticamente la misma en toda la tierra:

	Otoño	Invierno	Primavera	Verano
Proporción poblacional de casos en cada estación	**.30**	**.40**	**.10**	**.10**

Los expertos creen que debido al cambio climático se ha modificado esa distribución. Para analizar el tema se estudian muestras de N = 40 enfermos en cuatro regiones y en cada una se mira la estación en la que ocurrió cada caso. Se han anotado las distancias de cada FR de la muestra (p_i) a la frecuencia relativa de la población en 1950 (Π_i):

© Todos los derechos reservados

		Otoño	Invierno	Primavera	Verano
FR poblacional de casos en cada estación en **1950**	π_i=	**.30**	**.40**	**.10**	**.10**
América: FR en la muestra actual de N = **40**	P_i= $P_i - \pi_i$=	.30 0.00	.35 -0.05	.20 0.00	.15 +0.05
Asia: FR en la muestra actual de N = **40**	P_i= $P_i - \pi_i$=	.20 -0.10	.35 -0.05	.25 +0.05	.20 +0.10
Australia: FR en la muestra actual de N = **40**	P_i= $P_i - \pi_i$=	.20 -0.10	.25 .0.15	.35 +0.15	.20 +0.10
Europa: FR en la muestra actual de N = **40**	P_i= $P_i - \pi_i$=	.05 -0.25	.15 -0.25	.50 +0.30	.30 +0.20

- Las frecuencias observadas en América son muy próximas a las poblacionales de 1950, por lo que esa muestra no es evidencia contra la H_0 que dice que las proporciones poblacionales actuales en América son como las de 1950.
- Las frecuencias observadas en Europa son muy lejanas a las poblacionales de 1950, por lo que esa muestra es fuerte evidencia contra la H_0 que dice que en Europa las proporciones poblacionales actuales son como las anteriores.
- Las FR observadas en Asia se alejan de las poblacionales de 1950 un poco más que las americanas. ¿Son los datos de Asía compatibles con la H_0 que dice que en Asia las proporciones poblacionales actuales son como las de 1950? Unos investigadores pueden pensar que esos datos son compatibles con la H_0 y por ello no constituyen evidencia contra ella; otros pueden inclinarse a pensar que las FR muestrales se alejan demasiado de las poblacionales de 1950, por lo que esa muestra es evidencia contra la H_0. La Inferencia Estadística puede ayudar calculando el valor P del test, que es la probabilidad de que en una muestra de ese tamaño aparezcan FR como las de nuestro estudio o aún más alejadas de las de 1950 si las FR poblacionales de Asia son como las de 1950.

Para resumir en un solo número las distancias de los pi a los π_i, de las FR muestrales a las poblacionales, se suma del cuadrado de esas distancias en cada celda divididas por los respectivos valores de π_i y todo ello multiplicado por N, el tamaño de la muestra:

$$X^2 = N \cdot \sum \frac{(\pi_i - p_i)^2}{\pi_i}$$

Este estadístico, X^2, que se lee "Chi-cuadrado", es cero si los p_i coinciden exactamente con los π_i y es mayor cuanto mayor sea la divergencia entre los p_i y los π_i.

Para esta muestra es:

$$X^2 = 40\ [0.1^2/0.3 + 0.05^2/0.4 + 0.05^2/0.2 + 0.1^2/0.1] = 40 \cdot 0.1521 = \mathbf{6.1}.$$

© Todos los derechos reservados

Pequeños valores de X^2, no constituyen evidencia contra la H_0. Pero esa evidencia es mayor cuanto mayor sea el valor X^2 de la muestra. El valor P de este test indica cuan probable es que, siendo cierta la H_0, salga por azar un valor X^2 como el de nuestra muestra o aún mayor, es decir, que salgan muestras con FR observadas tan discrepantes de las esperadas como las de nuestra muestra o aún más discrepantes. En la tabla siguiente se da el valor de P para cada una de las muestras estudiadas:

	Otoño	Invier.	Primav.	Verano		
Prop. poblacional en 1950	**.30**	**.40**	**.10**	**.10**	Chi-Chuadrado	P
América: N = **40**	.30	.35	.20	.15	1.3	0.74
Asia: N = **40**	.20	.35	.25	.20	6.1	0.11
Australia: N = **40**	.20	.25	.35	.20	12.0	0.007
Europa: N = **40**	.05	.15	.50	.30	48.6	0.00000001

El valor P = **0.74** dice que en el MA de una población donde se cumple H_0, es decir, se da la distribución estacional de la enfermedad en 1950 (las π_i son 0.3, 0.4, 0.2 y 0.1), en **74** de cada **100** muestras de N = 40 salen FR tan diferentes de las poblacionales como las de esa muestra o aún más diferentes. Y el valor P = **0.11** dice que en el MA de una población donde se cumple H_0, en **11** de cada **100** muestras de N = 40 salen FR tan diferentes de las poblacionales como las de esa muestra o aún más diferentes.

El valor P lo calculan los programas estadísticos. Si no se tiene a mano uno de ellos, la tabla de Chi-Cuadrado, al final de la asignatura, da los valores de X^2 correspondientes a algunos valores de P. Para usarla debe tenerse en cuenta la cantidad llamada "grados de libertad", gl, que en estos casos es el número de celdas menos uno. Con esas tablas no se puede calcular el valor P detalladamente, pero se puede saber si es mayor o menor que ciertas cantidades.

Fíjese en la explicación que aparece al pie de la tabla y compruebe que, por ejemplo, para la muestra de Asia con 3 gl el valor X^2 = 6.1 no aparece en la tabla, pero en ella se ve en la fila de 3 gl que el valor más próximo es X^2 = 6.25 que tiene asociado un valor P = 0.10 (en la cabecera de la columna). Como a medida que aumenta X^2 disminuye P, en nuestro ejemplo el valor P sería mayor de 0.10, suficiente como para concluir que los datos observados en la muestra de Asia son compatibles con la H_0.

12.7. ASPECTOS A RECORDAR

- Estudiamos situaciones en las que los individuos de una muestra quedan clasificados según la distribución conjunta de dos variables (o en más de dos grupos según el valor de una variable cualitativa o cuantitativa agrupada).
- Las Tablas RxC se generan cuando los individuos de una muestra se clasifican conjuntamente según dos variables, una con R categorías y la otra con C categorías; o cuando se toman R muestras y en cada una los individuos se clasifican según una variable

© Todos los derechos reservados

con C categorías; o cuando se toman C muestras y en cada una los individuos se clasifican según una variable con R categorías.

- En las Tablas RxC la Hipótesis nula planteada es que las dos variables son, en población, independientes, es decir, las DF de cada variable son iguales para cada valor de la otra variable. Si en la muestra la DF de una de las variables es "parecida" en los distintos niveles de la otra no rechazaremos la H_0, la aceptaremos como posible; en caso contrario rechazaremos H_0.
- En las Tablas RxC los valores Esperados son los valores que tendría que haber en la tabla para que, con los mismos tamaños de muestra, las distribuciones condicionadas fueran iguales, es decir, para que se cumpla en la muestra lo que propone la H_0 para la población. Cuanto mayor es la divergencia entre las distribuciones muestrales mayor es la distancia entre los valores Observados y los Esperados.
- La divergencia en Tablas RxC se cuantifica mediante el estadístico llamado Chi-Cuadrado, donde los sumatorios recorren uno las filas y el otro las columnas de modo que habrá tantos sumando como casillas haya en la tabla (RxC casillas).

$$X^2_{C-1} = \sum_{1}^{R} \sum_{1}^{C} \frac{(E_i - O_i)^2}{E_i}$$

- Las Tablas 1xC se generan cuando en una muestra hacemos la Distribución de Frecuencias, DF, de una variable en la que consideramos C categorías. Planteamos como Hipótesis que las frecuencias relativas de individuos en cada categoría tienen en población ciertos valores: $\pi_1, \pi_2..., \pi_C$. y vemos si las frecuencias encontradas en la muestra: $p_1, p_2,...p_C$ son compatibles con dichos valores poblacionales.

© Todos los derechos reservados

© TODOS LOS DERECHOS RESERVADOS

Resumen

© Todos los derechos reservados

© TODOS LOS DERECHOS RESERVADOS

Regresión lineal 13

Visión general

- Criterio de mínimos cuadrados.
- Coeficiente de correlación lineal y Coeficiente de determinación.
- Regresión lineal: Muestreo - Inferencia.
- Reg. Lineal: Intervalo de Confianza y Test de Significancia.

Objetivo

- Conocer el análisis de datos con regresión lineal.

© Todos los derechos reservados

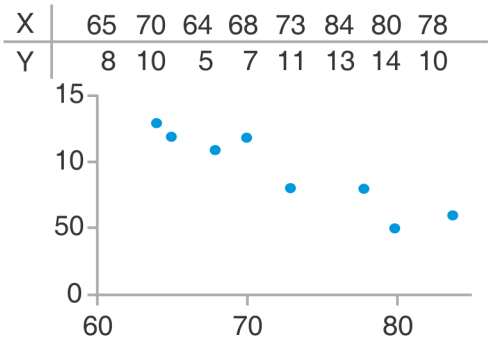
13.1. CRECIMIENTO Y DECRECIMIENTO LINEAL

En el Capítulo 3 se exponía un principio de aplicación general en el análisis de datos:

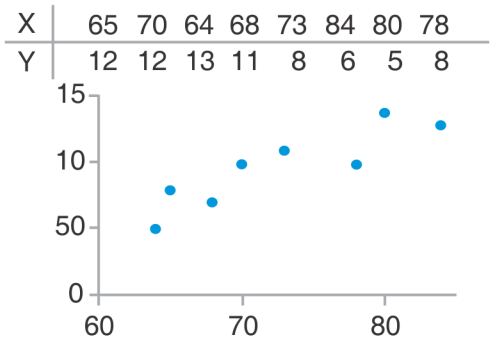
> Analizar la relación entre dos variables es estudiar el comportamiento de una de las variables (dando su media o proporción) en los distintos niveles de la otra.

Si para cada individuo de un colectivo se miden dos variables cuantitativas puede hacerse una representación gráfica de **nube de puntos** que permite una apreciación visual de la relación entre ellas. Puede que valores elevados de una se acompañen de valores elevados de la otra (correlación positiva) o que valores elevados de la primera se acompañen de valores bajos de la segunda (correlación negativa) o que haya otro tipo de asociación o ausencia de ella.

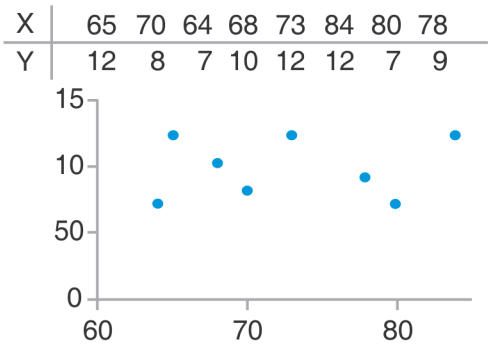
En cada uno de los ejemplos que siguen se han medido dos variables, X e Y, en 8 individuos. En el primer ejemplo, el primer individuo tiene X = 65, e Y = 8, el segundo tiene X = 70 e Y = 10, etc., y se ha representado cada par de valores (X, Y) correspondiente a cada individuo por un punto.



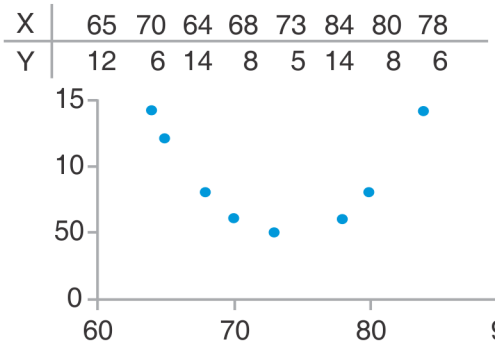
Correlación negativa, aproximadamente lineal



Correlación positiva, aproximadamente lineal



No correlación



Correlación no lineal

De especial interés, por la simplicidad de cálculo y de la interpretación biológica, son las relaciones que llamamos **lineales**, de modo que a igual incremento de X corresponde igual incremento o decremento de Y, como en los dos primeros ejemplos.

- Imaginemos una situación en la que medidas dos variables cuantitativas en 6 individuos se obtuvieran los datos de las dos primeras filas de la tabla siguiente:

X, en años	5		6		7		8		9		10
Y	17		15		13		11		9		7
Incremento de Y por año		–2		–2		–2		–2		–2	

Cuando se pasa de 5 a 6 años, el valor de Y pasa de 17 a 15 (2 unidades menos), y cuando pasamos de 6 a 7, los valores Y pasan de 15 a 13 (2 unidades menos); y lo mismo ocurre siempre que aumentamos una unidad en los valores de X. Decimos entonces que Y decrece **linealmente** con la edad a razón de **2** unidades por año. La "pendiente de la recta" o "Coeficiente de Regresión" es –2 unidades/año.

- Y si la variable Y varía con la edad según estos valores:

X, en años	5		6		7		8		9		10
Y	17		20		25		32		41		53
Incremento de Y por año		3		5		7		9		11	

© Todos los derechos reservados

Decimos que Y crece **no linealmente** con la edad, puesto que el incremento de Y por cada año no es el mismo en los distintos años.

- Si Y varía con la edad según estos otros valores, también en 6 individuos:

X, en años	5		6		7		8		9		10
Y	17		20		23		26		29		32
Incremento de Y por año		3		3		3		3		3	

Vemos que al aumentar un año, Y aumenta siempre 3 unidades. Decimos que Y crece **linealmente** con la edad a razón de **3** unidades por año. La "pendiente de la recta" o "Coeficiente de Regresión" es 3. Los valores de Y se relacionan con los de X por esta expresión: **Y = 2+3 X** que se llama ecuación de la recta y hace corresponder un valor de Y a cada valor de X. Así, para X = 5 es Y = 2+3 · 5 = 17 y para X = 6 es Y = 2+3 · 6 = 20, etc.

En general, decimos que los valores Y dependen linealmente de los valores X si la relación es **Y = a + b X**, donde '**a**' y '**b**' son dos constantes, que pueden ser mayores o menores de la unidad y positivos o negativos. Si se calcula el valor Y que corresponde a distintos valores de X con esa expresión y cada pareja de valores X e Y se representa por un punto en un sistema de ejes perpendiculares (los valores de X en el eje horizontal y los de Y en el vertical), los puntos quedan sobre una recta. La recta es ascendente hacia la derecha si es b > 0 y descendente hacia la derecha si es b < 0. La recta tiene más pendiente (más inclinada) cuanto mayor es b. La tabla que sigue da los valores de Y correspondientes a X=0, X=1, X=2 y X=3 en dos rectas.

X	0	1	2	3
Y = 3+1 X	3 (= a)	4	5	6
Y = 2+0.5 X	23 (= a)	2.5	3	3.5

Dibujando en los ejes un punto por cada uno de los pares (X, Y), se obtiene la representación gráfica de las dos rectas: para la primera recta (0,3), (1,4), (2,5) y (3,6) y para la segunda (0,2), (1,2.5), (2,3) y (3,3.5). La recta con coeficiente de regresión 1 tiene mayor pendiente que la recta con coeficiente de regresión 0.5. La constante "**a**" es el valor de Y cuando X vale 0 y se le llama ordenada en el origen.

Si X = 0 es Y = a

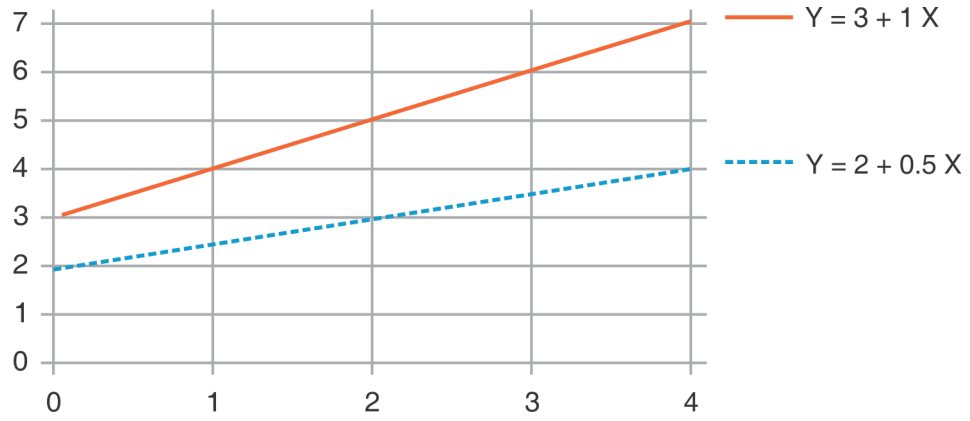
A la constante "**b**" se le llama "**Pendiente de la recta**" o "**Coeficiente de regresión**" y dice cuánto varía Y cuando X aumenta una unidad.

$Y_{x+1} - Y_x = a+b\,(x+1) - a+bx = bx+b - bx = b$

El valor de "b" depende de las unidades en que se expresen las dos magnitudes que se estén relacionando. Por ejemplo, si Y indica la distancia en kilómetros recorrida por un vehículo y X el tiempo en horas transcurrido, la relación Y = 0+120 X, nos dice que por cada hora que pasa, el camino andado aumenta 120 km. Si el tiempo se midiera en minutos, la misma relación vendría

© Todos los derechos reservados

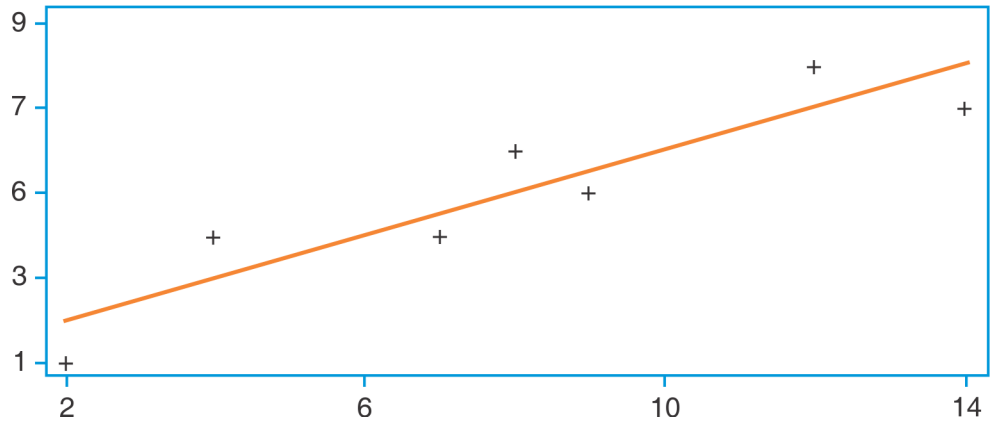
expresada por Y = 0+2 X, y si, además, la distancia se expresara en metros, la relación sería Y = 0+2.000 X. Es decir, 120 km/hora = 2 km/minuto = 2.000 m/minuto.



13.2. RECTA MÁS AJUSTADA A LOS DATOS. CRITERIO DE MÍNIMOS CUADRADOS

Solamente hay una ecuación de la forma Y = **a+b** X que relacione exactamente los valores Y con los de X si los puntos de la representación gráfica están exactamente sobre una línea recta. Cuando los puntos no quedan sobre una recta (no hay una ecuación lineal que relacione exactamente los valores Y con los de X) se busca la función lineal que a partir de los valores de X produzca valores de Y' *lo más próximos* que sea posible a los valores Y observados, es decir, una recta que pase lo más cerca posible de los puntos. Veamos como ejemplo las variables X, edad en años, e Y, fuerza en newtons en el brazo izquierdo, medida en cada uno de N = 7 niños.

X →	2	4	7	8	9	12	14	M = 8
Y →	1	4	4	6	5	8	7	M = 5



© Todos los derechos reservados

- La tabla de la izquierda a continuación contiene los valores X e Y observados. El primer niño tuvo X = 2 años e Y = 1 newton, el segundo X = 4 años e Y = 4 newtons, etc.

  Vamos a comparar tres ecuaciones lineales, es decir, tres rectas, para seleccionar entre ellas la que produzca valores calculados por dichas rectas, Y' = **a+b** X, más próximos a los valores Y observados en la muestra.

- La segunda tabla, a la derecha de los datos, contiene los cálculos para la recta **Y' = -2+1 · X**. La columna de la izquierda da los valores Y' calculados. Al primer niño con X = 2, corresponde valor calculado **Y'** = –2+1 · 2 = 0. La segunda columna da las distancias de cada Y' al Y observado, llamadas "errores de estimación", d = Y–Y'. Para el primer niño es **d** = 1–0 = 1. La tercera columna contiene los errores al cuadrado. Para el primer niño es $d^2 \equiv (Y - Y')^2 = 1^2 = 1$. La suma de todos los d^2 está en la última fila, es $\Sigma d^2 = 39$.

Datos			Y' = -2+1 · X			Y' = 1+0.5 · X			Y' = -3+1 · X		
X	Y		Y'	d = Y – Y'	$(Y - Y')^2$	Y'	d = Y – Y'	$(Y - Y')^2$	Y'	d = Y – Y'	$(Y - Y')^2$
2	1		0	+1	1	2	–1	1	-1	+2	4
4	4		2	+2	4	3	+1	1	1	+3	9
7	4		5	–1	1	4.5	–0.5	0.25	4	0	0
8	6		6	0	0	5	+1	1	5	+1	1
9	5		7	–2	4	5.5	–0.5	0.25	6	–1	1
12	8		10	–2	4	7	+1	1	9	–1	1
14	7		12	–5	25	8	–1	1	11	–4	16
		Σ		–7	39		0	5.5		0	32

- La tabla central contiene los cálculos para la recta, **Y' = 1+0.5 X** y la tabla de la derecha los cálculos para **Y' = -3+1 X**.

- En cada ecuación, para cada valor X de la tabla el valor Y' calculado indica la altura de la recta para ese valor de X y el valor Y observado indica la *altura del punto* observado, por ello d = Y–Y' es la *distancia vertical* del punto observado a la recta. Por ejemplo, el primer niño tiene X = 2 e Y = 1 y con la recta Y' = 1+0.5 · X el estimado es 2. El error de estimación es d = 1–2 = –1, lo que indica que el punto observado para X = 2 está 1 unidad por debajo de la recta. El penúltimo niño tiene X = 12, por tanto Y' = 1+0.5 · 12 = 7 y como el observado es Y = 8, el error de estimación es 8 –7 = 1. El punto observado queda 1 unidad por encima de la recta.

¿Cuál de las tres rectas produce valores calculados más próximos a los observados en la muestra, es decir, menores errores de estimación? Para unos niños el error de estimación es menor con una ecuación y para otros con otra. Por ejemplo, para el primero, con X = 2, la tercera recta produce mayor error de estimación (d = 4) que las otras dos (d = 1), pero en el tercero, con X = 4, el error de estimación es menor con la tercera recta (d = 0), y mayor (en valor absoluto) con la primera (d = –1). Es decir, ninguna recta produce mejor estimación que otra en todos y cada uno de los valores de X. Por ello buscaremos la recta que da menor "error promedio".

© Todos los derechos reservados

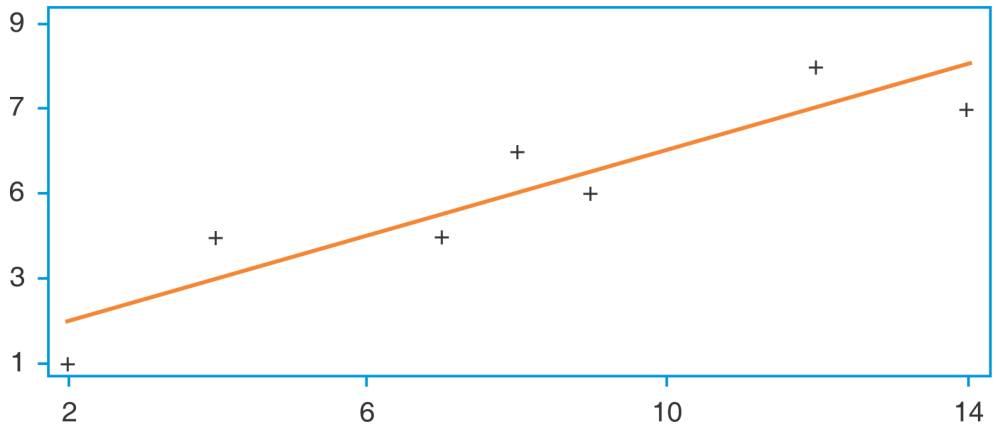
La media de los errores es la suma de ellos dividida por el número de casos. La media es menor con la recta que de menor suma. Podría elegirse la recta que diera menor suma de los errores en valor absoluto, sin signo. Pero es más útil al hacer Inferencia la recta que hace mínima la suma de los errores al cuadrado, Σd^2. Se le llama **Recta de Mínimos Cuadrados**. El criterio de Mínimos Cuadrados es muy utilizado en muchos campos del Análisis Estadístico. En nuestro ejemplo la Σd^2 es 39 para la primera recta, 5.5 para la segunda y 32 para la tercera. Por tanto, entre las tres estudiadas la mejor es **Y' = 1 + 0.5 X**.

El investigador busca la recta de mínimos cuadrados entre todas las posibles. Ello podría parecer una tarea ingente porque el número de rectas candidatas es infinito, ya que tanto "**a**" como "**b**" pueden tomar infinitos valores posibles. Por ejemplo, una recta distinta de Y' = 1+0.5 X es Y' = 0.99+0.5 X y distinta de las dos anteriores es Y' = 0.99+0.501 X y distinta de esa es Y' = 0.991+0.501 X. Variando el valor de **b** y el de **a** se van obteniendo distintas rectas y cada una de ellas podría dar menor la suma de cuadrados de error. Sin embargo, una vez más, la matemática viene al rescate y permite encontrar directamente, sin tanteos reiterados, la recta que obtiene los valores Y' más próximos a los Y, en el sentido de que se minimiza la suma de los cuadrados de los errores de estimación.

13.3. RECTA DE MÍNIMOS CUADRADOS EN LA MUESTRA

En nuestro ejemplo de los niños la media de edad es 8 años y la media de fuerza es 5 newtons.

X, en años	2	4	7	8	9	12	14	$\Sigma X = 56 \rightarrow M_X = 8$
$x = X - M_X$	–6	–4	–1	0	1	4	6	$\Sigma_X = 0$
Y, en Newtons	1	4	4	6	5	8	7	$\Sigma Y = 40 \rightarrow M_Y = 5$
y = Y–M	–4	–1	–1	1	0	3	2	$\Sigma_y = 0$



Usamos las letras "x" e "y" (minúsculas) para indicar las distancias de cada valor original de X o de Y a su media aritmética. Para el primer niño es x = 2–8 = –6, es decir, 6 años por debajo de la

© TODOS LOS DERECHOS RESERVADOS

media del grupo; y tiene una fuerza de 1, y = 1–5 = –4, es decir, 4 Newton de fuerza por debajo de la media. Ocurre que en toda lista de valores la suma de las distancias a la media es cero, $\Sigma x = 0$ y $\Sigma y = 0$.

"x" (minúscula) = $X - M_X \rightarrow \Sigma x \equiv \Sigma (X - M_X) = 0 \rightarrow S^2_X = \Sigma x^2/(N-1)$.

"y" (minúscula) = $Y - M_Y \rightarrow \Sigma y \equiv \Sigma (Y - M_Y) = 0 \rightarrow S^2_Y = \Sigma y^2/(N-1)$.

Para obtener la recta de mínimos cuadrados hay que calcular en la muestra tres cantidades:

$\Sigma(X-M_X)^2$	$= \Sigma x^2 = 106$
$\Sigma(Y-M_Y)^2$	$= \Sigma y^2 = 32$
$\Sigma(X-M_X)\,(Y-M_Y)$	$= \Sigma xy$ _ = 53

A partir de esas tres cantidades se calculan las constantes "a" y "b" y otros valores de la recta de mínimos cuadrados:

"b", Coeficiente de Regresión Muestral: b = $\Sigma xy/\Sigma x^2$ = 53/106 = **0.5**

Indica cuánto varía Y, por término medio, en la muestra, por cada unidad que aumenta X. Por cada año más que tiene el niño, su fuerza aumenta en 0.5 Newton. El valor de b depende de las unidades empleadas. Por ejemplo, es lo mismo 0.5 Newton/año que 1 newton/bienio o que 2.5 Newton/quinquenio.

"a", Ordenada en el origen: a = M_Y - b M_X = 5–0.5 · 8 = **1**.

Es el valor de Y que correspondería a un individuo con X = 0 en la recta de mínimos cuadrados. Indica el punto en que la recta de mínimos cuadrados corta al eje vertical.

Por tanto, nuestra recta de Mínimos Cuadrados es **Y' = 1+0.5 X**.

Vemos que a partir de los estadísticos Σx^2, Σy^2 y Σxy se encuentra directamente la recta buscada. El investigador no necesita recordar las fórmulas. Los programas informáticos las aplican y dan como salida los valores de "a", de "b" y otros estadísticos relacionados con la recta de mínimos cuadrados que veremos en seguida.

Tampoco necesita el investigador entender por qué esas fórmulas permiten hallar la recta de mínimos cuadrados. Pero en las lecturas complementarias se demuestra que esa solución es la correcta porque para quien no detesta las matemáticas elementales puede ser interesante ver un caso sencillo de una herramienta tan potente como es el cálculo de máximos y mínimos.

Veamos algunos estadísticos importantes relacionados con esta recta:

© Todos los derechos reservados

X	2	4	7	8	9	12	14	**Media = 8**
Y	1	4	4	6	5	8	7	**Media = 5**
Y' = 1+0.5 X	2	3	4.5	5	5.5	7	8	
d = Y – Y'	–1	+1	–0.5	+1	–.5	+1	–1	$\Sigma d = 0$
$d^2 = (Y - Y')^2$	1	1	0.25	1	0.25	1	1	Σd^2 = **5.5**

- Σd^2 = **Suma de errores al cuadrado** = $\Sigma(Y–Y')^2 = \Sigma(Y–\{1+0.5X\})^2 = 5.5$.

  Es la suma de los cuadrados de los errores de estimación, es decir, de los cuadrados de las distancias verticales de los puntos a la recta. La Σd^2 es mínima con esta recta. Se calcula directamente con:

  $\Sigma d^2 = \Sigma y^2 – (\Sigma xy)^2/\Sigma x^2 = 32 – 53^2/106 = 5.5$.

- $S^2_{y.x} = \Sigma\, d^2/(N–2)$ se llama **Varianza de la estimación**.

  La cantidad $\Sigma d^2/N$ es la media de los errores de estimación al cuadrado. Podría usarse para resumir en una sola cifra la cuantía de esos errores. Pero se usa una ligera modificación de ella, que consiste en dividir la suma Σd^2 por "N – 2", en vez de dividir por N.

- $S_{y.x} = [\Sigma d^2/(N–2)]^{½} = [5.5/5]^{½} = 1.05$. Es el **Error estándar de la regresión**.

  Indica cuánto, por término medio, nos equivocamos al estimar en la muestra los valores de Y a partir de los valores de X, utilizando la expresión de la recta.

- $S_b = S_{y.x}/(\Sigma x^2)^{½} = 1.05/106^{½} = 0.102$. Es el **Error estándar de la "b"**.

  Es la cantidad utilizada para hacer tests respecto al Coeficiente de Regresión poblacional y calcular los Intervalos de Confianza para ese parámetro. Estima la desviación estándar entre los b´s muestrales si se hiciera MA.

13.4. COEFICIENTE DE CORRELACIÓN LINEAL Y COEFICIENTE DE DETERMINACIÓN

El valor numérico de **b**, de Σy^2 y de Σd^2 depende de las unidades en que se mide la variable Y. Por ejemplo, si Y es una longitud expresada inicialmente en metros, al expresarla en centímetros el número es cien veces mayor y el valor de Σy^2 y de Σd^2 es 10.000 veces mayor, por estar elevados al cuadrado. Pero el valor del cociente $\Sigma d^2/\Sigma y^2$ es el mismo cualquiera que sea la unidad usada.

Se llama **Coeficiente de Determinación**, r^2, a la cantidad $1–\Sigma d^2/\Sigma y^2$. Puede interpretarse como la proporción de variación entre los valores de Y "explicada por" la X. En nuestro ejemplo es $r^2 = 1–\Sigma d^2/\Sigma y^2 = 1–5.5/32 = 1–0.172 = 0.828$. Con un poco de álgebra se puede ver que también es $r^2 = (\Sigma xy)^2/(\Sigma x^2 \cdot \Sigma y^2)$, es decir, $53^2/32 \cdot 106 = 0.828$.

La fuerza no es igual en todos los niños de esta muestra. Las diferencias entre los valores de Y se cuantifican por Σy^2 ("y" minúscula, es decir, $Y–M_Y$). Las diferencias entre las fuerzas se deben en

© Todos los derechos reservados

parte a que las edades difieren sensiblemente, es decir, las diferencias entre las X (las edades) "explican" buena parte de las diferencias entre las Y (las fuerzas). Pero no todas las diferencias entre las fuerzas se deben a las diferencias de edad. Niños con igual edad no tienen todos la misma fuerza, sino que hay diferencias entre ellos. La cantidad $\Sigma d^2/\Sigma y^2 = 0.172$ puede interpretarse como la proporción de la variación entre las fuerzas que **no** se debe a las diferencias de edades y la cantidad $r^2 = 1-\Sigma d^2/\Sigma y^2 = 1-0.172 = 0.828$ es la proporción de la variación entre las fuerzas que se debe a las diferencias de edades. En las lecturas complementarias se comenta un poco más extensamente este punto.

A la raíz cuadrada del Coeficiente de Determinación se le llama **Coeficiente de Correlación**, $r = \Sigma xy/(\Sigma x^2 \cdot \Sigma y^2)^{1/2}$ y se usa con más frecuencia que el de Determinación. En nuestro ejemplo es $r = 0.828^{1/2} = \mathbf{0.91}$. El Coeficiente de Correlación toma valores entre –1 y +1. Vale cero cuando los puntos no se agrupan en torno a una recta ascendente ni descendente, es decir, a mayores valores de X no corresponden valores de Y mayores ni menores, sino parecidos.

Cuanto más se aproximan los puntos a una recta, más se aproxima r a la unidad. Vale 1 cuando todos los puntos quedan exactamente sobre una recta ascendente y vale –1 cuando todos los puntos quedan exactamente sobre una recta descendente.

Resumiendo: el Coeficiente de Regresión, **b**, mide la pendiente de la recta de mínimos cuadrados, dice cuanto varía Y por cada unidad que aumenta X y su valor depende de las unidades en que se expresen X e Y. El Coeficiente de Correlación, **r**, mide la proximidad de los puntos a esa recta, toma valores entre –1 y +1 y su valor no depende de las unidades en que se expresen X e Y.

© Todos los derechos reservados

13.5. REGRESIÓN LINEAL: MODELO POBLACIONAL Y MUESTREO

En este, como en otros muchos campos del análisis de datos, el estudio de la muestra observada se hace generalmente para intentar conocer ese tipo de relaciones en la población muestreada. En la población se llama *Línea de Regresión* de la variable Y sobre X a la línea formada por las *medias de la variable* Y para cada valor de X. Es decir, hay que considerar todos los individuos con un mismo valor de X y tener en cuenta que la variable Y no tiene el mismo valor en todos ellos, sino que forman una distribución. El modelo más usado se describe por tres propiedades:

a) La media de Y para todos los individuos con cada valor concreto de X, μ, crece o decrece linealmente al aumentar los valores de X, es decir, $\mu = \alpha + \beta X$.

b) La desviación estándar entre los valores de Y para todos los individuos con el mismo valor X, $\sigma_{Y \cdot X}$ es igual para todos los valores de X, es decir

$$\sigma_{Y \cdot X1} = \sigma_{Y \cdot X2} = \sigma_{Y \cdot X3} \ldots \equiv \sigma.$$

c) Los valores Y correspondientes a cada valor X se distribuyen de modo Normal.

Muchas situaciones reales se aproximan bien a este modelo que queda especificado por los tres parámetros α, β y σ. El investigador no conoce el valor de esos parámetros y debe estimarlos a

partir de la muestra. ¿Qué estadísticos muestrales estiman esos parámetros? Si se sacan millones de muestras, todas del mismo tamaño y con los mismos valores de X, y en cada una se calcula la recta de Mínimos Cuadrados, ocurre que:

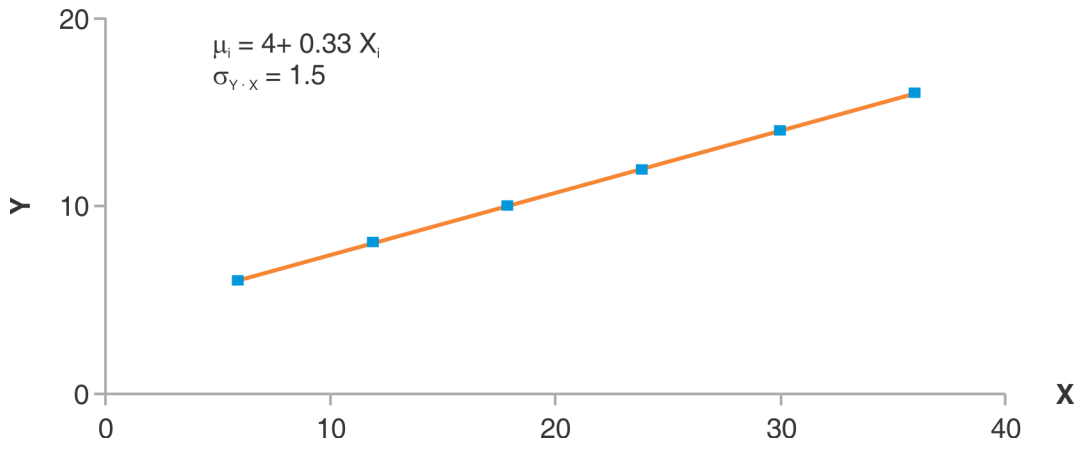
- **b** es estimador insesgado de β, es decir, la media de las b's en el MA es β.
- **a** es estimador insesgado de α, es decir, la media de las a's en el MA es α.
- $S^2_{Y\cdot X}$ es estimador insesgado de $\sigma^2_{Y\cdot X}$, la media de las $S^2_{Y\cdot X}$ en el MA es $\sigma^2_{Y\cdot X}$.

Es decir, los estadísticos a, b y $S_{Y\cdot X}$ calculados en la muestra estiman los parámetros α, β y σ desconocidos. Con esos estadísticos se calculan los IC y se hacen test para los parámetros, usando para ello la distribución de los estadísticos en el MA.

Ejemplo 13.1

Supongamos que en *población* la relación entre dos variables cuantitativas X e Y es μ = **4+0.33 X**, y que es $\sigma^2_{Y\cdot X} \equiv \sigma^2$ = **2.25**. Es decir, la recta de regresión poblacional corta al eje vertical en el valor Y = 4. Ese sería el valor medio de Y en los individuos con X = 0. Y la pendiente es 0.33, es decir, la media de Y para los individuos con cada valor de X aumenta 0.33 por cada unidad que aumenta X. Y la varianza entre las Y's de los individuos con el mismo valor de X es 2.25. $\rightarrow \sigma$ = 1.5.

Ejemplo de algunos valores μ con la a = 4 y b = 0.33				
X	0	1	2	3
$\mu \equiv \mu_{Y\|.X=X}$	4	4.33	4.66	4.99



Si de esa población tomamos una muestra de N = 6 individuos con valores de X: 6, 12, 18, 24, 30 y 36, para el individuo con X = 12, por ejemplo, no sabemos el valor Y que va a salir. Saldrá una cantidad al azar, dentro de las que forman la distribución de todos los Y que corresponden a X = 12 y que tienen media **4 + 0.33 · 12 = 8**. Si hacemos MA tomando en cada muestra siempre 6 individuos con esos mismos valores X, la variable Y tomará en cada

© Todos los derechos reservados

individuo un valor más o menos cercano, pero generalmente no exactamente igual, a **4 + 0.33 X**. A cada muestra corresponde una recta muestral de Mínimos Cuadrados cuya pendiente y ordenada en el origen tendrán valores próximos a los de la recta de regresión poblacional. La tabla que sigue reproduce algunas de las muestras obtenidas en el MA y los estadísticos más relevantes de cada una.

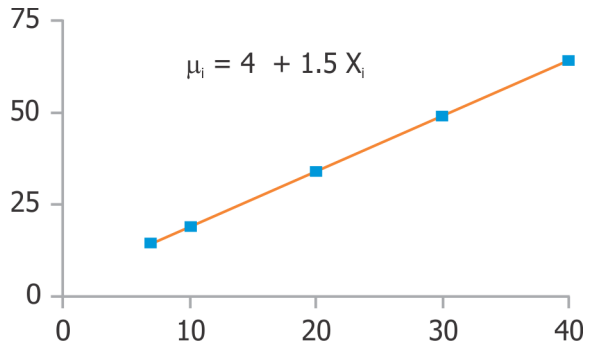
Población		→	Muestra 1		...	Muestra 10		...	Muestra 1000		...	Muestra 10000	
X	μ		X	Y		X	Y		X	Y		X	Y
6	6		6	7.2		6	5.3		6	5.1		6	6.6
12	8		12	9.8		12	8.4		12	7.0		12	7.6
18	10		18	8.3		18	9.6		18	8.3		18	11.0
24	12		24	10.5		24	10.9		24	12.4		24	12.5
30	14		30	13.9		30	13.2		30	16.3		30	11.3
36	16		36	15.1		36	17.1		36	15.1		36	18.0
β estimada por			b:	0.28		0.59			0.49	0.15		→	0.33
α estimada por			a:	5.7		3.4			2.8	4.2		→	4
$\sigma^2_{Y \cdot X}$ estim.			$S^2_{Y \cdot X}$:	2.9		1.3			4.1	2.1		→	2.25

© Todos los derechos reservados

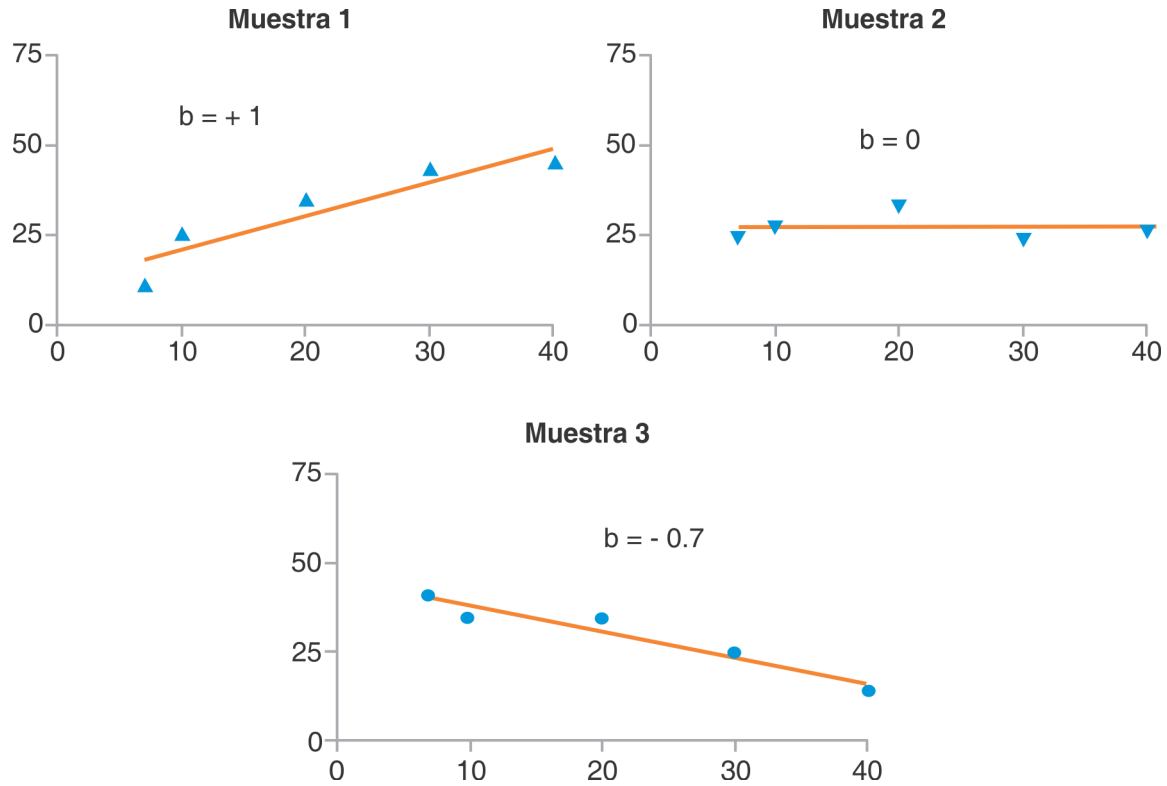
13.6. INFERENCIA EN LA REGRESIÓN LINEAL

Si de una población con, por ejemplo, $\beta = 1.5$ se toma una muestra, la b muestral puede tomar valores como +1 o 0 o –0.7, como se muestra en las figuras:

Modelo poblacional →



El modelo dice que en la población, por cada unidad de aumento de la X, las medias de Y aumentan 1.5 unidades. Si tomamos muestras de esta población, en ellas no ocurrirá exactamente lo mismo. Pueden ser muestras como las que siguen:



Pero en la práctica no sabemos lo que ocurre realmente en la población, solo tenemos los datos de una muestra, a partir de la cual intentaremos saber cómo es la relación entre las dos variables en la población. Aunque realmente no haya relación lineal entre dos variables en la población, en la muestra siempre aparecerá, por la VRM, un cierto grado de asociación positiva o negativa. Para ver hasta qué punto la asociación encontrada en la muestra podría ser por la VRM se hace test estadístico.

Como ejemplo haremos test e intervalo de confianza en el ejemplo de la fuerza medida en el brazo, Y, y la edad en años, X, en cada uno de N = 7 niños:

X, en años →	2	4	7	8	9	12	14	M = 8
Y, en Newtons →	1	4	4	6	5	8	7	M = 5

Con estos datos se calculó la recta de mínimos cuadrados, **Y' = 1+0.5 X** con coeficiente de regresión b = 0.5, error estándar de la regresión $S_{y.x}$ = 1.05 y error estándar de la b S_b = 0.102.

Intervalo de Confianza para β

$$IC_{(\beta)} = b \pm S_b \, t_{N-2}$$

Donde Sb es el error estándar de la b calculada con los datos de la muestra. Si se quiere confianza 95% el valor "t_{N-2}" es el valor de la distribución t de Student con N–2 grados de libertad, gl, que deja fuera de él, el 2.5% de los valores de esa distribución en cada cola (5% entre las dos colas).

© Todos los derechos reservados

En nuestro ejemplo con N =7, utilizaremos el valor que aparece donde se cruzan la fila de 5 gl y la columna con cabecera BIL = 0.05, que es t = 2.571. Si se quiere confianza 99% se toma el valor que deja fuera de él el 0.5% de los valores de esa distribución en cada cola (1% entre las dos colas), (para este ejemplo sería t = 4.032). s_b es la estimación que se hace con esta muestra de la desviación estándar entre las b's muestrales si se hiciera MA.

$IC_{95\%} = 0.5 \pm 0.102 \times 2.571 = 0.5 \pm 0.26 =$ **0.24 y 0.76**.

En la muestra es b = 0.5, es decir, por cada año de más, la fuerza aumenta en promedio en 0.5 Newtons. No sabemos cuánto es ese aumento en la población, pero tenemos 95% de confianza en que se encuentre entre 0.24 y 0.76 Newtons.

Tests de significación para β:

a) La Hipótesis Nula, H_0, habitualmente planteada es que en población es **β = 0**, es decir, no hay regresión lineal, al aumentar la X la media de Y no crece ni disminuye linealmente. Esta H_0 implica también que en población es **r = 0**.

El estadístico a calcular es: $t_{N-2} = b/S_b$

En nuestro ejemplo la hipótesis planteada dice que en población no hay variación lineal de la fuerza con el aumento de la edad. Encontramos t5 = 0.5 / 0.102 = 4.9 → mirando en la tabla "t-Student" con grados de libertad, gl = N–2 = 5, encontramos que el valor más próximo a él es 4.032, al que le corresponde P_{UNIL} = 0.005. Al valor t = 4.9 obtenido en la muestra le corresponderá **P_{UNI} < 0.005**. Es decir, si en población no hay regresión y repitiéramos este estudio con 7 niños millones de veces, en menos de 5 estudios cada 1.000 aparecería b = 0.50 o mayor. Los datos de la muestra no son fácilmente compatibles con la hipótesis H_0, por lo que tendemos a rechazar H_0 y pensar que en población sí hay relación lineal entre la edad y la fuerza.

b) También puede plantearse como H_0 que β tiene cierto valor, es decir, que en población al aumentar la X la media de Y crece o disminuye cierta cantidad β.

El estadístico a calcular es: $tN-2 = |\beta - b| / Sb$

En el ejemplo, para **H_0: β = 0.7** es t_5 = |0.7–0.5|/0.102 = **1.96**. En la tabla "t de Student" se encuentra que con gl = 5 al valor t = 2.015 le corresponde P_{UNI} = 0.05, por tanto al valor t = 1.96 (que es menor que 2.015) le corresponderá P_{UNI} > 0.05.

Cuando los cálculos del test los hace un programa informático la salida incluye el valor P, que en este caso es **P_{UNI} = 0.053**. Es decir, si en población es **β = 0.7** y repitiéramos este estudio millones de veces, en 53 de cada 1.000 veces aparecería por la VRM valor **b** igual o menor que 0.5. Es decir, no es muy raro encontrar unos datos con un coeficiente de regresión como el obtenido (b = 0.5) si en la población fuera de 0.7 (β = 0.7), por lo que decimos que nuestros datos son compatibles con la hipótesis planteada para la población; no rechazamos H_0 admitiendo el valor β = 0.7 como un valor posible para la población.

© Todos los derechos reservados

El numerador del estadístico calculado en el test mide la diferencia entre la pendiente de la recta que propone la H_0 para la población y la encontrada en la muestra.

Si, por ejemplo, se propone como H0 que en población Y disminuye 0.3 unidades por cada unidad que aumenta X, es decir, β **= –0.3,** (recuerde que en la muestra aumenta 0.5 unidades, b = 0.5) es t_5 = |-0.3–0.5|/0.102 = 0.8/0.102 = 7.84 y P_{UNIL} **< 0.0005**.

Rechazamos H_0, β no es –0.3, será β > –0.3.

13.7. ASPECTOS A RECORDAR

- El estudio de la relación entre dos variables se hace de distinta forma dependiendo del tipo de variables implicadas:
  - Si las dos son cualitativas, una con C categorías y otra con K categorías, se forma una tabla de contingencia C x K. La situación es la misma si una de las variables es cuantitativa y sus valores se agrupan en intervalos o si las dos variables son cuantitativas y los datos se agrupan en intervalos en ambas. Se hace Inferencia utilizando el test de chi-cuadrado.
  - Si una es cuantitativa y la otra cualitativa, se calcula la media y desviación de la cuantitativa para cada categoría de la cualitativa y se hace ANOVA.
  - Si las dos son cuantitativas (con valores no agrupados en intervalos) indagamos la posible Regresión Lineal, RL: **Y = a+b X** a la que gráficamente corresponde una línea recta.
- Los datos X e Y medidos en cada individuo de una muestra se representan en una gráfica llamada nube de puntos que asocia un punto a cada individuo, representando su valor X en el eje horizontal y su valor Y en el vertical.
- Se busca la recta que mejor se ajusta a la nube de puntos. Tiene por expresión Y = a+b · X donde la **ordenada en el origen**, "a", indica el valor de Y que corresponde a X = 0, es el punto en que la recta corta al eje vertical. El **coeficiente de regresión** o **pendiente** de la recta, "b", indica cuando varía Y por cada unidad que aumenta X. El valor de b depende de las unidades en que se midan las variables.
- Calculando los coeficientes 'b' y 'a' de la recta con estas expresiones: **b** = $\Sigma_{x \cdot y} / \Sigma_x^2$ y **a** = $M_Y - b \cdot M_X$ se hace mínima la suma de los cuadrados de los errores de estimación, $\Sigma \mathbf{d}^2 = \Sigma(Y-\{a+bX\})^2$.
- El **Coeficiente de Correlación** mide la cercanía de los puntos observados a la recta de regresión de Mínimos Cuadrados. Vale **cero** cuando los puntos no se agrupan en torno a una recta. Cuanto más se aproximan los puntos a una recta, más se aproxima a **+1** (recta ascendente) o –1 (descendente).
- El **error estándar de la regresión**, $S_{y.x} = [\Sigma d^2/(N-2)]^{1/2}$, dice cuánto, por término medio, nos equivocamos al estimar en la muestra los valores de Y a partir de los valores de X. Estima la $\sigma^2_{Y.X} \equiv \sigma^2$ poblacional.
- **El error estándar de la b, S_b** = $S_{y.x}/(\Sigma_x^2)^{1/2}$ estima la desviación estándar entre los b´s muestrales si se hiciera MA. Se utiliza para hacer tests e IC para β.

© Todos los derechos reservados

Universidad
Europea
del Atlántico

Resumen

© Todos los derechos reservados

© Todos los derechos reservados

Bibliografía

Bibliografía actual

[1] Arpajón Peña, Y., & Sosa Pérez, A. L. (2015). Procedimientos para la estimación por intervalos de confianza en las investigaciones biomédicas. *Revista Cubana De Angiología Y Cirugía Vascular, 16*(1), 92-100.

[2] Atanasio, J. M. P. (2015). Principios de bioestadística para comprender los resultados reportados en los artículos científicos. *Principios, 11*(2).

[3] Ayçaguer, L. C. S. (2014). Manejo de la temporalidad en estudios observacionales para establecer relaciones causales. Temas y controversias en bioestadística Medwave 2014;14(4).

[4] Chiavegatto Filho, A. D. P., Wang, Y. P., Malik, A. M., Takaoka, J., Viana, M. C., & Andrade, L. H. (2015). Determinantes do uso de serviços de saúde: análise multinível da Região Metropolitana de São Paulo. *Revista de Saúde Pública, 49*, 1-12.

[5] de Oliveira, A. P., Malhado, C. H. M., Barbosa, L. T., Martins Filho, R., & Carneiro, P. L. S. (2015). Inferência bayesiana na avaliação genética de bovinos da raça tabapuã do nordeste brasileiro. *Revista Caatinga, 28*(4), 227-234.

[6] Espírito-Santo, H., Garcia, I. Q., Monteiro, B., Carolino, N., & Daniel, F. (2016). Avaliação breve do défice executivo em pessoas idosas com Acidente Vascular Cerebral: Validação da Bateria de Avaliação Frontal.

[7] Figueiró, R., Vargas, A. B., & Vieira, V. (2013). A Bioestatística em contexto para o estudante de graduação: relato de experiência de uma prática interdisciplinar para o curso de Ciências Biológicas. *Revista Práxis, 5*(10).

[8] Iglesias, J. C. A., & comunitario en Vigo, F. (2012). Introducción a la bioestadística. *Farmacéuticos comunitarios, 4*(1), 25-30.

[9] Lopes, B., Ramos, I. C. D. O., Ribeiro, G., Correa, R., Valbon, B. D. F., Luz, A. C. D., & Ambrósio Junior, R. (2014). Biostatistics: fundamental concepts and practical applications. *Revista Brasileira de Oftalmologia, 73*(1), 16-22.

[10] Méndez Martínez, C., & Rondón Sepúlveda, M. A. (2012). Introducción al análisis factorial exploratorio. *Revista Colombiana De Psiquiatria, 41*(1), 197-207.

[11] Navarro, F., & González de Dios, J. (2014). Palabras y expresiones inglesas de traducción difícil o engañosa en investigación clínica, bioestadística y medicina basada en la evidencia. *Emergencias, 26*(5), 375-392.

[12] Navarro, S. R. M., & Tobar, J. A. R. (2015). Uso de modelos de regresión para la determinación de factores de riesgo. Temas y controversias en bioestadística Medwave 2015 Jun;15(5).

© Todos los derechos reservados

[13] Oliveira, F. L. D. (2013). Estudo das técnicas inferenciais de bioestatística com aplicação a dados de natureza biológica.

[14] Reissner, C. V. D., & Martínez, G. I. R. (2015). Fundamentos para la aplicación de Bioestadística en Odontología. *Revista de Salud Pública del Paraguay,5*(1), 33-39.

[15] Silva, S. W. S. D. (2014). Aplicação dos modelos lineares generalizados com distribuição binomial para dados binários.

Referencias Bibliográficas

[1] Suárez, E., Pérez, C. M., & Guzmán, J. (2000). Métodos Alternos de Análisis Estadístico en Epidemiología: Frecuentista, Verosimilitud y Bayesiano. *Puerto Rico Health Sciences Journal, 19*(2).

[2] Díaz, F. R., Lopez, F. J. B., Font, E. S., & Guijosa, L. P. (2010). Bioestadística: métodos y aplicaciones.

[3] Rubiales, A. S., del Valle, M. L., Vecino, A., & Flores, L. A. (2003). Bioestadística para clínicos. Más allá de "p< 0, 05". *Medicina paliativa, 10*(4), 208-214.

[4] Calvache, J. A., Barón, F. J., & Shoemaker, R. G. La bioestadística y su aplicación a la investigación en salud.

[5] Sánchez, A. C., García, S. H., & Aleaga, M. E. (2009). Herramientas básicas de la estadística matemática para un investigador. *Innovación Tecnológica, 15*(2), 1-10.

© Todos los derechos reservados

Universidad Europea del Atlántico

Anexo

Tablas Estadísticas

Distribución Binomial

$$P(X=x)=\binom{N}{x}\cdot\pi^{x}\cdot(1-\pi)^{N-x}$$

© Todos los derechos reservados

N	X	π	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
2	0		.9801	.9025	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1		.0198	.0950	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2		.0001	.0025	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0		.9703	.8574	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1		.0294	.1354	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2		.0003	.0071	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3		.0000	.0001	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0		.9606	.8145	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1		.0388	.1715	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2		.0006	.0135	.0486	.0975	.1636	.2109	.2646	.3105	.3456	.3675	.3750
	3		.0000	.0005	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4		.0000	.0000	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0		.9510	.7738	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0312
	1		.0480	.2036	.3280	.3915	.4096	.3855	.3602	.3124	.2592	.2059	.1562
	2		.0010	.0214	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3		.0000	.0011	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4		.0000	.0000	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5		.0000	.0000	.0000	.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0		.9415	.7351	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1		.0571	.2321	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2		.0014	.0305	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3		.0000	.0021	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4		.0000	.0001	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5		.0000	.0000	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0938
	6		.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0018	.0041	.0083	.0156

N	X	π	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
7	0		.9321	.6983	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1		.0659	.2573	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0574
	2		.0020	.0406	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3		.0000	.0036	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4		.0000	.0002	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5		.0000	.0000	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6		.0000	.0000	.0000	.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7		.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0016	.0037	.0078
8	0		.9227	.6634	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1		.0746	.2793	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0312
	2		.0026	.0515	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3		.0001	.0054	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188
	4		.0000	.0004	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5		.0000	.0000	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6		.0000	.0000	.0000	.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7		.0000	.0000	.0000	.0000	.0001	.0004	.0012	.0033	.0079	.0164	.0312
	8		.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0007	.0017	.0039
9	0		.9135	.6302	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1		.0830	.2985	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2		.0034	.0629	.1722	.2597	.3020	.3003	2688	.2162	.1612	.1110	.0703
	3		.0001	.0077	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641
	4		.0000	.0006	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461
	5		.0000	.0000	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6		.0000	.0000	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7		.0000	.0000	.0000	.0000	.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8		.0000	.0000	.0000	.0000	.0000	.0001	.0004	.0013	.0035	.0083	.0176
	9		.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0008	.0020
10	0		.9044	.5987	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1		.0914	.3151	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2		.0042	.0746	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3		.0001	.0105	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172
	4		.0000	.0010	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
	5		.0000	.0001	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6		.0000	.0000	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7		.0000	.0000	.0000	.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8		.0000	.0000	.0000	.0000	.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9		.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0005	.0016	.0042	.0098
	10		.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0010

Para cada valor de n (primera columna) y de π (primera fila), el interior de la tabla da la probabilidad de que haya X individuos "positivos" en una muestra de tamaño n tomada de una población con proporción de positivos igual a π. Ejemplo: si de una población con 20% de enfermos (π = 0.2) se toma una muestra de n = 8, la probabilidad de que en ella haya 3 enfermos es 0.1468 y P (7 enfermos) = 0.0001.

© Todos los derechos reservados

Universidad Europea del Atlántico

Distribución de Poisson

$$P(X=x)=e^{-\lambda}\frac{\lambda^x}{x!}$$

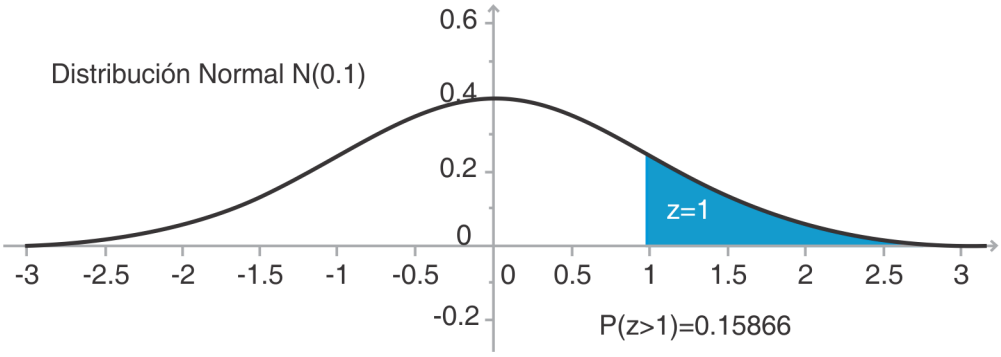
$\lambda\downarrow$ $X\rightarrow$	0	1	2	3	4	5	6	7	8	9	10	11
0.1	.9048	.0905	.0045	.0002	.0000	.0000	.0000					
0.2	.8187	.1637	.0164	.0011	.0001	.0000	.0000					
0.3	.7408	.2222	.0333	.0033	.0002	.0001						
0.4	.6703	.2681	.0536	.0072	.0007	.0002						
0.5	.6065	.3033	.0758	.0126	.00016							
0.6	.5488	.3293	.0988	.0198	.0030	.0004	.0000	.0000	.0000			
0.7	.4966	.3476	.1217	.0284	.0050	.0007	.0001	.0000				
.08	.4493	.3595	.1438	.0383	.0077	.0012	.0002	.0000				
.09	.4066	.3659	.1647	.0494	.0111	.0020	.0003	.0001				
1.0	.3679	.3679	.1839	.0613	.0153	.0030	.0005					
1.1	.3329	.3662	.2014	.0738	.0203	.0045	.0008	.0001	.0000	.0000		
1.2	.3012	.3614	.2169	.0867	.0260	.0062	.0012	.0002	.0000	.0000		
1.3	.2725	.3543	.2303	.0998	.0324	.0084	.0018	.0003	.0001	.0000		
1.4	.2466	.3452	.2417	.1128	.0395	.0111	.0026	.0005	.0001			
1.5	.2231	.3347	.2510	.1255	.0471	.0141	.0035	.0008	.0001			
1.6	.2019	.3230	.2584	.1378	.0551	.0176	.0047	.0011	.0002	.0000	.0000	
1.7	.1827	.3106	.2640	.1496	.0636	.0216	.0061	.0015	.0003	.0001	.0000	
1.8	.1653	.2975	.2678	.1607	.0723	.0260	.0078	.0020	.0005	.0001	.0000	
1.9	.1496	.2842	.2700	.1710	.0812	.0309	.0098	.0027	.0006	.0001	.0000	
2.0	.1353	.2707	.2707	.1804	.0902	.0361	.0120	.0034	.0009	.0002		
2.2	.1108	.2438	.2681	.1966	.1082	.0476	.0174	.0055	.0015	.0004	.0001	.0000
2.4	.0907	.2177	.2613	.2090	.1254	.0602	.0241	.0083	.0025	.0007	.0002	.0000
2.6	.0743	.1931	.2510	.2176	.1414	.0735	.0319	.0118	.0038	.0011	.0003	.0001
2.8	.0608	.1703	.2384	.2225	.1557	.0872	.0407	.0163	.0057	.0018	.0005	.0001
3.0	.0498	.1494	.2240	.2240	.1680	.1008	.0504	.0216	.0081	.0027	.0008	.0002
3.2	.0408	.1304	.2087	.2226	.1781	.1140	.0608	.0278	.0111	.0040	.0013	.0004
3.4	.0334	.1135	.1929	.2186	.1858	.1264	.0716	.0348	.0148	.0056	.0019	.0006
3.6	.0273	.0984	.1771	.2125	.1912	.1377	.0826	.0425	.0191	.0076	.0028	.0009
3.8	.0224	.0850	.1615	.2046	.1944	.1477	.0936	.0508	.0241	.0102	.0039	.0013
4.0	.0183	.0733	.1465	.1954	.1954	.1563	.1042	.0595	.0298	.0132	.0053	.0019
5.0	.0067	.0337	.0842	.1404	.1755	.1755	.1462	.1044	.0653	.0363	.0181	.0082
6.0	.0025	.0149	.0446	.0892	.1339	.1606	.1606	.1377	.1033	.0688	.0413	.0225
7.0	.0009	.0064	.0223	.0521	.0912	.1277	.1490	.1490	.1304	.1014	.0710	.0452
8.0	.0003	.0027	.0107	.0286	.0573	.0916	.1221	.1396	.1396	.1241	.0993	.0722
9.0	.0001	.0011	.0050	.0150	.0337	.0607	.0911	.1171	.1318	.1318	.1186	.0970
10.0	.0000	.0005	.0023	.0076	.0189	.0378	.0631	.0901	.1126	.1251	.1251	.1137

© Todos los derechos reservados

λ↓ X→	12	13	14	15	16	17	18	19	20	21	22	23
3.2	.0001	.0000										
3.4	.0002	.0000										
3.6	.0003	.0000										
3.8	.0004	.0001										
4.0	.0006	.0002										
5.0	.0034	.0013	.0005	.0002	.0003	.0001	.0002	.0001	.0002	.0001	.0001	.0002
6.0	.0113	.0052	.0022	.0009	.0014	.0006	.0009	.0004	.0006	.0003	.0004	
7.0	.0264	.0142	.0071	.0033	.0045	.0021	.0029	.0014	.0019	.0009		
8.0	.0481	.0296	.0169	.0090	.0109	.0058	.0071	.0037				
9.0	.0728	.0504	.0324	.0194	.0217	.0128						
10.0	.0948	.0729	.0521	.0347								

Para cada valor λ (primera columna), el interior de la tabla da la probabilidad del valor "x" (primera fila) en una Distribución de Poisson de parámetro λ, P(X=x). Ejemplo, si de una población con 0.3% (π = **0.003**) de enfermos se toma una muestra de n = 2 000 el valor esperado es $\lambda = 0.003 \cdot 2\,000 =$ **0.1606**, es decir, en aproximadamente 16 de cada 100 muestras habrá 6 enfermos. Y la probabilidad de que en ella haya justamente 2 enfermos es $P(X = 2 | \lambda = 6) =$ **0.0446**. Y la probabilidad de que haya más de 13 enfermos es $P(X \geq 14 | \lambda = 6) = 0.0035$. (Los valores en blanco indican que la probabilidad es menor de 0.0001).

Distribución Normal



Z	Área más allá de Z	Z	Área más allá de Z	Z	Área más allá de Z	Z	Área más allá de Z
0.00	.5000	0.80	.2119	1.60	.0548	2.40	.0082
0.02	.4920	0.82	.2061	1.62	.0526	2.42	.0078
0.04	.4840	0.84	.2061	1.64	.0505	2.44	.0073
0.06	.4761	0.86	.1949	1.66	.0485	2.46	.0069
0.08	.4681	0.88	.1894	1.68	.0465	2.48	.0066

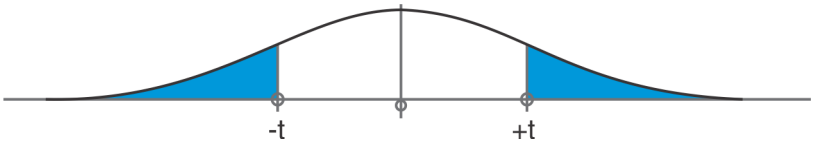
© Todos los derechos reservados

Z	Área más allá de Z	Z	Área más allá de Z	Z	Área más allá de Z	Z	Área más allá de Z
0.10	.4602	0.90	.1841	1.70	.0446	2.50	.0062
0.12	.4522	0.92	.1788	1.72	.0427	2.52	.0059
0.14	.4443	0.94	.1736	1.74	.0409	2.54	.0055
0.16	.4364	0.96	.1685	1.76	.0392	2.56	.0052
0.18	.4286	0.98	.1635	1.78	.0375	2.58	.0049
0.20	.4207	1.00	.1587	1.80	.0359	2.60	.0047
0.22	.4129	1.02	.1539	1.82	.0344	2.62	.0044
0.24	.4052	1.04	.1492	1.84	.0329	2.64	.0041
0.26	.3974	1.06	.1446	1.86	.0314	2.66	.0039
0.28	.3897	1.08	.1401	1.88	.0301	2.68	.0037
0.30	.3821	1.10	.1357	1.90	.0287	2.70	.0035
0.32	.3745	1.12	.1314	1.92	.0274	2.72	.0033
0.34	.3669	1.14	.1271	1.94	.0262	2.74	.0031
0.36	.3594	1.16	.1230	1.96	.0250	2.76	.0029
0.38	.3520	1.18	.1190	1.98	.0939	2.78	.0027
0.40	.3446	1.20	.1151	2.00	.0228	2.80	.0026
0.42	.3372	1.22	.1112	2.02	.0217	2.82	.0024
0.44	.3300	1.24	.1075	2.04	.0207	2.84	.0023
0.46	.3228	1.26	.1038	2.06	.0197	2.86	.0021
0.48	.3156	1.28	.1003	2.08	.0188	2.88	.0020
0.50	.3085	1.30	.0968	2.10	.0179	2.90	.0019
0.52	.3015	1.32	.0934	2.12	.0170	2.92	.0018
0.54	.2946	1.34	.0901	2.14	.0162	2.94	.0016
0.56	.2877	1.36	.0869	2.16	.0154	2.96	.0015
0.58	.2810	1.38	.0838	2.18	.0146	2.98	.0014
0.60	.2743	1.40	.0808	2.20	.0139	3.00	.0013
0.62	.2676	1.42	.0778	2.22	.0132	3.10	.0010
0.64	.2611	1.44	.0749	2.24	.0125	3.20	.0007
0.66	.2546	1.46	.0721	2.26	.0119	3.30	.0005
0.68	.2483	1.48	.0694	2.28	.0113	3.40	.0003
0.70	.2420	1.50	.0668	2.30	.0107	3.50	.0002
0.72	.2358	1.52	.0643	2.32	.0102	3.70	.0001
0.74	.2296	1.54	.0618	2.34	.0096	3.90	.00005
0.76	.2236	1.56	.0594	2.36	.0091	4.00	.00003
0.78	.2177	1.58	.0571	2.38	.0087	5.00	$3\text{-}10^{-7}$
						6.00	$1\text{-}10^{-11}$

Para cada uno de los valores Z (1ª columna), la 2ª columna da la probabilidad de un valor igual o superior a él en una Distribución Normal Estandarizada, N(0,1). Es el área sombreada en el gráfico. Ejemplo: en una Normal con Media μ = 200 y Desviación estándar σ = 10, la proporción de casos con valor mayor de 215 es P (X > 215) = P (Z >1.5) = 0.0668 y la proporción de casos con X mayor de 190 es P (X > 190) = P (Z > –1) = 0.5 + 0.1587 = 0.6587. En una Normal con Media μ = 80 y Desviación estándar σ = 5, la proporción de casos con valor menor de 65 es P (X < 65) = P (Z < -3) = P (Z > 3) = 0.0013. Y la proporción de casos con valor entre 70 y 85 es P (70 < X < 85) = P (–2 < Z < 1) = (0.4772 + 0.3413) = 0.8185.

© Todos los derechos reservados

Distribución t de Student



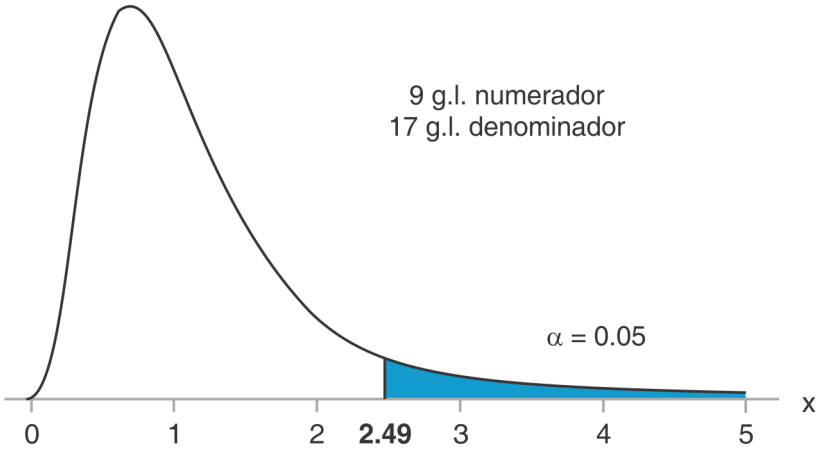
BIL →	0.30	0.20	0.10	0.05	0.02	0.01	0.001
UNI →	0.15	0.10	0.05	0.025	0.010	0.005	0.0005
gl ↓ 1	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	1.250	1.636	2.353	3.182	4.541	5.841	12.941
4	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	1.083	1.356	1.782	2.179	2.618	3.055	4.318
13	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	1.050	1.303	1.684	2.021	2.423	2.704	3.551
50	1.047	1.229	1.676	2.009	2.403	2.678	3.497
60	1.046	1.296	1.671	2.000	2.390	2.660	3.460
80	1.043	1.292	1.664	1.990	2.374	2.639	3.417
100	1.042	1.290	1.660	1.984	2.364	2.626	3.391
∞	1.036	1.282	1.645	1.960	2.326	2.576	3.291

Cada fila da los datos para la Distribución t-Student con los grados de libertad, gl, indicados en la primera columna. Todas las distribuciones t-Student tienen media cero y son simétricas.

© Todos los derechos reservados

- La distribución con, por ejemplo, gl = 2, tiene el 0.15 de los valores por encima de 1.386 y otro 0.15 por debajo de –1.386, de modo que deja 30% en la dos colas.
- La distribución con, por ejemplo, gl = 15, tiene el 0.005 de los valores por encima de 2.947 y otro 0.005% por debajo de –2.947, de modo que deja 0.01 en la dos colas.
- La distribución con, por ejemplo, gl = 40, tiene el 0.01 de los valores por encima de 2.423 y otro 0.01 por debajo de –2.423, de modo que deja 0.02 en la dos colas.
- Si en un test se obtiene, por ejemplo, t-Student = 3.3 con 16 gl, vemos que 3,3 está entre 2.921 y 3.819, por lo que es $P_{UNIL} < 0.005$.
- Si en un test se obtiene, por ejemplo, t-Student = 1.2 con 29 gl, vemos que 1.2 está entre 1.055 y 1.311, por lo que es $P_{UNIL} > 0.10$.

Distribución F de Snedecor



alfa = 0.05 **alfa = 0.01**	**Grados de libertad del numerador**													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**20**	**60**	**100**	**10000**
1	161,45 4052,2	199,50 4999,3	215,71 5403,5	224,58 5624,3	230,16 5764,0	233,99 5859,0	236,77 5928,3	238,88 5981,0	240,54 6022,4	241,88 6055,9	248,02 6208,7	252,20 6313,0	253,04 6333,9	254,30 6365,6
2	18,51 98,50	19,00 99,00	19,16 99,16	19,25 99,25	19,30 99,30	19,33 99,33	19,35 99,36	19,37 99,38	19,38 99,39	19,40 99,40	19,45 99,45	19,48 99,48	19,49 99,49	19,50 99,50
3	10,13 34,12	9,55 30,82	9,28 29,46	9,12 28,17	9,01 28,24	8,94 27,91	8,89 27,67	8,85 27,49	8,81 27,34	8,79 27,23	8,66 26,69	8,57 26,32	8,55 26,24	8,53 26,13
4	7,71 21,20	6,94 18,00	6,59 16,69	6,39 15,98	6,26 15,52	6,16 15,21	6,09 14,98	6,04 14,80	6,00 14,66	5,96 14,55	5,80 14,02	5,69 13,65	5,66 13,58	5,63 13,46
5	6,61 16,26	5,79 13,27	5,41 12,06	5,19 11,39	5,05 10,97	4,95 10,67	4,88 10,46	4,82 10,29	4,77 10,16	4,74 10,05	4,56 9,55	4,43 9,20	4,41 9,13	4,37 9,02
6	5,99 13,75	5,14 10,92	4,76 9,78	4,53 9,15	4,39 8,75	4,28 8,47	4,21 8,26	4,15 8,10	4,10 7,98	4,06 7,87	3,87 7,40	3,74 7,06	3,71 6,99	3,67 6,88
7	5,59 12,25	4,74 9,55	4,35 8,45	4,12 7,85	3,97 7,46	3,87 7,19	3,79 6,99	3,73 6,84	3,68 6,72	3,64 6,62	3,44 6,16	3,30 5,82	3,27 5,75	3,23 5,65
8	5,32 11,26	4,46 8,65	4,07 7,59	3,84 7,01	3,69 6,63	3,58 6,37	3,50 6,18	3,44 6,03	3,39 5,91	3,35 5,81	3,15 5,36	3,01 5,03	2,97 4,96	2,93 4,86

© Todos los derechos reservados

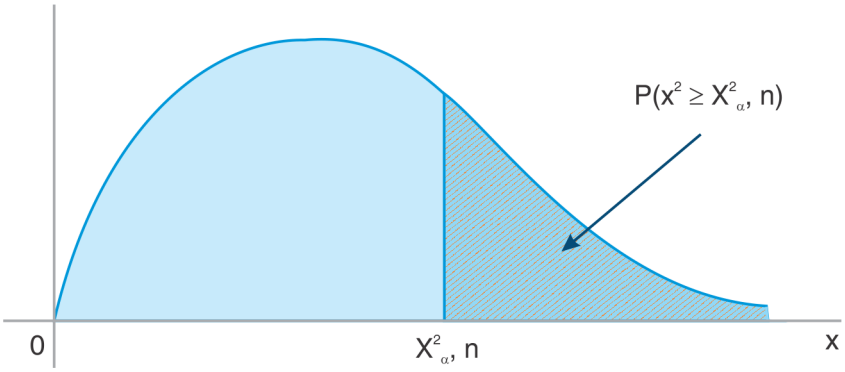
alfa = 0.05 **alfa = 0.01**	Grados de libertad del numerador													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**20**	**60**	**100**	**10000**
9	5,12 10,56	4,26 8,02	3,86 6,99	3,63 6,42	3,48 6,06	3,37 5,80	3,29 5,61	3,23 5,47	3,18 5,35	3,14 5,26	2,94 4,81	2,79 4,48	2,76 4,41	2,71 4,31
10	4,94 10,04	4,10 7,56	3,71 6,55	3,48 5,99	3,33 5,64	3,22 5,39	3,14 5,20	3,07 5,06	3,02 4,94	2,98 4,85	2,77 4,41	2,62 4,08	2,59 4,01	2,54 3,91
11	4,84 9,65	3,98 7,21	3,59 6,22	3,36 5,67	3,20 5,32	3,09 5,07	3,01 4,89	2,95 4,74	2,90 4,63	2,85 4,54	2,65 4,10	2,49 3,78	2,46 3,71	2,41 3,60
12	4,75 9,33	3,89 6,93	3,49 5,95	3,26 5,41	3,11 5,06	3,00 4,82	2,91 4,64	2,85 4,50	2,80 4,39	2,75 4,30	2,54 3,86	2,38 3,54	2,35 3,47	2,30 3,36
13	4,67 9,07	3,81 6,70	3,41 5,74	3,18 5,21	3,03 4,86	2,92 4,62	2,83 4,44	2,77 4,30	2,71 4,19	2,67 4,10	2,46 3,66	2,30 3,34	2,26 3,27	2,21 3,17
14	4,60 8,86	3,74 6,51	3,34 5,56	3,11 5,04	2,96 4,69	2,85 4,46	2,76 4,28	2,70 4,14	2,65 4,03	2,60 3,94	2,39 3,51	2,22 3,18	2,19 3,11	2,13 3,01
15	4,54 8,68	3,68 6,36	3,29 5,42	3,06 4,89	2,90 4,56	2,79 4,32	2,71 4,14	2,64 4,00	2,59 3,89	2,54 3,80	2,33 3,37	2,16 3,05	2,12 2,98	2,07 2,87
16	4,49 8,53	3,63 6,23	3,24 5,29	3,01 4,77	2,85 4,44	2,74 4,20	2,66 4,03	2,59 3,89	2,54 3,78	2,49 3,69	2,28 3,26	2,11 2,93	2,07 2,86	2,01 2,75
17	4,45 8,40	3,59 6,11	3,20 5,19	2,96 4,67	2,81 4,34	2,70 4,10	2,61 3,93	2,55 3,79	2,49 3,68	2,45 3,59	2,23 3,16	2,06 2,83	2,02 2,76	1,96 2,65
18	4,41 8,29	3,55 6,01	3,16 5,09	2,93 4,58	2,77 4,25	2,66 4,01	2,58 3,84	2,51 3,71	2,46 3,60	2,41 3,51	2,19 3,08	2,02 2,75	1,98 2,68	1,92 2,57
19	4,38 8,18	3,52 5,93	3,13 5,01	2,90 4,50	2,74 4,17	2,63 3,94	2,54 3,77	2,48 3,63	2,42 3,52	2,38 3,43	2,16 3,00	1,98 2,67	1,94 2,60	1,88 2,49
20	4,35 8,10	3,49 5,85	3,10 4,94	2,87 4,43	2,71 4,10	2,60 3,87	2,51 3,70	2,45 3,56	2,39 3,46	2,35 3,37	2,12 2,94	1,95 2,61	1,91 2,54	1,84 2,42
21	4,32 8,02	3,47 5,78	3,07 4,87	2,84 4,37	2,68 4,04	2,57 3,81	2,49 3,64	2,42 3,51	2,37 3,40	2,32 3,31	2,10 2,88	1,92 2,55	1,88 2,48	1,81 2,36
22	4,30 7,95	3,44 5,72	3,08 4,82	2,82 4,31	2,66 3,99	2,55 3,76	2,46 3,59	2,40 3,45	2,34 3,35	2,30 3,26	2,07 2,83	1,89 2,50	1,85 2,42	1,78 2,31
23	4,28 7,88	3,42 5,66	3,03 4,76	2,80 4,26	2,64 3,94	2,53 3,71	2,44 3,54	2,37 3,41	2,32 3,30	2,27 3,21	2,05 2,78	1,86 2,45	1,82 2,37	1,76 2,26
24	4,26 7,82	3,40 5,61	3,01 4,72	2,78 4,22	2,62 3,90	2,51 3,67	2,42 2,50	2,36 3,36	2,30 3,26	2,25 3,17	2,03 2,74	1,84 2,40	1,80 2,30	1,73 2,21
25	4,24 7,77	3,39 5,57	2,99 4,68	2,76 4,18	2,60 3,85	2,49 3,63	2,40 3,46	2,34 3,32	2,28 3,22	2,24 3,13	2,01 2,70	1,82 2,36	1,78 2,29	1,71 2,17
26	4,23 7,72	3,37 5,53	2,98 4,64	2,74 4,14	2,59 3,82	2,47 3,59	2,39 3,42	2,32 3,29	2,27 3,18	2,22 3,09	1,99 2,66	1,80 2,33	1,76 2,25	1,69 2,13
27	4,21 7,68	3,35 5,49	2,96 4,60	2,73 4,11	2,57 3,78	2,46 3,56	2,37 3,39	2,31 3,26	2,25 3,15	2,20 3,06	1,97 2,63	1,79 2,29	1,74 2,22	1,67 2,10
28	4,20 7,64	3,34 5,45	2,95 4,57	2,71 4,07	2,56 3,75	2,45 3,53	2,36 3,36	2,29 3,23	2,24 3,12	2,19 3,03	1,96 2,60	1,77 2,26	1,73 2,19	1,65 2,07
29	4,18 7,60	3,33 5,42	3,93 4,54	2,70 4,04	2,55 3,73	3,43 3,50	2,35 3,33	2,28 3,20	2,22 3,09	2,18 3,00	1,94 2,57	1,75 2,23	1,71 2,16	1,64 2,04
30	4,17 7,56	3,32 5,39	2,92 4,51	2,69 4,02	2,53 3,70	2,42 3,47	2,33 3,30	2,27 3,17	2,21 3,07	2,16 2,98	1,93 2,55	1,74 2,21	1,70 2,13	1,62 2,01
40	4,08 7,31	3,23 5,18	2,84 4,31	2,61 3,83	2,45 3,51	2,34 3,29	2,25 3,12	2,18 2,99	2,12 2,89	2,08 2,80	1,84 2,37	1,64 2,02	1,59 1,94	1,51 1,81
50	4,03 7,17	3,18 5,06	2,79 4,20	2,56 3,72	2,40 3,41	2,29 3,19	2,20 3,02	2,13 2,89	2,07 2,78	2,03 2,70	1,78 2,27	1,58 1,91	1,52 1,82	1,44 1,68

© Todos los derechos reservados

alfa = 0.05 **alfa = 0.01**	Grados de libertad del numerador													
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**20**	**60**	**100**	**10000**
100	3,94 6,90	3,09 4,82	2,70 3,98	2,46 3,51	2,31 3,21	2,19 2,99	2,10 2,82	2,03 2,69	1,97 2,59	1,93 2,50	1,68 2,07	1,45 1,69	1,39 1,60	1,28 1,43
500	3,86 6,69	3,01 4,65	2,62 3,82	2,39 3,36	2,23 3,05	2,12 2,84	2,03 2,68	1,96 2,55	1,90 2,44	1,85 2,36	1,59 1,92	1,35 1,52	1,28 1,41	1,12 1,17
1000	3,85 6,66	3,00 4,63	2,61 3,80	2,38 3,34	2,22 3,04	2,11 2,82	2,02 2,66	1,95 2,53	1,89 2,43	1,84 2,34	1,58 1,90	1,33 1,50	1,26 1,38	1,08 1,12
2000	2,85 6,65	3,00 4,62	2,61 3,79	2,38 3,33	2,22 3,03	2,10 2,81	2,01 2,65	1,94 2,52	1,88 2,42	1,84 2,33	1,58 1,89	1,32 1,48	1,25 1,37	1,06 1,09
10000	3,84 6,64	3,00 4,61	2,61 3,78	2,37 3,32	2,21 3,02	2,10 2,80	2,01 2,64	1,94 2,51	1,88 2,41	1,83 2,32	1,57 1,88	1,32 1,48	1,25 1,36	1,03 1,05

La cabecera de las columnas indica los Grados de Libertad (gl) del numerador de la Razón de Varianzas F y en las filas los gl del denominador. En cada casilla aparece los valores de F para α = 0.05 y α = 0.01 (este último en negrita). Supongamos un ANOVA en que se comparan 5 grupos con 3 individuos en cada uno. Tenemos gl_{Entre} = 4 y gl_{Dentro} = 10. En el cruce de la columna con cabecera "4" y la fila con cabecera "10" vemos las cantidades 3,33 y 5,64. Al valor F = 3.33 le corresponde P 0.05 y al valor F = 5.64 le corresponde P 0.01. Si F es menor de 3.33, P es mayor de 0.05, y si F es mayor de 5.64, P es menor de 0.01. Si F está entre 3.33 y 5.64, P está entre 0.05 y 0.01. Por ejemplo, si es F = 2.5, es P > 0.05 y si es F = 9.3, es P < 0.01. Los programas informáticos dan el valor exacto de P.

Distribución Chi - cuadrado



© Todos los derechos reservados

Nivel de significación					
Grados libertad	**0,1**	**0,05**	**0,025**	**0,01**	**0,005**
1	2,71	3,84	5,02	6,63	7,88
2	4,61	5,99	7,38	9,21	10,60
3	6,25	7,81	9,35	11,34	12,84
4	7,78	9,49	11,14	13,28	14,86
5	9,24	11,07	12,83	15,09	16,75
6	10,64	12,59	14,45	16,81	18,55
7	12,02	14,07	16,01	18,48	20,28
8	13,36	15,51	17,53	20,09	21,95
9	14,68	16,92	19,02	21,67	23,59
10	15,99	18,31	20,48	23,21	25,19
11	17,28	19,68	21,92	24,73	26,76
12	18,55	21,03	23,34	26,22	28,30
13	19,81	22,36	24,74	27,69	29,82
14	21,06	23,68	26,12	29,14	31,32
15	22,31	25,00	27,49	30,58	32,80
16	23,54	26,30	28,85	32,00	34,27
17	24,77	27,59	30,19	33,41	35,72
18	25,99	28,87	31,53	34,81	37,16
19	27,20	30,14	32,85	36,19	38,58
20	28,41	31,41	34,17	37,57	40,00
21	29,62	32,67	35,48	38,93	41,440
22	30,81	33,92	36,78	40,29	42,80
23	32,01	35,17	38,08	41,64	44,18
24	33,20	36,42	39,36	42,98	45,56
25	34,38	37,65	40,65	44,31	46,93
26	35,56	38,89	41,92	45,64	48,29
27	36,74	40,11	43,19	46,96	49,65
28	37,92	41,34	44,46	48,28	50,99
29	39,09	42,56	45,72	49,59	52,34
30	40,26	43,77	46,98	50,58	53,67
40	51,81	55,76	59,34	63,69	66,77
50	63,17	67,50	71,42	76,15	79,49
60	74,40	79,08	83,30	88,38	91,95
70	85,53	90,53	95,02	100,43	104,21
80	96,58	101,88	106,63	112,33	116,32
90	107,57	113,15	118,14	124,12	128,30
100	118,50	124,34	129,56	135,81	140,17

Cada fila da los datos para la Distribución Chi-Cuadrado con los grados de libertad, gl, indicados en la primera columna. Solamente se consideran las proporciones a la derecha de cada valor, como indica la figura. Por ejemplo, la distribución con, gl = 2, tiene el 0.10 de los valores por encima de 4.61, la distribución con gl = 15, tiene el 0.005 de los valores por encima de 32.80 y la distribución con gl = 40, tiene el 0.01 de los valores por encima de 63.39. Si un test da, por ejemplo, Chi-Cuadrado = 33.4 con 21 gl, vemos que 33.4 está entre 32.67 y 35.48, por lo que es $0.05 > P_{UNIL} > 0.025$.

© Todos los derechos reservados