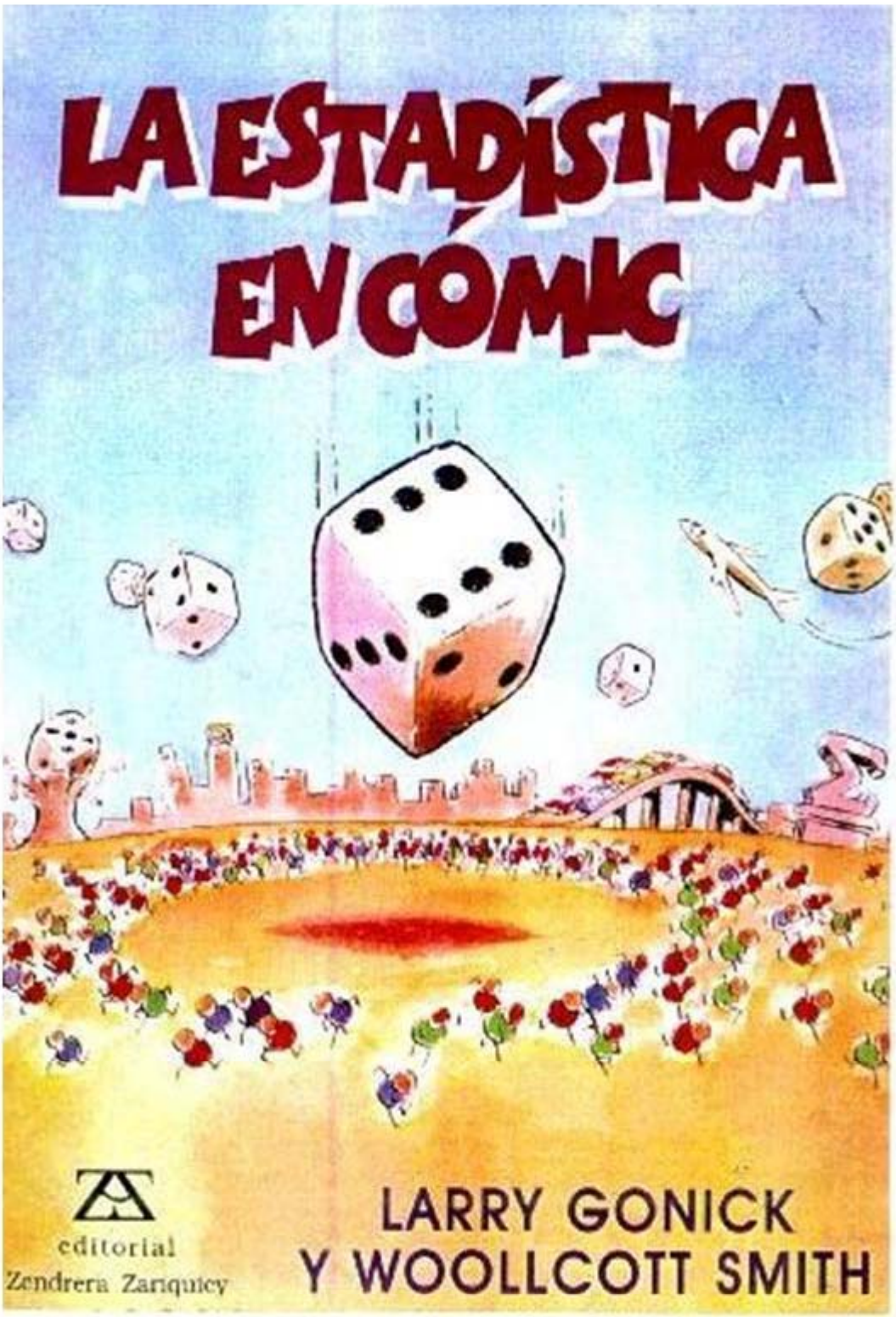
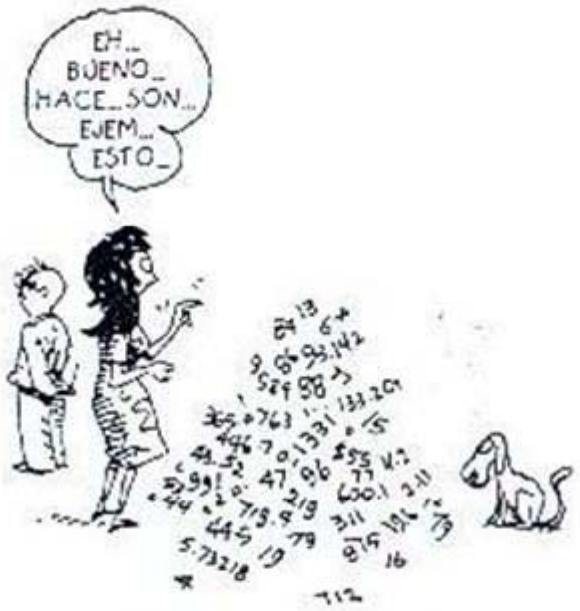
LA ESTADÍSTICA EN CÓMIC



LARRY GONICK
Y WOOLLCOTT SMITH

editorial
Zendrera Zariquiey

LA ESTADÍSTICA EN CÓMIC



LARRY GONICK
Y WOOLLCOTT SMITH

editorial
Zeralda Zariquiey

Edición original publicada en 1993 con el título:
The Cartoon Guide to Statistics

© Editorial Zendrera Zariquiey
Cardenal Vives i Tutó, 59
08034 Barcelona
Tel.: 932801234

Ilustraciones de: Larry Gonick

Traducido por: Laura Manero
Revisado por: Erik Cobo, Guadalupe Gómez y Pilar Muñoz
Primera edición: octubre 1999
ISBN: 84-8418-041-7
Depósito Legal: B.41462-1999
Producción: Addenda, s.c.c.l., Pau Claris, 92, 08010 Barcelona
Impresión: Edim, s.c.c.l., Badajoz, 147, 08018 Barcelona

♦ CONTENIDO ♦

Agradecimientos

NOS GUSTARÍA DAR LAS GRACIAS A CAROL COHEN, DE HARPERCOLLINS, POR HABERNOS SUGERIDO ESTE PROYECTO, A NUESTRA EDITORA ERICA SPABERG POR HABER DADO EL VISTO BUENO EN EL ÚLTIMO INSTANTE, Y A VICKY BIJUR, NUESTRA AGENTE LITERARIA, POR HABER INICIADO LA COLABORACIÓN GONICK/SMITH AL PRESENTAR A LOS COAUTORES.

LOS COMENTARIOS DE WILLIAM FAIRLEY Y LEAH SMITH SIRVIERON PARA MEJORAR LOS PRIMEROS BORRADORES DE ESTE LIBRO.

DONNA OKINO NOS PROPORCIONÓ AYUDA Y CONSEJO IMPAGABLES EN LA CREACIÓN DE LAS PÁGINAS DE CÓMIC. AFIRMA QUE ESCRIBIR UNA GUÍA EN CÓMIC ES MÁS DURO QUE CORRER UN MARATÓN, Y ELLA DEBERÍA SABERLO... PORQUE YA HA HECHO LAS DOS COSAS.

LA COMPAÑÍA ALTSYS CREÓ EL FONTOGRAPHER, EL MARAVILLOSO PROGRAMA DE SOFTWARE QUE NOS HA PERMITIDO SIMULAR TEXTO MANUSCRITO Y FÓRMULAS EN UN MACINTOSH.

Y, COMO LA DEDICATORIA ES SIEMPRE UNA CALLE DE DOBLE DIRECCIÓN, NOS QUITAMOS EL SOMBRERO ANTE LOS SUFRIDOS DISCÍPULOS DE SMITH DE LA UNIVERSIDAD DE TEMPLE, Y EN ESPECIAL ANTE EL GRUPO DE ESTUDIO DEL OTOÑO DE 1992, ORGANIZADO POR ADRIANA TORRES. EL FUTURO ES SUYO.

♦ Capítulo 1 ♦

¿QUÉ ES LA ESTADÍSTICA?

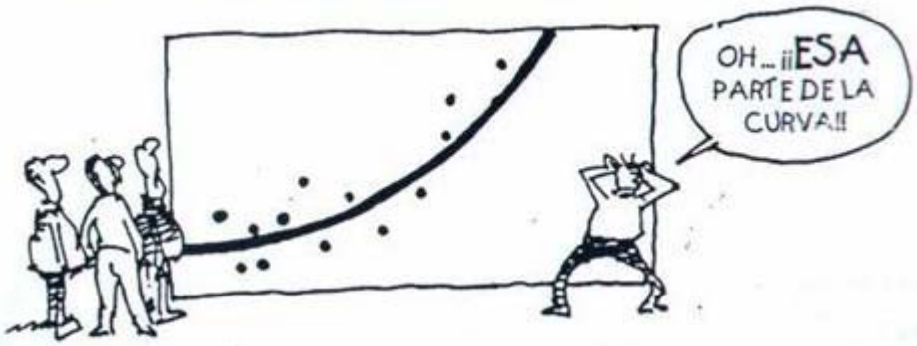
VAGAMOS POR LA VIDA TOMANDO DECISIONES BASADAS EN UNA INFORMACIÓN INCOMPLETA...

LA MAYORÍA DE NOSOTROS VIVIMOS CÓMODOS CON CIERTO NIVEL DE INCERTIDUMBRE.

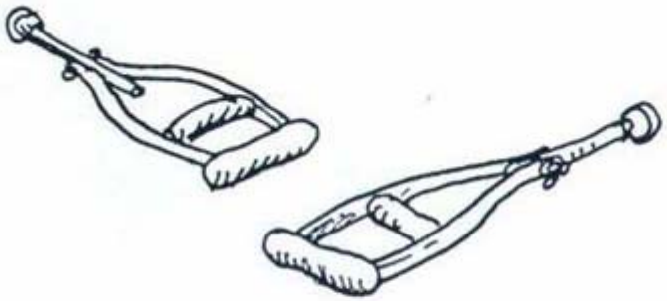
LO ESPECIAL DE LA ESTADÍSTICA, PARA SER PRECISOS, ES SU HABILIDAD DE **CUANTIFICAR** LA INCERTIDUMBRE. ESTO PERMITE A LOS ESTADÍSTICOS HACER **AFIRMACIONES CATEGÓRICAS** CON UNA SEGURIDAD TOTAL SOBRE EL NIVEL DE INCERTIDUMBRE.

¡NO ES SÓLO CUESTIÓN DE PEDIR UNA SOPA! LA ESTADÍSTICA TAMBIÉN TRATA ASUNTOS DE VIDA O MUERTE...

POR EJEMPLO, EN 1986, LA LANZADERA ESPACIAL **CHALLENGER** EXPLOTÓ CON SIETE ASTRONAUTAS DENTRO. LA DECISIÓN DE LANZAR LA NAVE A UNA TEMPERATURA DE −2 °C SE TOMÓ SIN REALIZAR UN SIMPLE ANÁLISIS SOBRE LA FIABILIDAD DE LOS DATOS A BAJAS TEMPERATURAS.



UN EJEMPLO MÁS POSITIVO ES EL DE LA **VACUNA DE SALK** CONTRA LA POLIO. EN 1954, SE PROBÓ LA VACUNA EN 400.000 NIÑOS CON UN RIGUROSO CONTROL PARA EVITAR RESULTADOS SESGADOS. UN BUEN ANÁLISIS ESTADÍSTICO DE LOS RESULTADOS ESTABLECIÓ LA EFICACIA DE LA VACUNA. Y ACTUALMENTE LA POLIO ESTÁ CASI ERRADICADA.



PARA REALIZAR ESTAS HAZAÑAS DE PRESTIDIGITACIÓN MATEMÁTICA, LOS ESTADÍSTICOS SE BASAN EN TRES DISCIPLINAS QUE ESTÁN ESTRECHAMENTE RELACIONADAS:

El análisis de datos,

LA RECOPILACIÓN, ORGANIZACIÓN Y RESUMEN DE LOS DATOS:

La probabilidad,

LAS LEYES DEL AZAR DENTRO Y FUERA DEL CASINO:

La inferencia estadística,

LA CIENCIA QUE EXTRAE CONCLUSIONES ESTADÍSTICAS A PARTIR DE DATOS CONCRETOS BASÁNDOSE EN EL CÁLCULO DE PROBABILIDADES.

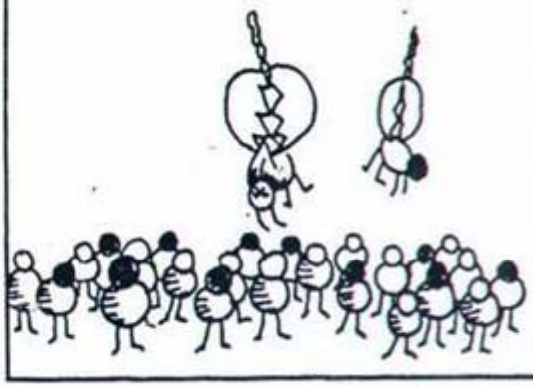
EN ESTE LIBRO, TRATAREMOS LAS TRES DISCIPLINAS Y LAS VEREMOS APLICADAS A UNA AMPLIA VARIEDAD DE SITUACIONES DEL MUNDO ACTUAL EN LAS QUE LA ESTADÍSTICA JUEGA UN PAPEL CLAVE.

EN EL CAPÍTULO 2 VEREMOS UN SIMPLE CONJUNTO DE DATOS, EL PESO DE UN GRUPO DE ESTUDIANTES UNIVERSITARIOS.

EN EL CAPÍTULO 3 ESTUDIAREMOS LAS LEYES DE LA PROBABILIDAD EN SU LUGAR DE NACIMIENTO, UN ANTRO DE JUEGO.

LOS CAPÍTULOS 4 Y 5 ENSEÑAN A DESCRIBIR EL MUNDO CON MODELOS DE PROBABILIDAD UTILIZANDO EL CONCEPTO DE VARIABLE ALEATORIA.

EL CAPÍTULO 6 PRESENTA UNO DE LOS PROCEDIMIENTOS ESENCIALES DE LA ESTADÍSTICA, TOMAR MUESTRAS DE UNA POBLACIÓN.



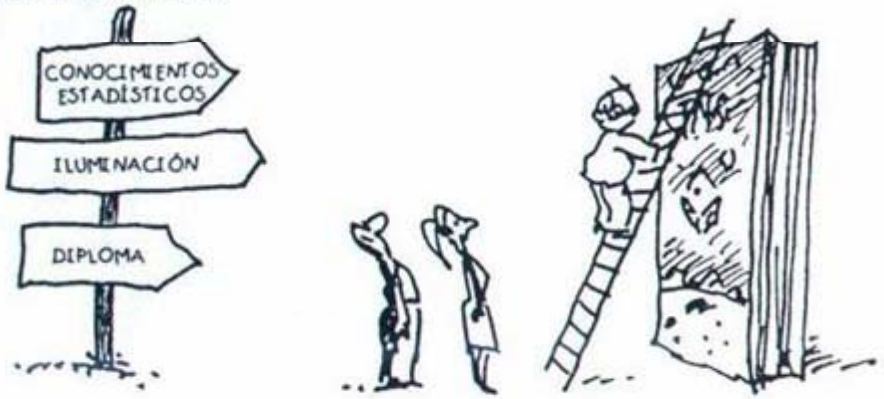
A PARTIR DEL CAPÍTULO 7 MOSTRAMOS CÓMO HACER INFERENCIA ESTADÍSTICA EN CAMPOS TAN COTIDIANOS COMO LOS SONDEOS DE OPINIÓN, EL CONTROL DE CALIDAD INDUSTRIAL, LAS PRUEBAS MÉDICAS, LOS PROGRAMAS DE SEGUIMIENTO PARA LA PROTECCIÓN MEDIO-AMBIENTAL, LA DISCRIMINACIÓN RACIAL Y LA LEY.

POR ÚLTIMO, CUANDO SE HABLA DE ESTA DISCIPLINA, RESULTA DIFÍCIL NO MENCIONAR ALGO MÁS: LA AMPLIA **DESCONFIANZA** EN LA ESTADÍSTICA DEL MUNDO ACTUAL. NO HAY QUIEN NO HAYA OÍDO HABLAR DE «ESTADÍSTICAS AMAÑADAS», Y EN LA VIDA COTIDIANA ES CASI IMPOSIBLE ENCONTRAR BUENOS ANÁLISIS ESTADÍSTICOS. ¡QUÉ LE VAMOS A HACER!*

*EL LIBRO *LYING WITH STATISTICS*- ES MUY POPULAR EN LOS EE.UU.: EN ÉSTE SE DA CUENTA DE FORMAS FRAUDULENTAS DE USAR LA ESTADÍSTICA. [N.T.]

NUESTRA HUMILDE OPINIÓN ES QUE NO SERÍA MALA IDEA APRENDER UN POCO **MÁS SOBRE EL TEMA**... Y POR ESO HEMOS ESCRITO ESTE LIBRO.

EN LOS SIGUIENTES CAPÍTULOS INTENTAMOS PRESENTAR LOS ELEMENTOS DE LA ESTADÍSTICA DE LA FORMA MÁS GRÁFICA E INTUITIVA POSIBLE. LO ÚNICO QUE NECESITAS PARA LEER ESTE LIBRO ES UN POCO DE **PACIENCIA**, ALGO DE **RAZONAMIENTO**, Y CIERTA TOLERANCIA AL **ÁLGEBRA**. O COMO MÍNIMO, ¡AL MENOS UNO DE ESTOS REQUISITOS!



♦ Capítulo 2 ♦

ESTADÍSTICA DESCRIPTIVA

LOS **DATOS** SON LA MATERIA PRIMA DE LOS ESTADÍSTICOS, LOS NÚMEROS QUE UTILIZAMOS PARA INTERPRETAR LA REALIDAD. EN TODO PROBLEMA ESTADÍSTICO HAY QUE RECOPILAR, DESCRIBIR Y ANALIZAR DATOS, O AL MENOS **PENSAR** EN LA RECOPILACIÓN, LA DESCRIPCIÓN Y EL ANÁLISIS DE LOS MISMOS.

ESTE CAPÍTULO SE CENTRA EN LA **DESCRIPCIÓN** DE DATOS. ¿CÓMO PODEMOS REPRESENTARLOS DE FORMA ÚTIL? ¿CÓMO DESCUBRIR LAS ESTRUCTURAS INTERNAS DE UN MONTÓN DE NÚMEROS DESNUDOS? ¿CÓMO SE PUEDE RESUMIR LA FORMA BÁSICA DE LOS DATOS?

BIEN, PARA DESCRIBIR DATOS, LO PRIMERO QUE NECESITAMOS SON DATOS QUE DESCRIBIR. ASÍ QUE, ¡VAMOS A RECOPILAR UNOS CUANTOS!

AQUÍ TENEMOS UNOS DATOS REALES: COMO PARTE DE UN EXPERIMENTO DE CLASE, SE RECOGIERON LOS PESOS DE 92 ESTUDIANTES DE PENNSYLVANIA. LOS RESULTADOS FUERON ÉSTOS:*

*EL PESO ESTÁ DADO EN LIBRAS. 1 LIBRA = 0.454 KG. [N.T.]

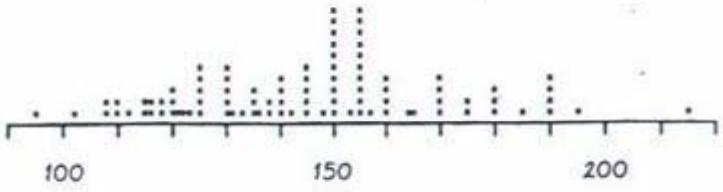
HOMBRES

140 145 160 190 155 165 150 190 195 138 160 155 153 145 170 175 175 170 180 135
170 157 130 185 190 155 170 155 215 150 145 155 155 150 155 150 180 160 135 160
130 155 150 148 155 150 140 180 190 145 150 164 140 142 136 123 155

MUJERES

140 120 130 138 121 125 116 145 150 112 125 130 120 130 131 120 118 125 135 125
118 122 115 102 115 150 110 116 108 95 125 133 110 150 108

PARA EMPEZAR, DIBUJAMOS UN **DIAGRAMA DE PUNTOS**: UN PUNTO POR CADA ESTUDIANTE SOBRE SU PESO CORRESPONDIENTE:



PESO EN LIBRAS

AQUÍ ENCONTRAMOS UN **PROBLEMA**: LA CONCENTRACIÓN EN *150* Y *155* LIBRAS. LOS ESTUDIANTES SUELEN DECIR SU PESO EN **MÚLTIPLOS DE CINCO LIBRAS**. EN SITUACIONES REALES COMO ÉSTA, REDONDEAR ASÍ LOS DATOS PUEDE LLEGAR A OCULTAR LAS ESTRUCTURAS GENERALES DE UN CONJUNTO DE DATOS... PERO, DE MOMENTO, LO PASAREMOS POR ALTO.

LOS DATOS SE PUEDEN RESUMIR EN UNA TABLA DE FRECUENCIAS. DIVIDIMOS LA LÍNEA DE NÚMEROS EN INTERVALOS IGUALES Y CONTAMOS EL NÚMERO DE PESOS QUE HAY ANOTADOS EN CADA INTERVALO. LA FRECUENCIA ES EL NÚMERO DE ANOTACIONES DE CADA INTERVALO. LA FRECUENCIA RELATIVA ES LA PROPORCIÓN DE PESOS DE CADA INTERVALO. ES DECIR, LA FRECUENCIA DIVIDIDA ENTRE EL TOTAL DE ESTUDIANTES.

INTERVALO DE CLASES	PUNTO MEDIO	FRECUENCIA	FRECUENCIA RELATIVA
87,5-102,4	95	2	0,022
102,5-117,4	110	9	0,098
117,5-132,4	125	19	0,206
132,5-147,4	140	17	0,185
147,5-162,4	155	27	0,293
162,5-177,4	170	8	0,087
177,5-192,4	185	8	0,087
192,5-207,4	200	1	0,011
207,5-222,4	215	1	0,011
TOTAL		92	1,000

NOTA: HEMOS INTENTADO QUE LOS LÍMITES DE LOS INTERVALOS NO COINCIDAN CON LOS MÚLTIPLOS DE CINCO LIBRAS. ASÍ EVITAMOS EL SESGO EN LOS DATOS QUE HAN FACILITADO LOS ESTUDIANTES.

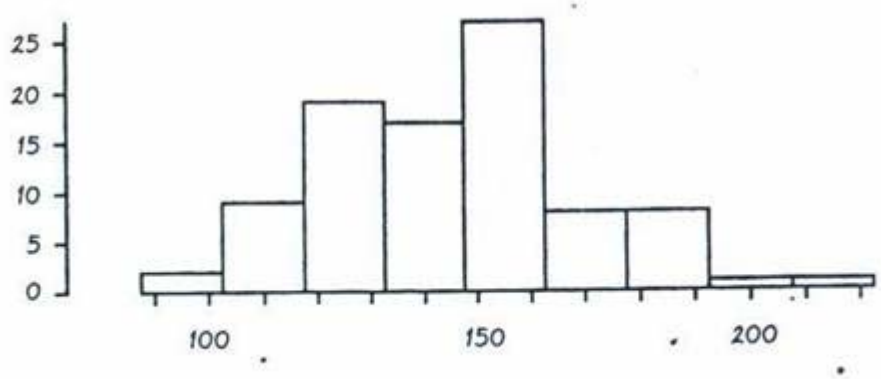
DIRECTRICES PARA ESTABLECER LOS INTERVALOS DE CLASE:

1) UTILIZA INTERVALOS IGUALES, QUE TENGAN SU PUNTO MEDIO EN NÚMEROS APROPIADOS.

2) CUANDO HAYA POCOS DATOS, UTILIZA POCOS INTERVALOS.

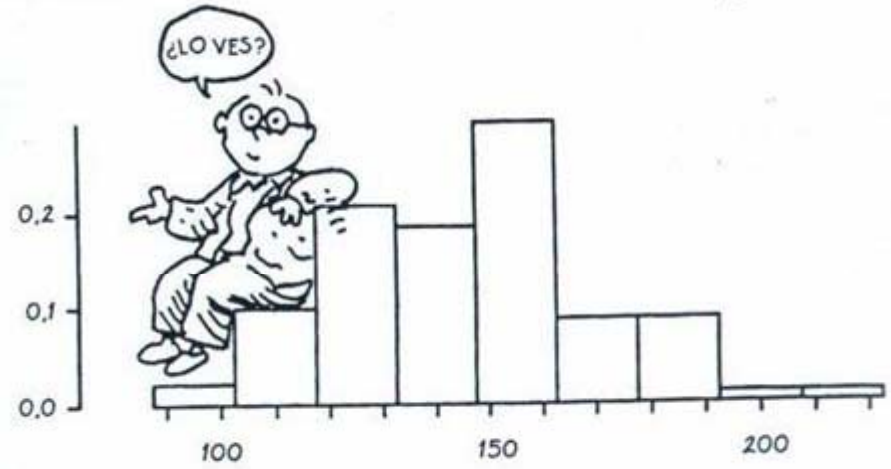
3) CUANDO LOS DATOS SEAN MUY NUMEROSOS, ¡UTILIZA TAMBIÉN MUCHOS INTERVALOS!

CON LA TABLA DE FRECUENCIAS MOSTRAMOS CUÁNTAS OBSERVACIONES HAY «ALREDEDOR» DE CADA VALOR. ESTA INFORMACIÓN TAMBIÉN SE PUEDE REPRESENTAR CON UN DIAGRAMA DE BARRAS QUE SE LLAMA **HISTOGRAMA**. CADA BARRA CUBRE UN INTERVALO Y TIENE SU PUNTO MEDIO EN EL CENTRO. LA ALTURA DE UNA BARRA REPRESENTA LA CANTIDAD DE PUNTOS, U OBSERVACIONES, DE CADA INTERVALO.



PESO EN LIBRAS

TAMBIÉN PODEMOS DIBUJAR UN **HISTOGRAMA DE FRECUENCIAS RELATIVAS**, REPRESENTANDO ÉSTAS EN FUNCIÓN DEL PESO. VISUALMENTE ES EL MISMO, SÓLO CAMBIA LA ESCALA VERTICAL.



PESO EN LIBRAS

EL ESTADÍSTICO JOHN TUKEY INVENTÓ UNA FORMA RÁPIDA PARA RESUMIR LOS DATOS Y MANTENER A LA VEZ TODAS LAS OBSERVACIONES INDIVIDUALES. LO LLAMÓ GRÁFICO DE TALLOS Y HOJAS.

EN ESTE CASO, EL TALLO ES UNA COLUMNA NUMÉRICA EN LA QUE SE REPRESENTA EL PESO DE DIEZ EN DIEZ LIBRAS, OMITIENDO EL ÚLTIMO DÍGITO.

9
10
11
12
13
14
15
16
17
18
19
20
21

LUEGO SE AÑADE EL ÚLTIMO DÍGITO DE CADA PESO EN LA LÍNEA CORRESPONDIENTE:

TALLO : HOJAS
9 :
10 :
11 : 628
12 : 0155005
13 : 080015
14 : 05
15 : 0
16 :
17 :
18 :
19 :
20 :
21 :

CUANDO SE HAN AÑADIDO TODOS LOS DATOS, EL DIAGRAMA TIENE ESTE ASPECTO:

9 : 5
10 : 288
11 : 628855060
12 : 01553005525
13 : 8500850600153
14 : 05505580502
15 : 505370550550505050050O
16 : 050004
17 : 055000
18 : 0500
19 : 00500
20 :
21 : 5]

POR ÚLTIMO, SE ORDENAN LAS «HOJAS».

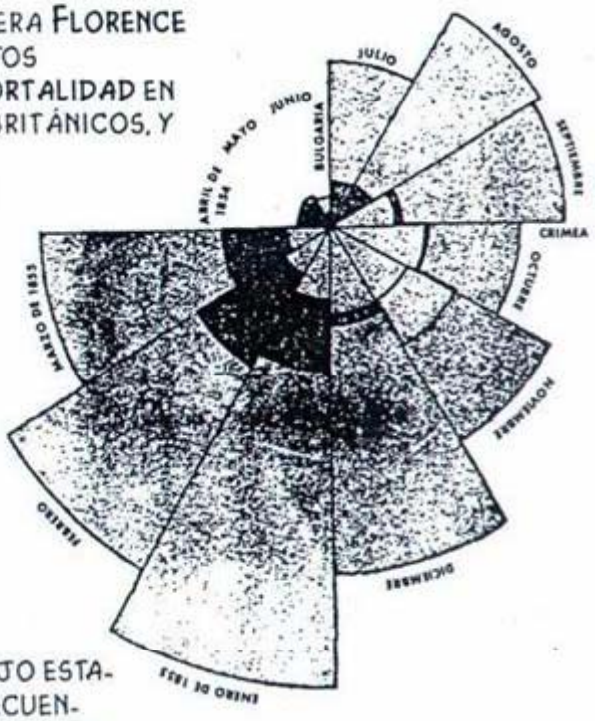
9 : 5
10 : 288
11 : 002556688
12 : 00012355555
13 : 0000013555688
14 : 00002555558
15 : 000000000035555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20 :
21 : 5

¡TODOS ESOS CEROS Y CINCOS DEMUESTRAN EL SESGO EN LA INFORMACIÓN QUE HAN PROPORCIONADO LOS ESTUDIANTES!

LA CONSTRUCCIÓN DE UN BUEN GRÁFICO TIENE ALGO DE ARTE Y ALGO DE CIENCIA.

¡Y, A VECES, ALGO DE POLÍTICA!

LA EMPRENDEDORA ENFERMERA **FLORENCE NIGHTINGALE** RECOPILÓ DATOS **ESTADÍSTICOS** SOBRE LA **MORTALIDAD** EN LOS HOSPITALES MILITARES BRITÁNICOS, Y CONSIGUIÓ HISTOGRAMAS TAN SORPRENDENTES COMO ÉSTE:
EL EJE RADIAL INDICA LAS MUERTES (TANTO EN LOS HOSPITALES COMO EN EL CAMPO DE BATALLA) DE LOS SOLDADOS BRITÁNICOS EN LA GUERRA DE CRIMEA.

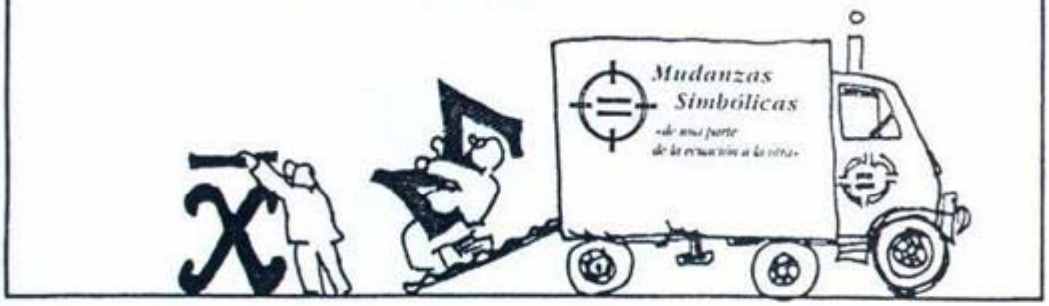


EL RESULTADO DE SU TRABAJO ESTADÍSTICO TUVO COMO CONSECUENCIA UNA INMEDIATA MEJORA DE LAS CONDICIONES HOSPITALARIAS Y UNA DISMINUCIÓN DE LA TASA DE MORTALIDAD.

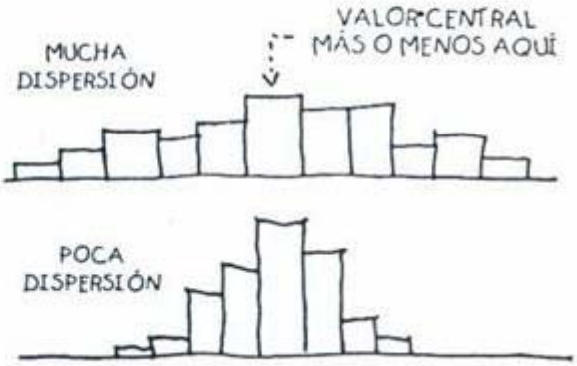
¡SALVADO POR LA ESTADÍSTICA!

RESUMEN NUMÉRICO

AHORA PASAMOS DE LOS GRÁFICOS A LAS FÓRMULAS. NUESTRO OBJETIVO ES CONSEGUIR CÁLCULOS SIMPLES DE LAS CARACTERÍSTICAS BÁSICAS DE UN CONJUNTO DE DATOS.



TODO CONJUNTO DE DATOS TIENE DOS PROPIEDADES PRINCIPALES: EL **VALOR CENTRAL**, O **TÍPICO**, Y LA **DISPERSIÓN** DE ESE VALOR. PUEDES HACERTE UNA IDEA CON ESTOS HISTOGRAMAS HIPOTÉTICOS.



SE PUEDE AVANZAR MUCHO CON POCA NOTACIÓN. SUPÓN QUE HEMOS HECHO UNA SERIE DE OBSERVACIONES... n OBSERVACIONES, PARA SER EXACTOS. ENTONCES ESCRIBIMOS

$$x_1, x_2, x_3, \ldots x_n$$

PARA CADA UNO DE LOS VALORES QUE HEMOS OBSERVADO. DE ESTA FORMA, n ES EL NÚMERO TOTAL DE DATOS, Y x_4, POR EJEMPLO, ES EL VALOR DEL CUARTO DATO.

LA **TABLA** ES UNA FORMA DE ORDENAR LOS DATOS:

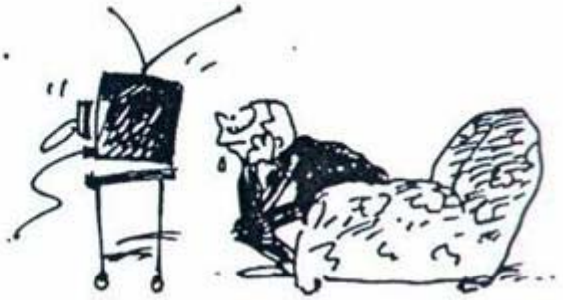
OBSERVACIÓN	1	2	3	4	...	n
VALOR DEL DATO	x_1	x_2	x_3	x_4	...	x_n

UN PEQUEÑO CONJUNTO DE $n = 5$ DATOS FACILITA LAS OPERACIONES. SUPÓN, POR EJEMPLO, QUE PREGUNTAMOS A CINCO PERSONAS CUÁNTAS HORAS DE TELEVISIÓN VEN A LA SEMANA, Y CONFECCIONAMOS LA SIGUIENTE TABLA:

OBSERVACIÓN	1	2	3	4	5
VALOR DEL DATO	5	7	3	38	7

ENTONCES $x_1 = 5$, $x_2 = 7$, $x_3 = 3$, $x_4 = 38$, Y $x_5 = 7$.

¿CUÁL ES EL «CENTRO» DE ESTOS DATOS? HAY DIFERENTES FORMAS DE CALCULARLO, PERO SÓLO VEREMOS DOS.



LA MEDIA

LA MEDIA SE REPRESENTA CON EL SÍMBOLO $\bar{x}$, Y SE OBTIENE DIVIDIENDO LA SUMA DE TODOS LOS DATOS ENTRE EL NÚMERO DE OBSERVACIONES:

$$\bar{x} = \frac{\text{SUMA DE LOS DATOS}}{n}$$

$$= \frac{x_1 + x_2 + \dots + x_n}{n}$$

EN NUESTRO EJEMPLO:

$$\bar{x} = \frac{5 + 7 + 3 + 38 + 7}{5} = \frac{60}{5}$$

$$= 12 \text{ HORAS}$$

ASÍ QUE, VAMOS A REPETIRLO. LA MEDIA DE UN CONJUNTO DE DATOS x_i ES

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{O BIEN} \quad \sum_{i=1}^{n} \frac{x_i}{n}$$

EN EL CASO DE NUESTROS 92 ALUMNOS DE PENNSYLVANIA, EL PESO MEDIO ES

$$\sum_{i=1}^{92} \frac{x_i}{92} = \frac{13.354}{92}$$

$$=$$

145,15 LIBRAS

LA MEDIANA

ES OTRO TIPO DE CENTRO. EL «PUNTO MEDIO» DE LOS DATOS. IGUAL QUE LA MEDIANA DE LA CARRETERA.

PARA ENCONTRAR LA MEDIANA DE UN CONJUNTO DE DATOS, ORDENAMOS LOS DATOS DE MENOR A MAYOR. LA MEDIANA ES EL VALOR QUE QUEDA EN EL CENTRO.

3 5 7 7 38

↑

MEDIANA

SI EL NÚMERO DE OBSERVACIONES ES PAR, EN CUYO CASO NO HAY NINGÚN PUNTO CENTRAL, HACEMOS LA MEDIA DE LOS DOS VALORES QUE QUEDAN EN EL CENTRO. ASÍ QUE SI LOS DATOS SON

3 5 7 7

↑

ESPACIO CENTRAL

HACEMOS LA MEDIA DE 5 Y 7: $\frac{5+7}{2} = 6$

ESTO NOS DA UNA REGLA GENERAL: ORDENAR LOS DATOS DE MENOR A MAYOR.

SI EL NÚMERO DE DATOS ES IMPAR, LA MEDIANA ES EL VALOR CENTRAL.

SI EL NÚMERO DE DATOS ES PAR, LA MEDIANA ES LA MEDIA DE LOS DOS DATOS MÁS CERCANOS AL CENTRO.

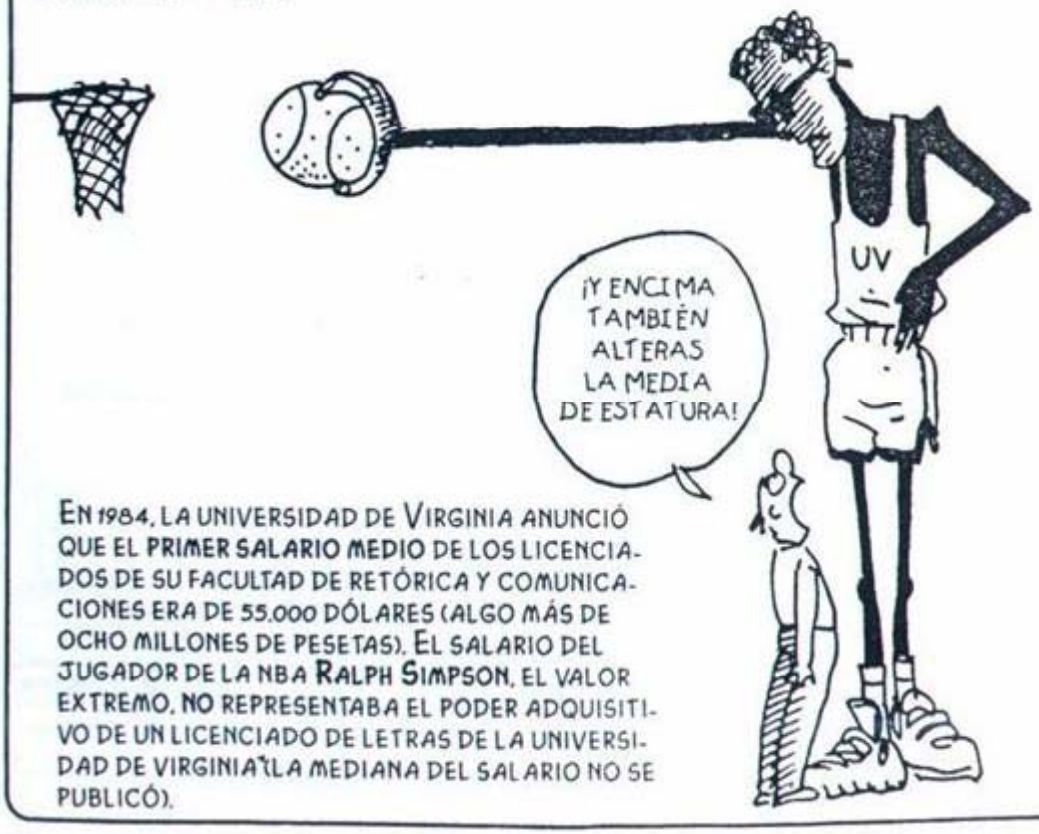
PARA ENCONTRAR LA MEDIANA DE LOS PESOS DE LOS ESTUDIANTES $n = 92$, PODEMOS UTILIZAR EL GRÁFICO DE TALLOS Y HOJAS ORDENADO: CUENTA HASTA LA OBSERVACIÓN NÚMERO 46. LA MEDIANA ES

$$\frac{x_{46} + x_{47}}{2} = \frac{145 + 145}{2}$$

$$= 145 \text{ LIBRAS}$$

```
 9 : 5
10 : 288
11 : 002556688
12 : 00012355555
13 : 0000013555688
14 : 000025555588
15 : 00000000003555555555557
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20 :
21 : 5
```

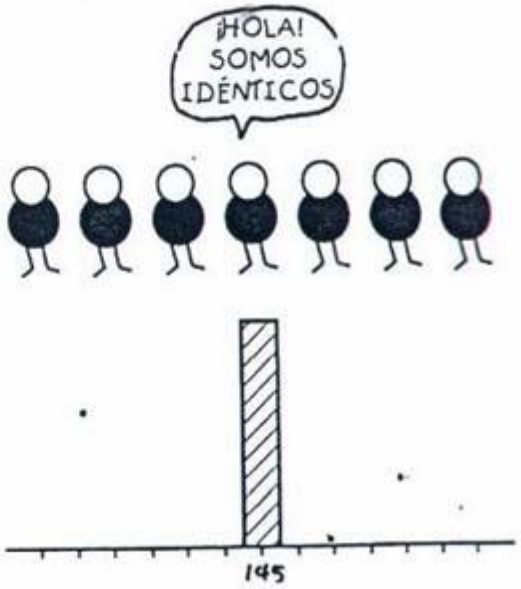
¿POR QUÉ EXISTEN DIFERENTES FORMAS DE CALCULAR EL CENTRO? CADA UNA TIENE SUS VENTAJAS. POR EJEMPLO, LA **MEDIANA** NO SE VE AFECTADA POR LOS DATOS MÁS ALEJADOS DEL CENTRO, LOS VALORES EXTREMOS QUE NO SON TÍPICOS DEL CONJUNTO DE DATOS. SUPONGAMOS QUE EN NUESTRO PEQUEÑO GRUPO DE TELESPECTADORES HAY UNA PERSONA QUE VE 200 HORAS DE TELEVISIÓN A LA SEMANA. NUESTROS DATOS SERÍAN ENTONCES 3, 5, 7, 7, 200. LA MEDIANA, 7, SIGUE SIENDO LA MISMA. ¡PERO AHORA LA MEDIA ES $\bar{x} = 45{,}8$!



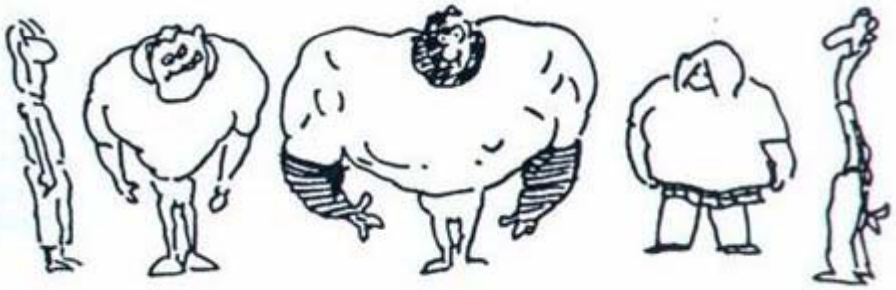
EN 1984, LA UNIVERSIDAD DE VIRGINIA ANUNCIÓ QUE EL **PRIMER SALARIO MEDIO** DE LOS LICENCIADOS DE SU FACULTAD DE RETÓRICA Y COMUNICACIONES ERA DE 55.000 DÓLARES (ALGO MÁS DE OCHO MILLONES DE PESETAS). EL SALARIO DEL JUGADOR DE LA NBA **RALPH SIMPSON**, EL VALOR EXTREMO, NO REPRESENTABA EL PODER ADQUISITIVO DE UN LICENCIADO DE LETRAS DE LA UNIVERSIDAD DE VIRGINIA (LA MEDIANA DEL SALARIO NO SE PUBLICÓ).

MEDIDAS DE DISPERSIÓN

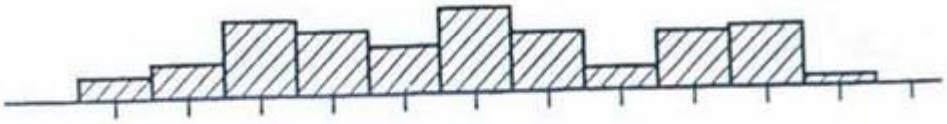
ADEMÁS DE CONOCER EL PUNTO CENTRAL DE UN CONJUNTO DE DATOS, TAMBIÉN QUEREMOS DESCRIBIR LA **DISPERSIÓN**, ES DECIR, A CUÁNTA DISTANCIA DEL CENTRO SE ENCUENTRAN LOS DATOS. POR EJEMPLO, SI TODOS LOS ESTUDIANTES PESARAN **EXACTAMENTE** 145 LIBRAS, NO HABRÍA DISPERSIÓN. NUMÉRICAMENTE, LA DISPERSIÓN SERÍA CERO Y EL HISTOGRAMA SERÍA MUY DELGADITO.



PERO SI MUCHOS DE LOS ESTUDIANTES PESARAN O BIEN MUCHO Y/O MUY POCO, VERÍAMOS OBVIAMENTE QUE HAY DISPERSIÓN. POR EJEMPLO, SI EL EQUIPO DE FÚTBOL HUBIERA PARTICIPADO EN EL EXPERIMENTO...



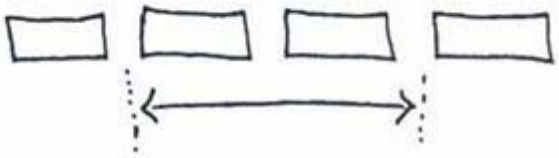
EL HISTOGRAMA SERÍA MÁS EXTENSO, ALGO ASÍ:



TAMBIÉN HAY VARIAS FORMAS DE MEDIR LA DISPERSIÓN. UNA ES EL

RECORRIDO INTERCUARTÍLICO

SE TRATA DE DIVIDIR LOS DATOS EN CUATRO GRUPOS IGUALES Y OBSERVAR LA DISTANCIA QUE SEPARA LOS GRUPOS EXTREMOS.



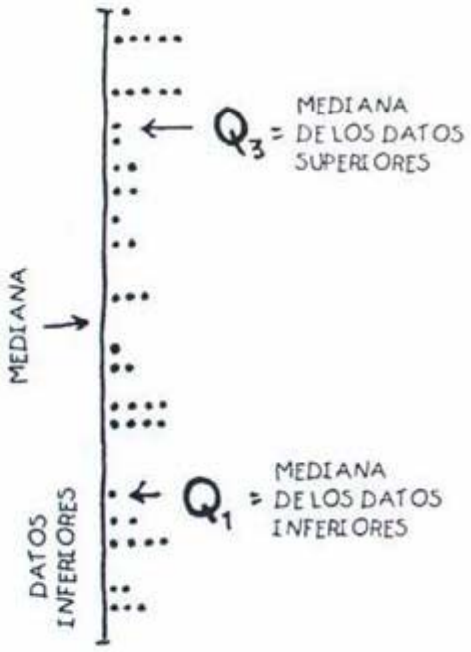
ÉSTA ES LA RECETA:

1) ORDENA LOS DATOS NUMÉRICAMENTE.

2) DIVIDE LOS DATOS POR LA MEDIANA EN DOS GRUPOS IGUALES (SI LA MEDIANA COINCIDE CON UN DATO, INCLÚYELO EN LOS DOS GRUPOS).

3) CALCULA LA MEDIANA DEL GRUPO INFERIOR. ÉSE ES EL PRIMER CUARTIL, O Q_1.

4) LA MEDIANA DEL GRUPO SUPERIOR ES EL TERCER CUARTIL, O Q_3.



EL RECORRIDO INTERCUARTÍLICO (IQR) ES LA DISTANCIA (O DIFERENCIA) QUE HAY ENTRE ELLOS:

$$IQR = Q_3 - Q_1$$

ESTOS SON LOS DATOS DE LOS PESOS, EN LOS QUE HEMOS DESTACADO LOS PUNTOS MEDIOS DEL GRUPO INFERIOR Y DEL SUPERIOR:

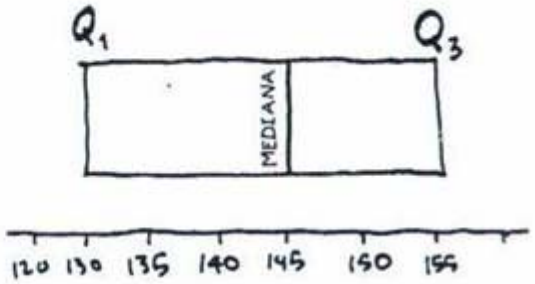
```
 9 : 5
10 : 288
11 : 002556688
12 : 000123555 55
13 : 0000013555688
14 : 00002555558
15 : 00000000003555555555 57
16 : 000045
17 : 000055
18 : 0005
19 : 00005
20:
21 : 5
```

Y VEMOS QUE

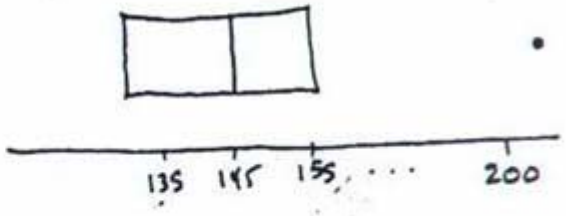
$$IQR = 156 - 125 = 31 \text{ LIBRAS}$$

ES DECIR, LA DIFERENCIA ENTRE LA MEDIANA DE LOS ESTUDIANTES QUE PESAN MUCHO Y LA MEDIANA DE LOS ESTUDIANTES QUE PESAN POCO.

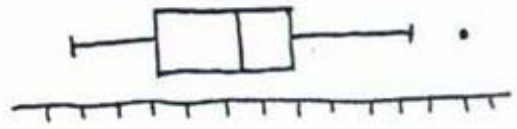
JOHN TUKEY INVENTÓ OTRO TIPO DE REPRESENTACIÓN PARA MOSTRAR EL IQR, EL GRÁFICO DE CAJA. LOS EXTREMOS DE LA CAJA SON LOS CUARTILES Q_1 Y Q_3. LA MEDIANA SE DIBUJA DENTRO DE LA CAJA.

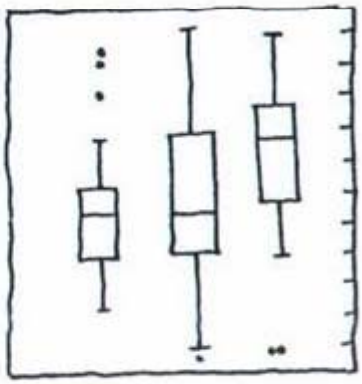


SI UN PUNTO SE ENCUENTRA A MÁS DE 1,5 IQR DE LOS EXTREMOS DE LA CAJA, SE CONSIDERA QUE ES UNA OBSERVACIÓN ATÍPICA, Y SE REPRESENTA INDIVIDUALMENTE.



POR ÚLTIMO, EXTENDEMOS LÍNEAS HASTA LOS PUNTOS MÁS ALEJADOS QUE NO SON OBSERVACIONES ATÍPICAS (ES DECIR, QUE SE ENCUENTRAN A MENOS DE 1,5 IQR DE LOS CUARTILES).



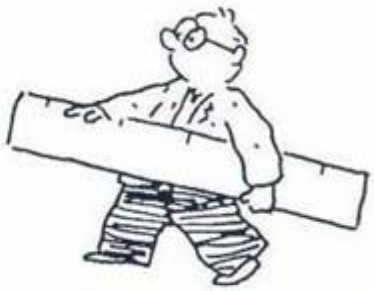


EL GRÁFICO CAJA ES MUY ÚTIL, EN ESPECIAL PARA REPRESENTAR LAS DIFERENCIAS ENTRE GRUPOS.

LA MEDIDA ESTÁNDAR DE LA DISPERSIÓN ES LA

DESVIACIÓN TÍPICA (TAMBIÉN DESVIACIÓN ESTÁNDAR)

A DIFERENCIA DEL IQR, QUE SE CALCULA A PARTIR DE LAS MEDIANAS, LA DESVIACIÓN TÍPICA MIDE LA DISPERSIÓN DE LOS DATOS DESDE LA MEDIA. UNA FORMA INTUITIVA DE VERLA ES COMO LA DISTANCIA MEDIA ENTRE LOS DATOS Y LA MEDIA $\bar{x}$...

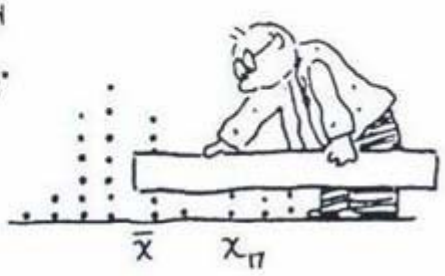


SIN EMBARGO, EN ESTA OCASIÓN UTILIZAMOS LAS DISTANCIAS ELEVADAS AL CUADRADO. O SEA, SI LA DISTANCIA AL CUADRADO ENTRE EL PUNTO x_i Y $\bar{x}$ ES $(x_i - \bar{x})^2$, ENTONCES

LA DISTANCIA CUADRÁTICA MEDIA = $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$

POR MOTIVOS TÉCNICOS, SE UTILIZA $n-1$ EN EL DENOMINADOR EN LUGAR DE n, Y DEFINIMOS ENTONCES LA VARIANZA MUESTRAL s^2:*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$



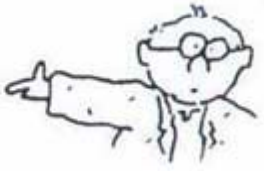
* TAMBIÉN ES CORRECTA VARIANCIA. [N.T.]

EN EL CONJUNTO DE DATOS {3 5 7 7 38}, DONDE $\bar{x} = 12$ Y $n = 5$, LA VARIANZA SE CALCULA ASÍ:

$$s^2 = \frac{(3-12)^2 + (5-12)^2 + (7-12)^2 + (7-12)^2 + (38-12)^2}{(5-1)}$$

$$= \frac{81 + 49 + 25 + 25 + 676}{4}$$

$$= 214$$



SIN EMBARGO, LA MEDIDA DE LA DISPERSIÓN DEBERÍA MEDIRSE EN LAS MISMAS **UNIDADES** QUE EL CONJUNTO INICIAL DE DATOS. EN EL EJEMPLO DE LOS PESOS, LA VARIANZA S^2 SE CALCULA EN LIBRAS AL CUADRADO... ¡UY!

LO MÁS LÓGICO ES HACER LA RAÍZ CUADRADA, Y ESO ES PRECISAMENTE LO QUE HACEMOS PARA DEFINIR LA...:

DESVIACIÓN TÍPICA

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

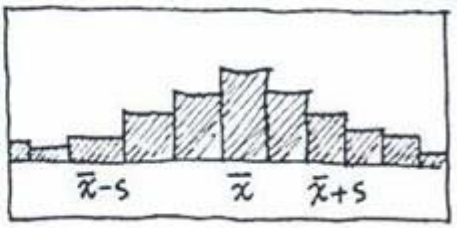
QUE EN NUESTRO PEQUEÑO CONJUNTO DE DATOS ES

$$s = \sqrt{214} = 14{,}63$$

INCLUSO CUANDO EL CONJUNTO DE DATOS ES PEQUEÑO, ¡LOS CÁLCULOS PUEDEN SER AGOTADORES! EN LA ACTUALIDAD BASTA CON APRETAR EL BOTÓN s DE LA CALCULADORA, O CONSULTAR SU VALOR CON LA AYUDA DE UN PAQUETE INFORMÁTICO.

Las propiedades de

$\bar{x}$ y s

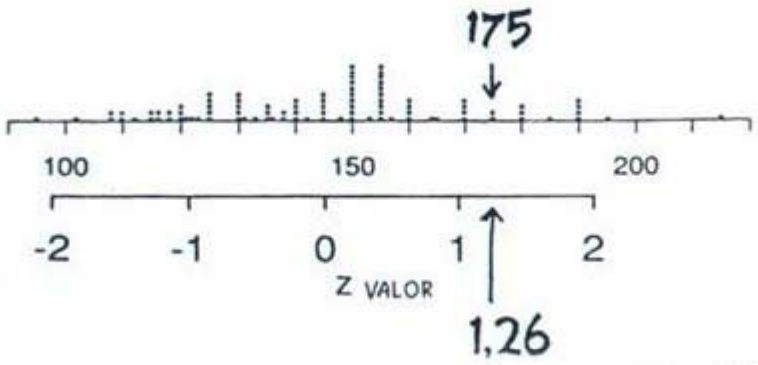


LA MEDIA Y LA DESVIACIÓN TÍPICA SON MUY ÚTILES PARA RESUMIR LAS PROPIEDADES DE **HISTOGRAMAS BASTANTE SIMÉTRICOS, SIN OBSERVACIONES ATÍPICAS**, O SEA, HISTOGRAMAS CON FORMA DE MONTAÑA.

A MENUDO RESULTA ÚTIL SABER CUÁNTAS DESVIACIONES TÍPICAS DISTA UN PUNTO DE LA MEDIA. ENTONCES DEFINIMOS z, O VALORES ESTANDARIZADOS, COMO LA DISTANCIA DESDE $\bar{x}$ POR DESVIACIÓN TÍPICA.

$$z_i = \frac{x_i - \bar{x}}{s} \quad \text{PARA CADA } i.$$

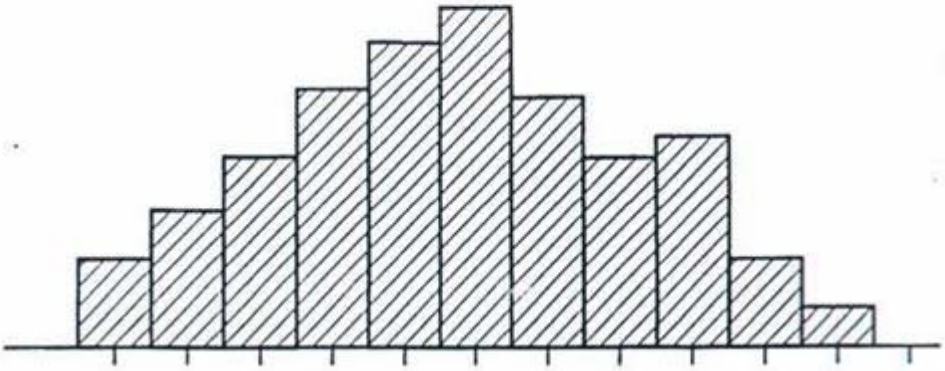
UNA z DE $+2$ QUIERE DECIR QUE LA OBSERVACIÓN SE ENCUENTRA DOS DESVIACIONES TÍPICAS POR ENCIMA DE LA MEDIA. EN LOS DATOS DE LOS PESOS DE LOS ESTUDIANTES ($\bar{x} = 145{,}2$ Y $s = 23{,}7$), PODEMOS REPRESENTAR LOS DATOS SIMULTÁNEAMENTE EN EL EJE x DEL PRINCIPIO Y UN EJE z.



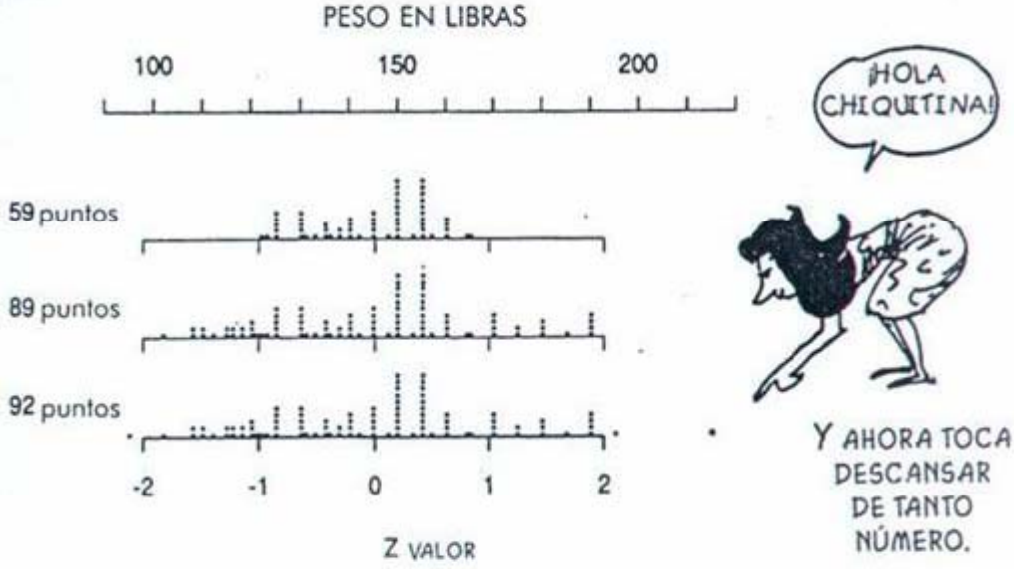
UN ESTUDIANTE QUE PESE 175 LIBRAS TIENE UNA z DE $\frac{175 - 145.2}{23.7} = 1{,}26$

una REGLA EMPÍRICA:

EN LOS CONJUNTOS DE DATOS CASI SIMÉTRICOS, CON FORMA DE MONTAÑA, ALREDEDOR DE UN **68%** DE LOS DATOS SE ENCUENTRA A MENOS DE UNA DESVIACIÓN TÍPICA DE LA MEDIA, Y EL **95%** ESTÁ A MENOS DE DOS DESVIACIONES TÍPICAS DE LA MEDIA.



SI MIRAMOS LOS PESOS, ESTA REGLA EMPÍRICA FUNCIONA BASTANTE BIEN: UN **64%** (= 59/92) DE LOS PESOS ESTÁN A MENOS DE UNA DESVIACIÓN TÍPICA DE LA MEDIA, Y UN **97%** (= 89/92) ESTÁN A MENOS DE DOS DESVIACIONES TÍPICAS DE LA MEDIA.



Y AHORA TOCA DESCANSAR DE TANTO NÚMERO.

¡HEMOS APRENDIDO MUCHO EN UN SOLO CAPÍTULO! EMPEZAMOS CON UN MONTÓN DE NÚMEROS DESORDENADOS, Y AHORA YA TENEMOS:

1) DIFERENTES FORMAS DE REPRESENTARLOS.

2) DOS CONCEPTOS DIFERENTES DEL CENTRO DE LOS DATOS, LA MEDIANA Y LA MEDIA.

3) DOS FORMAS DE CALCULAR LA DISPERSIÓN DE LOS DATOS ALREDEDOR DEL CENTRO.

4) HISTOGRAMAS EN FORMA DE MONTAÑA Y *z*, UNA VARIABLE QUE INDICA A CUÁNTAS DESVIACIONES TÍPICAS DE LA MEDIA SE ENCUENTRA UNA OBSERVACIÓN.

AHORA, PARA INVESTIGAR CON MÁS PROFUNDIDAD EL COMPORTAMIENTO DE LOS DATOS, VAMOS A DAR UN PEQUEÑO PASEO POR EL REINO DE LA ALEATORIEDAD... UNA TIERRA EN LA QUE TODO FUNCIONA **SIEMPRE** A LARGO PLAZO, Y DONDE NO HAY MÁS LEY QUE LA DEL **CASINO**...

♦ Capítulo 3 ♦

LA PROBABILIDAD

EN LA VIDA, NADA ES SEGURO. EN TODAS NUESTRAS ACCIONES, CALCULAMOS SIEMPRE LAS POSIBILIDADES DE UN BUEN RESULTADO, TANTO EN EL MUNDO DE LOS NEGOCIOS COMO EN LA MEDICINA O EL CLIMA. SIN EMBARGO, EN LA HISTORIA DE LA HUMANIDAD, LA **PROBABILIDAD**, EL ESTUDIO FORMAL DE LAS LEYES DEL AZAR, SE HA UTILIZADO PARA UNA SOLA COSA: EL JUEGO.

NADIE SABE CUÁNDO SE INVENTÓ EL JUEGO. COMO MÍNIMO SE REMONTA A TIEMPOS TAN ANCESTRALES COMO EL **ANTIGUO EGIPTO**. CUANDO HOMBRES Y MUJERES JUGABAN CON «ASTRÁGALOS» HECHOS CON LAS TABAS DE ANIMALES.

EL EMPERADOR ROMANO **CLAUDIO** (10 A. DE C. - 54 D. DE C.) ESCRIBIÓ EL PRIMER TRATADO SOBRE EL JUEGO. POR DESGRACIA, EL LIBRO *CÓMO GANAR A LOS DADOS* NO SE HA CONSERVADO.

LOS DADOS, TAL Y COMO LOS CONOCEMOS EN LA ACTUALIDAD, SE HICIERON MUY POPULARES EN LA EDAD MEDIA, A TIEMPO PARA QUE UN CALAVERA DEL RENACIMIENTO, **CHEVALIER DE MERE**, PROPUSIERA UN ENIGMA MATEMÁTICO:

CHEVALIER RAZONÓ QUE LA PROBABILIDAD DE OBTENER UNA TIRADA GANADORA ERA LA MISMA EN LOS DOS JUEGOS:

LA PROBABILIDAD DE UN SEIS = $\frac{1}{6}$

LA MEDIA EN CUATRO TIRADAS = $4 \cdot (\frac{1}{6}) = \frac{2}{3}$

LA PROBABILIDAD DE UN DOBLE SEIS EN UNA TIRADA = $\frac{1}{36}$

LA MEDIA EN 24 TIRADAS = $24 \cdot (\frac{1}{36}) = \frac{2}{3}$

ENTONCES, ¿¿POR QUÉ PERDÍA MÁS A MENUDO CON LA SEGUNDA APUESTA??

DE MERE LE PLANTEÓ LA PREGUNTA A SU AMIGO EL GENIO BLAISE PASCAL (1623-1666).

A PESAR DE QUE PASCAL HABÍA RENUNCIADO A LAS MATEMÁTICAS POR CONSIDERARLAS UNA FORMA DE DELEITE SEXUAL (!!), ACEPTÓ ESTUDIAR EL PROBLEMA DE DE MERE.

PASCAL ESCRIBIÓ A SU COMPAÑERO, TAMBIÉN GENIO, PIERRE DE FERMAT, Y EN EL TRANSCURSO DE UNAS CUANTAS CARTAS, LOS DOS YA HABÍAN DESARROLLADO LA TEORÍA DE LA PROBABILIDAD EN SU FORMA MODERNA (SIN VIÑETAS, CLARO).

DEFINICIONES BÁSICAS

MIENTRAS NUESTROS JUGADORES ECHAN UNA PARTIDA, NOSOTROS JUGAREMOS A SER CIENTÍFICOS Y ANALIZAREMOS LOS RESULTADOS:

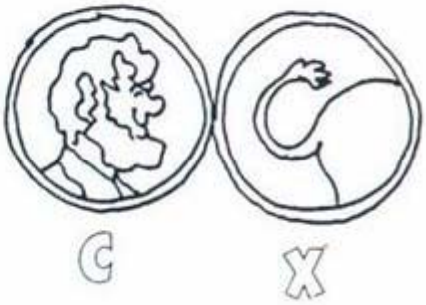
UN **experimento aleatorio** ES EL PROCESO DE OBSERVAR EL RESULTADO DE UN SUCESO CASUAL.

LOS **resultados elementales** SON TODOS LOS POSIBLES RESULTADOS INDIVISIBLES DEL EXPERIMENTO ALEATORIO.

EL **espacio muestral** ES EL CONJUNTO O EL COMPENDIO DE TODOS LOS RESULTADOS ELEMENTALES.

SI SE LANZA UNA MONEDA AL AIRE, POR EJEMPLO, EL EXPERIMENTO ALEATORIO CONSISTE EN TOMAR NOTA DE LOS RESULTADOS...

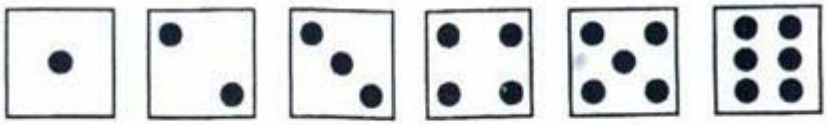
LOS RESULTADOS ELEMENTALES SON CARA O CRUZ...



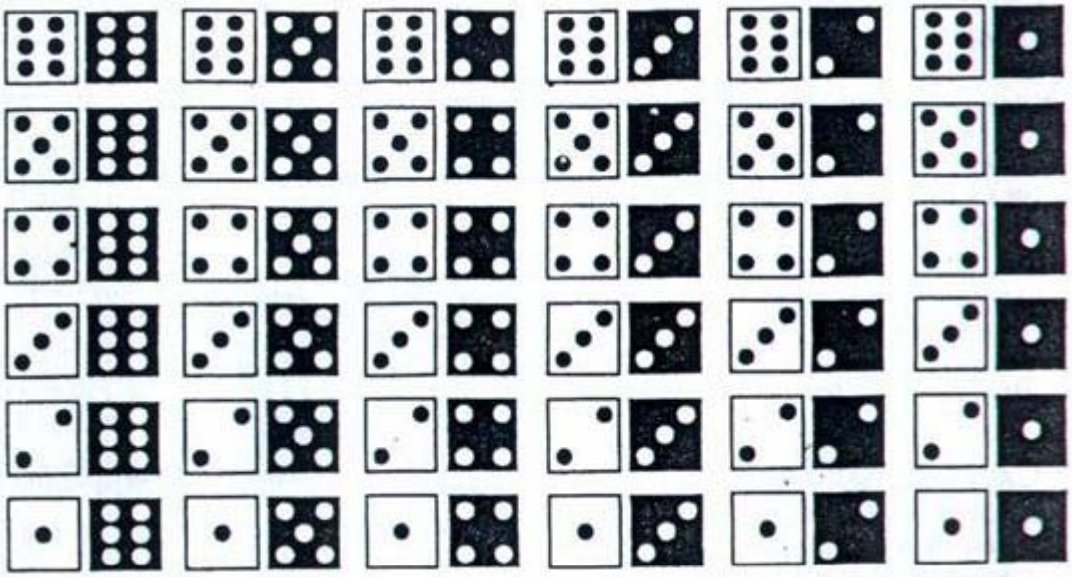
Y EL ESPACIO MUESTRAL SE ESCRIBE

$$\{C, X\}$$

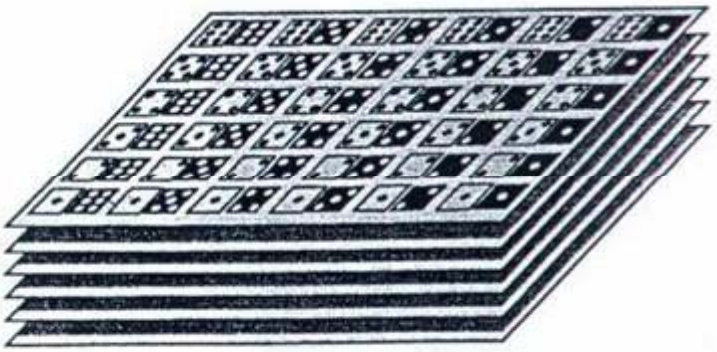
EL ESPACIO MUESTRAL DE UNA TIRADA CON **UN SOLO DADO** ES ALGO MAYOR.



Y CON **DOS** DADOS, EL ESPACIO MUESTRAL SERÍA ASÍ (HEMOS DIBUJADO UN DADO BLANCO Y OTRO NEGRO PARA DISTINGUIRLOS):

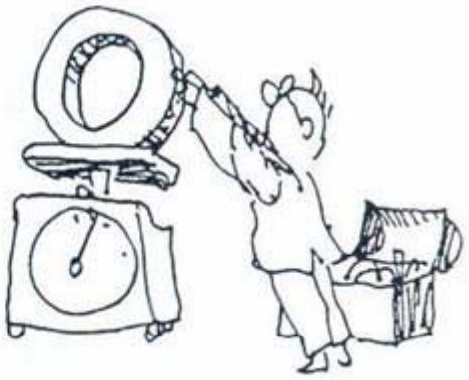


ESTE ESPACIO MUESTRAL TIENE 36 (6 × 6) RESULTADOS ELEMENTALES. CON TRES DADOS, EL ESPACIO TENDRÍA 216 ENTRADAS, COMO EN ESTE MONTÓN DE 6 × 6 × 6. ¿Y CON **CUATRO** DADOS?



EN ALGÚN MOMENTO TENDREMOS QUE DEJAR DE ENUMERAR Y EMPEZAR A RAZONAR...

VAMOS A IMAGINAR UN EXPERIMENTO ALEATORIO CON n RESULTADOS ELEMENTALES, $O_1, O_2, \ldots O_n$. QUEREMOS ASIGNARLES UN **PESO NUMÉRICO**, O **PROBABILIDAD**, A CADA UNO PARA MEDIR LA POSIBILIDAD DE QUE APAREZCA. ESCRIBIMOS LA PROBABILIDAD DE O_i COMO $P(O_i)$.



POR EJEMPLO, SI SE LANZA UNA MONEDA AL AIRE, HAY TANTAS POSIBILIDADES DE OBTENER CARA COMO CRUZ, Y LES ASIGNAMOS UNA PROBABILIDAD DE 0,5.

$$P(C) = P(X) = 0{,}5$$

CADA RESULTADO APARECE EN LA MITAD DE OCASIONES. ¡PREGUNTA A CUALQUIER **JUGADOR DE FÚTBOL**!

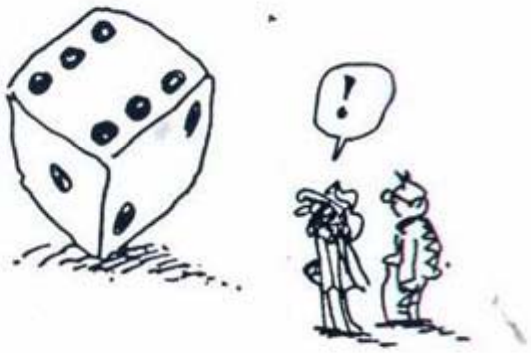
CUANDO SE TIRAN **DOS DADOS**, HAY 36 RESULTADOS ELEMENTALES Y TODOS TIENEN LA MISMA POSIBILIDAD DE APARECER.
ASÍ QUE LA PROBABILIDAD DE CADA UNO ES DE $\frac{1}{36}$.

POR EJEMPLO,

$$P(\text{NEGRO } 5,\ \text{BLANCO } 2) = \frac{1}{36}$$

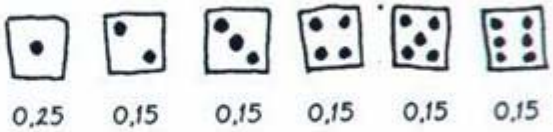
ES DECIR: SI TIRAMOS LOS DADOS MUCHAS VECES, A LARGO PLAZO OBTENDREMOS ESE RESULTADO EN $\frac{1}{36}$ DE LAS TIRADAS.

PERO, ¿QUÉ PASA SI NUESTRO JUGADOR HACE TRAMPAS Y TIRA UN DADO TRUCADO? A MODO DE EJEMPLO, SUPONDREMOS QUE EL UNO SALE UN 25% DE LAS VECES (A LARGO PLAZO).



EL ESPACIO MUESTRAL ES EL MISMO QUE EL DE UN DADO SIN TRUCAR

$\{1, 2, 3, 4, 5, 6\}$.



PERO LAS PROBABILIDADES SON DIFERENTES. AHORA $P(1) = 0{,}25$ Y EL RESTO DE PROBABILIDADES SUMAN EL 0,75 RESTANTE. SI 2, 3, 4, 5 Y 6 TUVIERAN LA MISMA PROBABILIDAD DE SALIR, CADA UNO TENDRÍA UNA PROBABILIDAD DE

$0{,}15 = \frac{1}{5}\ (0{,}75)$

EN GENERAL, LOS RESULTADOS ELEMENTALES NO TIENEN POR QUÉ TENER LA MISMA PROBABILIDAD.

¿QUÉ PODEMOS DECIR DE LA PROBABILIDAD DE $P(O_i)$ EN UN EXPERIMENTO ALEATORIO CUALQUIERA? EN PRIMER LUGAR.

$$P(O_i) \geq 0$$

LAS PROBABILIDADES NUNCA SON NEGATIVAS. UNA PROBABILIDAD DE CERO SIGNIFICA QUE ESE SUCESO NUNCA TENDRÁ LUGAR. UNA PROBABILIDAD POR DEBAJO DE CERO NO TIENE SENTIDO.

EN SEGUNDO LUGAR, SI UN SUCESO ES SEGURO, LE ASIGNAMOS UNA PROBABILIDAD DE 1. (A LARGO PLAZO, ESA ES LA PROPORCIÓN DE OCASIONES EN QUE OCURRIRÁ.) EN CONCRETO, LA PROBABILIDAD TOTAL DEL ESPACIO MUESTRAL DEBE SER 1. SI HACEMOS EL EXPERIMENTO, ¡ALGO TIENE QUE PASAR!

SI UNIMOS ESTOS DOS ÚLTIMOS PUNTOS, YA TENEMOS LAS PROPIEDADES CARACTERÍSTICAS DE LA PROBABILIDAD:

$$P(O_i) \geq 0$$

LA PROBABILIDAD NO ES NEGATIVA

$$P(O_1) + P(O_2) + \ldots + P(O_n) = 1$$

LA PROBABILIDAD TOTAL DE LOS RESULTADOS ELEMENTALES ES UNO.

IGUAL QUE HARÍA UN BUEN POLÍTICO, HEMOS EVITADO CIERTAS **PREGUNTAS INCÓMODAS** COMO: A) ¿QUÉ **SIGNIFICA** PROBABILIDAD?; Y B) ¿CÓMO ASIGNAMOS UNA PROBABILIDAD A UN RESULTADO?

AQUÍ TENEMOS DIFERENTES FORMAS DE VERLO:

LA PROBABILIDAD **Clásica**: ESTÁ BASADA EN EL JUEGO. LA SUPOSICIÓN FUNDAMENTAL ES QUE EL JUEGO ES JUSTO Y QUE TODOS LOS RESULTADOS ELEMENTALES TIENEN LA MISMA PROBABILIDAD.

LA **Frecuencia Relativa**: CUANDO UN EXPERIMENTO SE PUEDE REPETIR, LA PROBABILIDAD DE UN RESULTADO ES LA PROPORCIÓN DE OCASIONES EN LAS QUE APARECE A LARGO PLAZO.

LA PROBABILIDAD **Personal**: LA MAYORÍA DE LOS SUCESOS DE LA VIDA SON IRREPETIBLES. LA PROBABILIDAD PERSONAL ES LA VALORACIÓN PERSONAL QUE HACE UN INDIVIDUO DE LAS POSIBILIDADES DE OBTENER UN RESULTADO. SI UN JUGADOR CREE QUE UN CABALLO TIENE MÁS DE UN 50% DE POSIBILIDADES DE GANAR, HARÁ LA CORRESPONDIENTE APUESTA.

UN OBJETIVISTA UTILIZA LA DEFINICIÓN CLÁSICA DE PROBABILIDAD O LA DE LA FRECUENCIA RELATIVA. UN SUBJETIVISTA, O BAYESIANO, APLICA LAS LEYES FORMALES DEL AZAR A SUS PROBABILIDADES PERSONALES, O A LAS NUESTRAS.

OPERACIONES BÁSICAS

HASTA AHORA, SÓLO HEMOS HABLADO DE LA PROBABILIDAD DE LOS RESULTADOS ELEMENTALES. EN TEORÍA, CON ESO BASTARÍA PARA DESCRIBIR UN EXPERIMENTO ALEATORIO, PERO EN LA PRÁCTICA RESULTA UN POCO DIFÍCIL DE MANEJAR. POR EJEMPLO, ALGO TAN NORMAL COMO OBTENER UN SIETE NO ESTÁ CONTEMPLADO EN LOS RESULTADOS ELEMENTALES... ASÍ QUE TENEMOS QUE INTRODUCIR UNA NUEVA IDEA:

UN **SUCESO** ES UN CONJUNTO DE RESULTADOS ELEMENTALES. LA PROBABILIDAD DE UN SUCESO ES LA SUMA DE LAS PROBABILIDADES DE LOS RESULTADOS ELEMENTALES DEL CONJUNTO. POR EJEMPLO, ALGUNOS SUCESOS EN LA VIDA DE UN JUGADOR CON DOS DADOS SERÍAN:

DESCRIPCIÓN DEL SUCESO	RESULTADOS ELEMENTALES DEL SUCESO	PROBABILIDAD
A: TIRADA TOTAL SUMA 3	{(1,2), (2,1)}	$P(A) = \frac{2}{36}$
B: TIRADA TOTAL SUMA 6	{(1,5), (2,4), (3,3), (4,2), (5,1)}	$P(B) = \frac{5}{36}$
C: DADO BLANCO CAE EN 1	{(1,1), (1,2), (1,3), (1,4), (1,5), (1,6)}	$P(C) = \frac{6}{36}$
D: DADO NEGRO CAE EN 1	{(1,1), (2,1), (3,1), (4,1), (5,1), (6,1)}	$P(D) = \frac{6}{36}$

LO BELLO DE UTILIZAR SUCESOS EN LUGAR DE RESULTADOS ELEMENTALES ES QUE LOS PODEMOS COMBINAR PARA OBTENER OTROS DISTINTOS UTILIZANDO OPERACIONES LÓGICAS. LAS PALABRAS CLAVE SON **Y**, **O** Y **NO**.

ES DECIR, CON LOS SUCESOS E Y F, PODEMOS FORMAR NUEVOS SUCESOS:

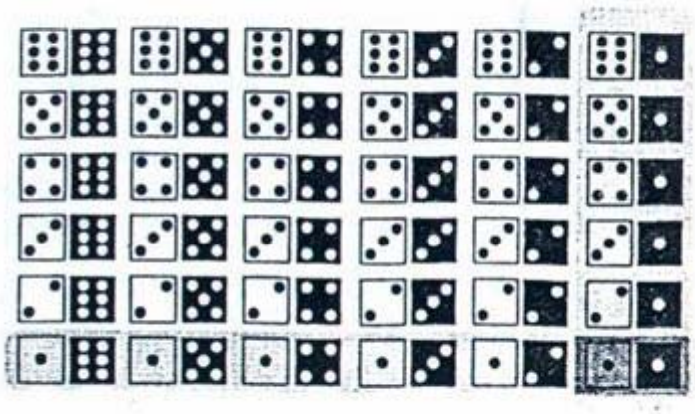
E **Y** F: TANTO EL SUCESO E COMO EL F OCURREN.

E **O** F: OCURRE EL SUCESO E, O EL F, O LOS DOS.

NO E: EL SUCESO E NO OCURRE.

SI COMBINAMOS LAS PRIMERAS DEFINICIONES DE PROBABILIDAD CON ESTAS OPERACIONES LÓGICAS, OBTENEMOS PODEROSAS FÓRMULAS PARA MANIPULAR LAS PROBABILIDADES.

VOLVAMOS AL EJEMPLO DE LOS DADOS. SI C ES EL SUCESO DADO BLANCO = 1, Y SI EL SUCESO D ES DADO NEGRO = 1, ENTONCES:



C O D ES LA ZONA SOMBREADA (DONDE UNO DE LOS DOS DADOS MUESTRA 1).

C Y D ES EL LUGAR EN QUE LAS DOS ZONAS SOMBREADAS SE SUPERPONEN (DONDE LOS DOS DADOS MUESTRAN 1).

ESTO ILUSTRA LA REGLA DE SUMA: PARA CUALESQUIERA DOS SUCESOS E, F,

$$\mathbf{P(E\ O\ F) = P(E) + P(F) - P(E\ Y\ F)}$$

SI SUMAMOS P(E) + P(F), SE DOBLAN LOS RESULTADOS ELEMENTALES QUE COMPARTEN E Y F. ASÍ QUE LUEGO HAY QUE RESTAR LO SOBRANTE, QUE ES P(E Y F).

EN EL EJEMPLO ANTERIOR,

$$P(C\ O\ D) = \frac{11}{36}$$

COMO PUEDES VER SI CUENTAS LOS RESULTADOS ELEMENTALES. DE IGUAL FORMA,

$$P(C\ Y\ D) = \frac{1}{36}$$

Y CONFIRMAMOS LA FÓRMULA:

$$P(C) + P(D) - P(C\ Y\ D)$$
$$= \frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36}$$
$$= P(C\ O\ D)^{*}$$

A VECES LA SUPERPOSICIÓN DE **E Y F** ESTÁ VACÍA, Y LOS DOS SUCESOS NO TIENEN RESULTADOS ELEMENTALES EN COMÚN. EN ESE CASO, E Y F SON **EXCLUYENTES** O INCOMPATIBLES Y ENTONCES P(E Y F) = 0. AQUÍ VEMOS LOS SUCESOS EXCLUYENTES **A**, TOTAL DE TIRADA SUMA 3, Y **B**, TOTAL DE TIRADA SUMA 6.

EN EL CASO DE SUCESOS EXCLUYENTES, SE DA UNA **REGLA ESPECIAL DE SUMA**: SI E Y F SON EXCLUYENTES, ENTONCES

P(E O F) = P(E) + P(F)

Y PODEMOS COMPROBAR QUE $P(A \text{ O } B) = \frac{7}{36} = \frac{2}{36} + \frac{5}{36} = P(A) + P(B)$

POR ÚLTIMO, LA **REGLA DE RESTA**: PARA CUALQUIER SUCESO E,

P(E) = 1 – P(NO E)

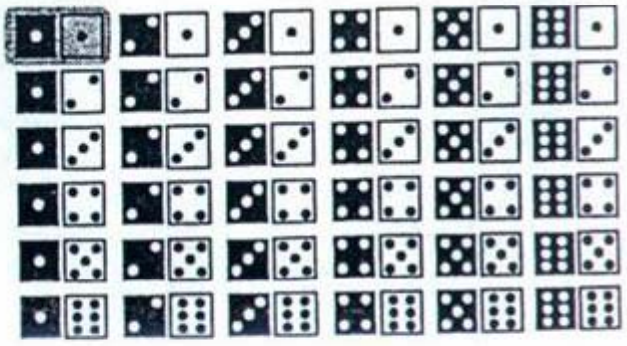
ESTA FÓRMULA RESULTA ÚTIL CUANDO P(NO E) ES MÁS FÁCIL DE CALCULAR QUE P(E). POR EJEMPLO, SI E ES EL SUCESO DE **NO OBTENER** UN DOBLE UNO. ENTONCES EL SUCESO NO E, OBTENER UN DOBLE UNO, TIENE UNA PROBABILIDAD $P(\text{NO E}) = \frac{1}{36}$.

ASÍ QUE

$P(E) = 1 - P(\text{NO E})$

$= 1 - \frac{1}{36}$

$= \frac{35}{36}$



SUPONGAMOS QUE ALTERAMOS UN POCO NUESTRO EXPERIMENTO, Y AHORA LANZAMOS EL DADO BLANCO ANTES QUE EL NEGRO. ¿QUÉ PROBABILIDAD HAY DE QUE LOS DOS SUMEN 3?

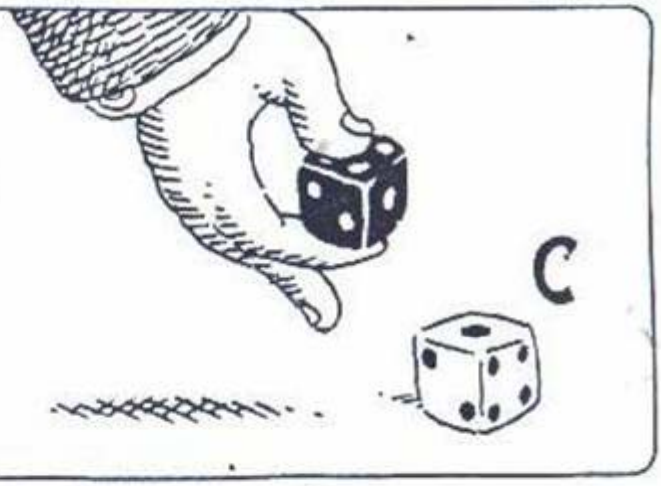
ANTES DE TIRAR LOS DADOS, LA PROBABILIDAD ES $P(A) = \frac{2}{36}$

PERO SI SUPONEMOS QUE EN EL DADO BLANCO HA SALIDO UN 1 (SUCESO C), ¿CUÁL ES AHORA LA PROBABILIDAD DE A?

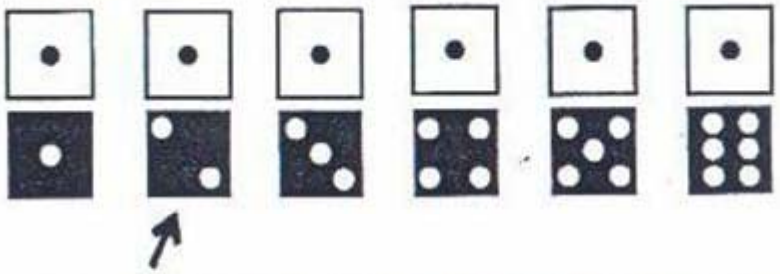
A ESTO SE LE LLAMA LA **PROBABILIDAD CONDICIONADA** DE QUE EL SUCESO A OCURRA, SABIENDO QUE EL SUCESO C YA SE HAYA DADO. ESCRIBIMOS

P(A | C)

Y DECIMOS «PROBABILIDAD DE A **CONDICIONADA** A C».



ANTES DE TIRAR LOS DADOS, EL ESPACIO MUESTRAL TENÍA 36 RESULTADOS, PERO AHORA QUE EL SUCESO C HA OCURRIDO, EL RESULTADO PERTENECE AL ESPACIO MUESTRAL C REDUCIDO.



EN EL ESPACIO MUESTRAL REDUCIDO DE SEIS RESULTADOS ELEMENTALES, SÓLO UNO (1, 2) SUMA 3. ASÍ QUE LA PROBABILIDAD CONDICIONADA ES 1/6.

EN GENERAL, PARA ENCONTRAR LA PROBABILIDAD CONDICIONADA P(E | F), CONTEMPLAMOS LOS SUCESOS E Y F COMO PARTE DEL ESPACIO MUESTRAL F REDUCIDO.

TRADUCIMOS ESTO A UNA DEFINICIÓN FORMAL:

LA PROBABILIDAD DE E CONDICIONADA A F ES

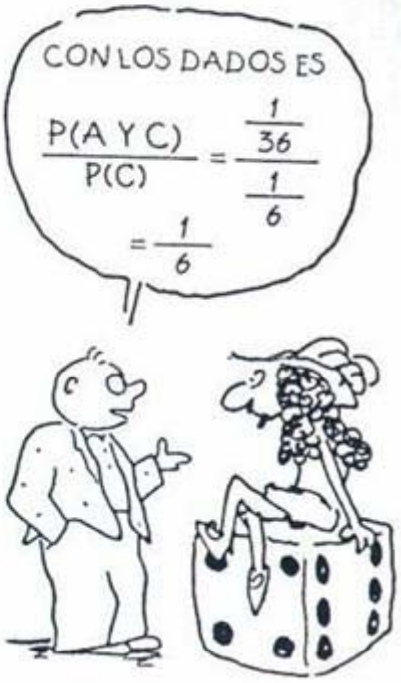
$$P(E|F) = \frac{P(E \text{ y } F)}{P(F)}$$

CON ESTA FÓRMULA SE PUEDEN VERIFICAR ALGUNOS HECHOS INTUITIVOS:

$P(E|E) = 1$ (UNA VEZ E HA OCURRIDO, YA ES SEGURO.)

CUANDO E Y F SON EXCLUYENTES,

$P(E|F) = 0$ (UNA VEZ F HA OCURRIDO, E ES IMPOSIBLE.)



SI REESCRIBIMOS LA DEFINICIÓN, OBTENEMOS LA REGLA DE LA MULTIPLICACIÓN:

$$P(E \text{ Y } F) = P(E|F)P(F)$$

QUE NOS GUSTARÍA REDUCIR HASTA UNA REGLA «ESPECIAL» DE MULTIPLICACIÓN, BAJO LAS CONDICIONES FAVORABLES DE QUE $P(E|F) = P(E)$. ¡SERÍA FANTÁSTICO!

LA INDEPENDENCIA y la regla especial de multiplicación.

DOS SUCESOS E Y F SON INDEPENDIENTES SI LA APARICIÓN DE UNO NO INFLUYE EN LA PROBABILIDAD DEL OTRO. POR EJEMPLO, LA TIRADA DE UN DADO NO TIENE NINGÚN EFECTO SOBRE LA DEL OTRO (¡A NO SER QUE ESTÉN PEGADOS, UNIDOS POR UN IMÁN, ETC.!).

EN TÉRMINOS DE PROBABILIDAD CONDICIONADA, ESTO EQUIVALE A DECIR QUE $P(E) = P(E|F)$ O, DE IGUAL FORMA, $P(F) = P(F|E)$. CUANDO E Y F SON INDEPENDIENTES, TENEMOS UNA REGLA ESPECIAL DE MULTIPLICACIÓN:

$$\mathbf{P(E\ Y\ F) = P(E)\,P(F)}$$

VAMOS A VERIFICAR LA INDEPENDENCIA DE LOS DADOS UTILIZANDO ESTAS FÓRMULAS. C ES EL SUCESO EL DADO BLANCO DÉ 1; D ES EL SUCESO EL DADO NEGRO DÉ 1, Y TENEMOS:

$$P(C|D) = \frac{P(C\ Y\ D)}{P(D)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} = P(C)$$

¡PERO QUE EL DADO BLANCO DÉ 1 AFECTA CLARAMENTE A LAS POSIBILIDADES DE QUE LOS DOS DADOS SUMEN 3!

$$P(A|C) = \frac{P(A\ Y\ C)}{P(C)} = \frac{P(1,2)}{P(C)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6} \neq P(A) = \frac{1}{18}$$

ASÍ QUE ESTOS DOS SUCESOS NO SON INDEPENDIENTES.

ANTES DE CONTINUAR, VAMOS A RESUMIR LAS REGLAS QUE HEMOS ACUMULADO:

REGLA DE SUMA:

$$P(E \text{ O } F) = P(E) + P(F) - P(E \text{ Y } F)$$

REGLA ESPECIAL DE LA SUMA: CUANDO E Y F SON MÚTUAMENTE EXCLUYENTES.

$$P(E \text{ O } F) = P(E) + P(F)$$

REGLA DE LA RESTA:

$$P(E) = 1 - P(\text{NO } E)$$

REGLA DE LA MULTIPLICACIÓN:

$$P(E \text{ Y } F) = P(E \mid F)P(F)$$

REGLA ESPECIAL DE LA MULTIPLICACION: CUANDO E Y F SON INDEPENDIENTES.

$$P(E \text{ Y } F) = P(E)P(F)$$

Y, POR FIN, EL PROBLEMA DE DE MERE... VAMOS A SUPONER QUE EL SUCESO E ES CONSEGUIR AL MENOS UN SEIS EN CUATRO TIRADAS DE UN SOLO DADO. ¿CUÁNTO ES P(E)? ESTE ES UNO DE LOS CASOS EN LOS QUE ES MÁS SENCILLO DESCRIBIR EL NEGATIVO: NO E ES EL SUCESO NO CONSEGUIR UN SEIS EN CUATRO TIRADAS.

SI A_i ES EL SUCESO NO CONSEGUIR UN SEIS EN LA TIRADA NÚMERO i, SABEMOS QUE $P(A_i) = \frac{5}{6}$. TAMBIÉN SABEMOS QUE LAS TIRADAS SON INDEPENDIENTES, ASÍ QUE

$$P(\text{NO } E) = P(A_1 \text{ Y } A_2 \text{ Y } A_3 \text{ Y } A_4)$$

REGLA DE MULTIPLICACIÓN →

$$= \left(\frac{5}{6}\right)^4 = 0{,}482.$$

ENTONCES,

$$P(E) = 1 - P(\text{NO } E) = 0{,}518$$

AHORA, A POR LA SEGUNDA PARTE: F ES EL SUCESO CONSEGUIR AL MENOS UN DOBLE SEIS EN 24 TIRADAS. DE NUEVO, NO F ES MÁS SENCILLO DE DESCRIBIR: ES EL SUCESO NO CONSEGUIR NINGÚN DOBLE SEIS.

SI B_i ES EL SUCESO, NO CONSEGUIR NINGÚN DOBLE SEIS EN LA TIRADA NÚMERO i, ASÍ QUE
NO F. = B_1 Y B_2 Y ... B_{24}.
LA PROBABILIDAD DE CADA B ES

$$P(B_i) = \frac{35}{36}$$, ASÍ QUE

$$P(\text{NO F}) = \left(\frac{35}{36}\right)^{24} = 0{,}509$$

(SEGÚN LA REGLA DE LA MULTIPLICACIÓN), Y PODEMOS LLEGAR A LA CONCLUSIÓN DE QUE

$$P(F) = 1 - P(\text{NO F}) = 1 - 0{,}509$$

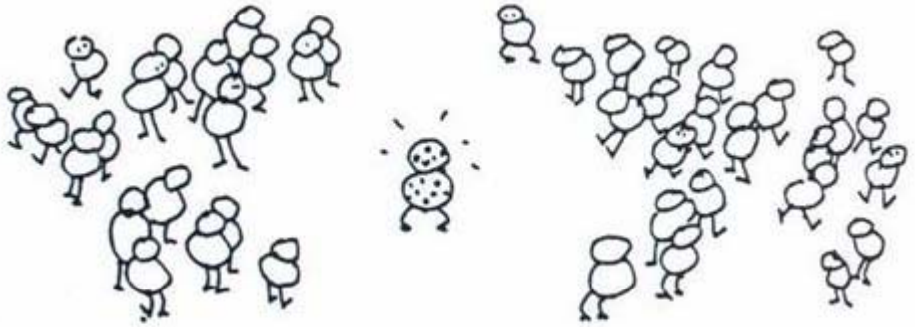
$$= 0{,}491$$

DE MERE LE DIJO A PASCAL QUE HABÍA OBSERVADO QUE EL SUCESO F APARECÍA CON MENOS FRECUENCIA QUE EL SUCESO E, PERO ERA INCAPAZ DE EXPLICAR POR QUÉ... DE LO QUE DEDUCIMOS QUE DE MERE JUGABA MUY A MENUDO Y ¡ANOTABA SUS JUGADAS!

EL TEOREMA DE BAYES y el caso de los falsos positivos

PARA CONTEMPLAR APLICACIONES MÁS SERIAS DE LA PROBABILIDAD CONDICIONADA, VAMOS A ADENTRARNOS EN UN TERRENO DE VIDA O MUERTE...

SUPONGAMOS QUE UNA EXTRAÑA ENFERMEDAD INFECCIOSA AFECTA A UNO DE CADA 1.000 HABITANTES DE UNA POBLACIÓN...



Y SUPONGAMOS QUE EXISTE UNA PRUEBA FIABLE, PERO NO INFALIBLE, PARA DETECTAR LA ENFERMEDAD: SI UNA PERSONA HA CONTRAÍDO LA ENFERMEDAD, LA PRUEBA RESULTA POSITIVA EN UN 99% DE LOS CASOS. POR OTRO LADO, LA PRUEBA TAMBIÉN PRODUCE FALSOS RESULTADOS POSITIVOS. UN 2% DE PACIENTES SANOS TAMBIÉN DAN UN RESULTADO POSITIVO, Y TÚ ACABAS DE RECIBIR EL TUYO: POSITIVO. ¿CUÁL ES LA PROBABILIDAD DE QUE TENGAS LA ENFERMEDAD?

TENEMOS DOS SUCESOS CON LOS QUE TRABAJAR:

A: EL PACIENTE PADECE LA ENFERMEDAD
B: EL PACIENTE DA UN RESULTADO POSITIVO.

LA INFORMACIÓN SOBRE LA EFICACIA DE LAS PRUEBAS SE PUEDE ESCRIBIR:

$P(A) = 0{,}001$ (UN PACIENTE DE CADA 1.000 PADECE LA ENFERMEDAD)

$P(B|A) = 0{,}99$ (LA PROBABILIDAD DE UN RESULTADO POSITIVO, DADA LA ENFERMEDAD, ES DE 0,99)

$P(B|\text{NO } A) = 0{,}02$ (LA PROBABILIDAD DE UN FALSO POSITIVO, DADO UN CASO DE NO INFECCIÓN, ES DE 0,02)

Y NOS PREGUNTAMOS:

$P(A|B) = ¿\text{QUÉ?}$ (LA PROBABILIDAD DE PADECER LA ENFERMEDAD, DADO UN RESULTADO POSITIVO)

YA QUE EL TRATAMIENTO DE LA ENFERMEDAD PRODUCE GRAVES EFECTOS SECUNDARIOS, LA DOCTORA, SU ABOGADA Y EL ABOGADO DE SU ABOGADA LLAMAN A JOE BAYES, QUE TIENE UN CONSULTORIO DE PROBABILIDADES, PARA QUE LES DÉ UNA RESPUESTA. JOE APLICA EL TEOREMA QUE DESARROLLÓ UN ANTEPASADO SUYO, EL RDO. THOMAS BAYES (1744-1809).

JOE EMPIEZA CON UNA TABLA DE 2 × 2, QUE DIVIDE EL ESPACIO MUESTRAL EN CUATRO CASOS EXCLUYENTES. REPRESENTA TODAS LAS COMBINACIONES POSIBLES DEL ESTADO DE LA ENFERMEDAD Y EL RESULTADO DE LA PRUEBA.

	A	NO A
B	A Y B	NO A Y B
NO B	A Y NO B	NO A Y NO B

VAMOS A ENCONTRAR LA PROBABILIDAD DE CADA CASO EN LA TABLA:

	A	NO A	SUMA
B	P(A Y B)	P(NO A Y B)	P(B)
NO B	P(A Y NO B)	P(NO A Y NO B)	P(NO B)
	P(A)	P(NO A)	1

LAS PROBABILIDADES DE LOS MÁRGENES SE CALCULAN SUMANDO FILAS Y COLUMNAS.

AHORA VAMOS A CALCULAR:

$$P(A \text{ Y } B) = P(B \mid A)\,P(A) = (0{,}99)(0{,}001) = 0{,}00099$$

$$P(\text{NO } A \text{ Y } B) = (B \mid \text{NO } A)\,P(\text{NO } A) = (0{,}02)(0{,}999) = 0{,}01998$$

Y ASÍ PODEMOS RELLENAR ALGUNAS ENTRADAS:

	A	NO A	SUMA
B	0,00099	0,01998	0,02097
NO B	P(A Y NO B)	P(NO A Y NO B)	P(NO B)
	0,001	0,999	1

Y ENCONTRAMOS LAS OTRAS PROBABILIDADES RESTANDO EN CADA COLUMNA Y DESPUÉS SUMANDO CADA FILA.

LA TABLA FINAL ES:

	A	NO A		
B	0,00099	0,01998	0,02097	P(B)
NO B	0,00001	0,97902	0,97903	P(NO B)
	0,001	0,999	1	
	P(A)	P(NO A)		

Y DE AHÍ DEDUCIMOS DIRECTAMENTE QUE

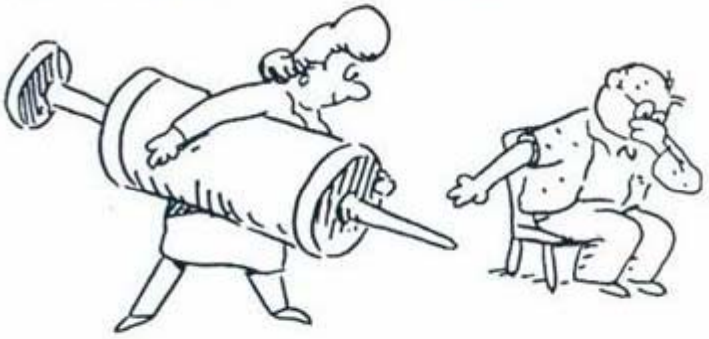
$$P(A|B) = \frac{P(A \text{ Y } B)}{P(B)} = \frac{0{,}0009}{0{,}0209} = \mathbf{0{,}0472}$$

A PESAR DE LA GRAN PRECISIÓN DE LA PRUEBA, **MENOS DE UN 5%** DE LOS QUE DAN POSITIVO PADECEN LA ENFERMEDAD. A ESTO SE LE LLAMA LA **PARADOJA DEL FALSO POSITIVO.**

ESTA TABLA MUESTRA LO QUE PASA EN UN GRUPO DE MIL PACIENTES. COMO MEDIA, SÓLO 21 PACIENTES DARÁN POSITIVO (¡Y SÓLO UNO PADECERÁ LA ENFERMEDAD!), Y 20 FALSOS POSITIVOS SE ENCONTRARÁN EN EL GRAN GRUPO DE NO AFECTADOS.

	ENFERMEDAD	NO ENFERMEDAD	
RESULTADO POSITIVO	1	20	21
RESULTADO NEGATIVO	0	979	979
	1	999	1.000

¿QUÉ DEBE HACER LA DOCTORA? JOE BAYES LE ACONSEJA NO EMPEZAR EL TRATAMIENTO BASÁNDOSE SÓLO EN ESA ÚNICA PRUEBA. SIN EMBARGO, LA PRUEBA APORTA CIERTA INFORMACIÓN: CON UN RESULTADO POSITIVO, LAS PROBABILIDADES DE QUE EL PACIENTE PADEZCA LA ENFERMEDAD HAN AUMENTADO DE 1 ENTRE 1.000 A 1 ENTRE 23. LA DOCTORA CONTINÚA HACIENDO OTRAS PRUEBAS.



JOE BAYES COBRA EL CHEQUE POR LA CONSULTA ANTES DE CONFESAR QUE TODOS ESOS PASOS QUE HA DADO SE PUEDEN COMPRIMIR EN UNA SOLA FÓRMULA, EL TEOREMA DE BAYES:

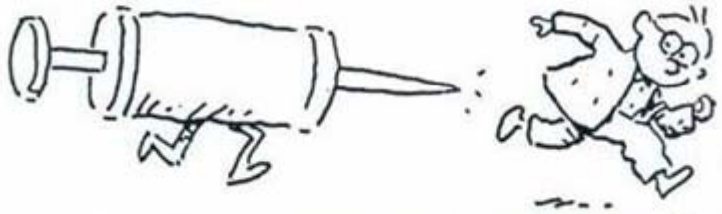
$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A)+P(\text{NO } A)P(B \mid \text{NO } A)}$$

ESTA FÓRMULA CALCULA P(A|B) A PARTIR DE P(A) Y LAS DOS PROBABILIDADES CONDICIONADAS P(B|A) Y P(B|NO A). SE PUEDE CALCULAR TENIENDO EN CUENTA QUE ESA GRAN FRACCIÓN TAMBIÉN SE PUEDE EXPRESAR COMO

$$\frac{P(A \text{ y } B)}{P(A \text{ y } B)+P(\text{NO } A \text{ y } B)} = \frac{P(A \text{ y } B)}{P(B)} = P(A \mid B)$$

EN ESTE CAPÍTULO, HEMOS HABLADO DE LOS ASPECTOS ESENCIALES DE LA PROBABILIDAD: SU DEFINICIÓN, LOS ESPACIOS MUESTRALES Y LOS RESULTADOS ELEMENTALES, LA PROBABILIDAD CONDICIONADA Y ALGUNAS FÓRMULAS BÁSICAS PARA CALCULAR LAS PROBABILIDADES. HEMOS ILUSTRADO ESTAS IDEAS CON UN ESPACIO MUESTRAL DE DOS DADOS. PARA EL JUGADOR MODERNO, LA PROBABILIDAD ES UNA PODEROSA HERRAMIENTA DE ELECCIÓN.

POR ÚLTIMO, EN EL EJEMPLO MÉDICO, HEMOS ENSEÑADO CÓMO ESTAS IDEAS ABSTRACTAS PUEDEN AYUDAR A TOMAR BUENAS DECISIONES CUANDO SE TIENE **INFORMACIÓN IMPERFECTA** Y **RIESGOS REALES**: EL OBJETIVO ÚLTIMO DE LA ESTADÍSTICA.



PERO ESTO NO ES MÁS QUE EL PRINCIPIO. PARA NOSOTROS, LA PROBABILIDAD ES SÓLO UNA **HERRAMIENTA** (UNA HERRAMIENTA ESENCIAL, POR SUPUESTO) PARA EL ESTUDIO DE LA ESTADÍSTICA. EN LOS CAPÍTULOS SIGUIENTES, EXPLORAREMOS LA SUTIL RELACIÓN ENTRE LA PROBABILIDAD, LAS VARIACIONES EN LOS DATOS ESTADÍSTICOS Y NUESTRA CONFIANZA EN LA INTERPRETACIÓN DEL SENTIDO DE LAS OBSERVACIONES QUE HAGAMOS.

♦ Capítulo 4 ♦
VARIABLES ALEATORIAS

EN EL CAPÍTULO 2, VIMOS QUE LAS OBSERVACIONES BASADAS EN DATOS NUMÉRICOS, COMO LOS PESOS DE LOS ESTUDIANTES, SE PUEDEN REPRESENTAR MEDIANTE GRÁFICOS Y RESUMIR EN TÉRMINOS DE PUNTO MEDIO, DISPERSIÓN, OBSERVACIONES ATÍPICAS, ETC. EN EL CAPÍTULO 3, HEMOS VISTO CÓMO SE PUEDEN ASIGNAR PROBABILIDADES A LOS RESULTADOS DE UN EXPERIMENTO ALEATORIO.

SI IMAGINAMOS QUE UN EXPERIMENTO ALEATORIO SE REPITE MUCHAS VECES, ESPERAMOS QUE LOS RESULTADOS ACABEN OBEDECIENDO A SUS PROBABILIDADES. LA PROBABILIDAD CONFORMA UN MODELO PARA EXPERIMENTOS REALES... ASÍ QUE, ¿POR QUÉ NO HACEMOS CON EL MODELO LO QUE YA HEMOS HECHO CON LOS DATOS QUE DESCRIBE?

LA IDEA PRINCIPAL ES LA **VARIABLE ALEATORIA**, QUE ESCRIBIMOS CON UNA

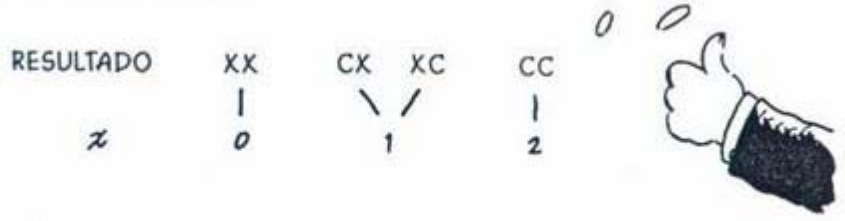
X

MAYÚSCULA.

UNA **VARIABLE ALEATORIA** SE DEFINE COMO EL **RESULTADO NUMÉRICO DE UN EXPERIMENTO ALEATORIO.**

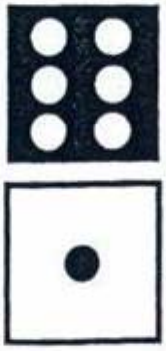
POR EJEMPLO, IMAGINEMOS QUE ESCOGEMOS A UN ESTUDIANTE AL AZAR DE TODO EL GRUPO. ÉSE ES EL EXPERIMENTO ALEATORIO. **ALTURA, PESO, INGRESOS FAMILIARES, NOTA DE SELECTIVIDAD Y NOTA MEDIA** SERÍAN LAS **VARIABLES** NUMÉRICAS QUE DESCRIBEN A ESTE ESTUDIANTE. TODAS ELLAS SON **VARIABLES ALEATORIAS.**

OTRO EJEMPLO: LANZA DOS MONEDAS (EL EXPERIMENTO ALEATORIO) Y ANOTA EL **NÚMERO** DE CARAS: 0, 1, O 2.



¡CUIDADO CON LA NOTACIÓN! LA VARIABLE SE ESCRIBE CON UNA X MAYÚSCULA. LA FILA INFERIOR, x, REPRESENTA UN SOLO VALOR DE X. POR EJEMPLO $x = 2$, SI SALEN DOS CARAS.

OTRO EJEMPLO SERÍA EL DE LA FAMOSA TIRADA DE DADOS. Y REPRESENTA LA SUMA DE LOS PUNTOS DE LOS DOS DADOS. EN ESTA VARIABLE ALEATORIA, Y PUEDE SER CUALQUIER NÚMERO ENTRE 2 Y 12.



$y = 7$

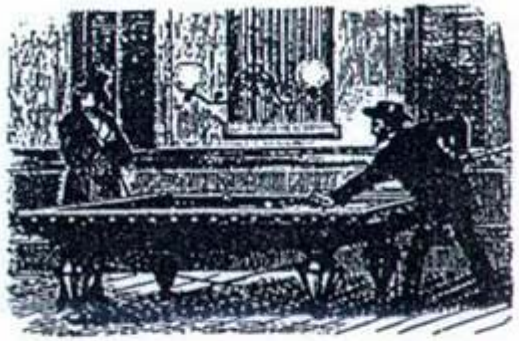
AHORA QUEREMOS SABER LAS **PROBABILIDADES** DE LOS RESULTADOS. PARA LA PROBABILIDAD DE QUE LA VARIABLE X TENGA EL VALOR x, ESCRIBIMOS $Pr(X = x)$, O SIMPLEMENTE $P(x)$. CON LA VARIABLE X DEL LANZAMIENTO DE LA MONEDA PODEMOS CONFECCIONAR UNA TABLA:

x	0	1	2
$Pr(X=x)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

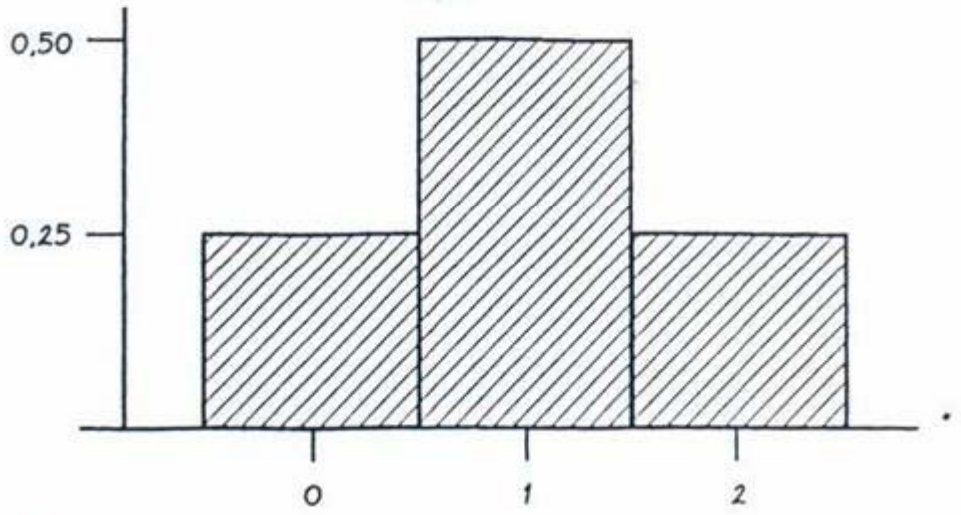
ESTA TABLA SE LLAMA LA **DISTRIBUCIÓN DE PROBABILIDAD** DE LA VARIABLE ALEATORIA X.

PARA LA VARIABLE ALEATORIA Y (LA SUMA DE LOS DOS DADOS), LA DISTRIBUCIÓN DE PROBABILIDAD ES ASÍ:

y	2	3	4	5	6	7	8	9	10	11	12
$Pr(Y=y)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

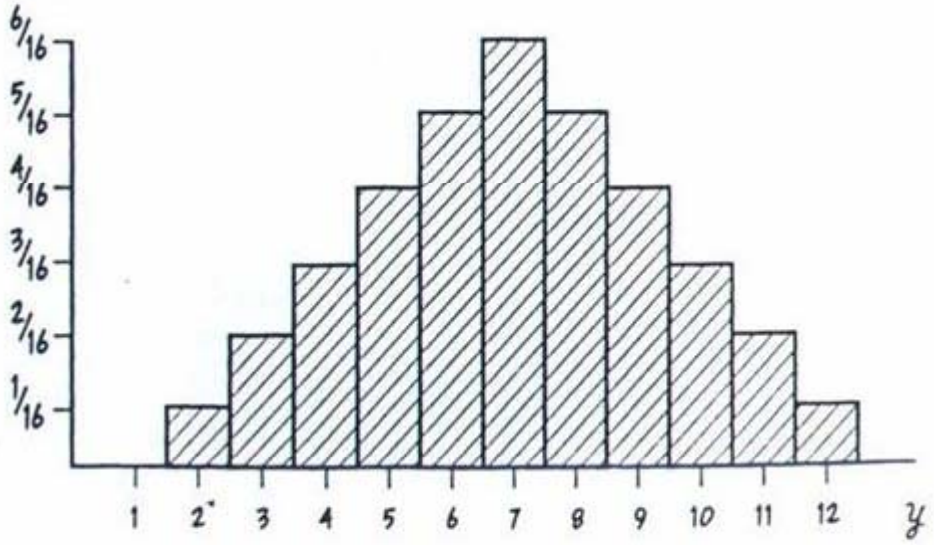


AHORA VAMOS A DIBUJAR GRÁFICOS, O HISTOGRAMAS, PARA REPRESENTAR ESTAS DISTRIBUCIONES DE PROBABILIDAD. POR CADA VALOR DE X, DIBUJAMOS UNA BARRA CON LA ALTURA IGUAL A $p(x)$.



ES FÁCIL VER QUE EL ÁREA TOTAL DE LAS CAJAS ES 1: TODAS LAS CAJAS TIENEN BASE 1 Y ALTURA $p(x)$, ASÍ QUE EL ÁREA TOTAL ES LA SUMA DE LAS PROBABILIDADES DE TODOS LOS RESULTADOS, ES DECIR, 1.

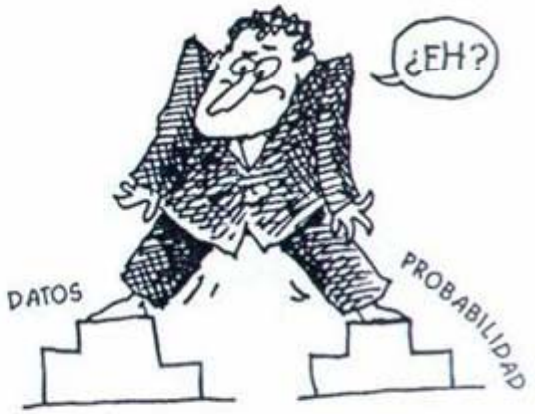
ÉSTE ES EL HISTOGRAMA DE LA PROBABILIDAD DE LA VARIABLE ALEATORIA Y, Y MUESTRA LA DISTRIBUCIÓN DE PROBABILIDAD DE LA SUMA DE DOS DADOS:



¿POR QUÉ LLAMAMOS A TODOS ESTOS GRÁFICOS HISTOGRÀMAS? SEGURO QUE RECUERDAS QUE, EN EL CAPÍTULO 2, UN HISTOGRAMA ERA UN GRÁFICO EN EL QUE SE REPRESENTABA CUÁNTOS DATOS PERTENECÍAN A CADA INTERVALO:

A PARTIR DE ESTE HISTOGRAMA DE **FRECUENCIAS**, DESARROLLAMOS EL HISTOGRAMA DE LA FRECUENCIA **RELATIVA**, EN EL QUE SE VEÍA LA **PROPORCIÓN** DE DATOS QUE TENÍA CADA INTERVALO:

PERO TAMBIÉN RECORDARÁS QUE, SEGÚN UNA DE SUS DEFINICIONES, LA PROBABILIDAD ES LA **FRECUENCIA RELATIVA DE UN SUCESO «A LARGO PLAZO»**. SI REPETIMOS EL EXPERIMENTO ALEATORIO MUCHAS VECES, EL HISTOGRAMA DE LA **FRECUENCIA RELATIVA** DE LOS RESULTADOS DEBERÍA PARECERSE MUCHO AL HISTOGRAMA DE **PROBABILIDAD** DE LA VARIABLE ALEATORIA.

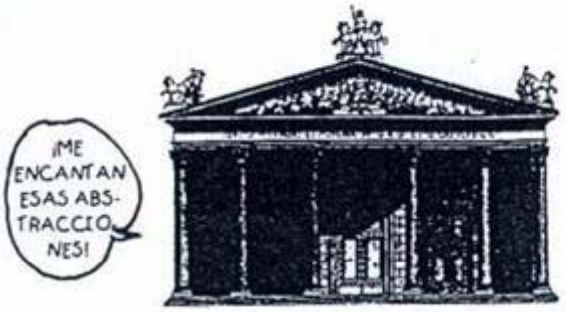


YA CONOCEMOS LA DISTRIBUCIÓN DE PROBABILIDAD DE X, Y TAMBIÉN QUE LAS TIRADAS DE LAS MONEDAS SE CORRESPONDERÁN MÁS O MENOS CON LAS PROBABILIDADES. DESPUÉS DE 1.000 TIRADAS, LA LANZADORA LOCA HACE UN RECUENTO DE LOS DATOS:

MODELO DE PROBABILIDAD		DATOS OBSERVADOS	
$p(x)$	x	n_x = NÚMERO DE OCURRENCIAS	$\frac{n_x}{n}$ = FRECUENCIA RELATIVA
0,25	0	260	0,260
0,5	1	517	0,517
0,25	2	223	0,223

Y ASÍ VEMOS QUE EL HISTOGRAMA DE PROBABILIDAD DE X ES COMO LA «FORMA PURA», O MODELO DEL HISTOGRAMA DE FRECUENCIA RELATIVA DE LOS DATOS.

PARA EXTENDER LA ANALOGÍA ENTRE LA FRECUENCIA RELATIVA Y LOS DATOS, DEBERÍAMOS HABLAR AHORA DE LA MEDIA Y LA VARIANZA (O DESVIACIÓN TÍPICA) DE UNA DISTRIBUCIÓN DE PROBABILIDAD...



Y PARA RECORDAR QUE NOS ENCONTRAMOS EN EL REINO DE LO ABSTRACTO, VAMOS A SOLTAR UNAS CUANTAS **LETRAS GRIEGAS...**

MEDIA Y VARIANZA DE LAS VARIABLES ALEATORIAS

UTILIZAMOS TERMINOLOGÍA Y SÍMBOLOS ESPECIALES PARA DISTINGUIR LAS PROPIEDADES DE LOS CONJUNTOS DE DATOS DE LAS DISTRIBUCIONES DE PROBABILIDAD:

LAS PROPIEDADES DE LOS DATOS SE LLAMAN PROPIEDADES MUESTRALES O ESTADÍSTICOS, MIENTRAS QUE LAS PROPIEDADES DE LA DISTRIBUCIÓN DE PROBABILIDAD SE LLAMAN PARÁMETROS DEL MODELO O POBLACIONALES. PARA LA MEDIA POBLACIONAL UTILIZAMOS LA LETRA GRIEGA μ (MU), Y σ (SIGMA MINÚSCULA) PARA LA DESVIACIÓN TÍPICA POBLACIONAL. (PARA LOS DATOS, UTILIZAMOS LOS SÍMBOLOS ROMANOS $\bar{x}$ Y s.)

LA MEDIA MUESTRAL SE DEFINÍA CON LA ECUACIÓN

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

PUEDE QUE ALGUNOS DE ESTOS DATOS x_i TENGAN VALORES IGUALES. ACUÉRDATE DE LA LANZADORA LOCA DE MONEDAS: LOS ÚNICOS VALORES POSIBLES ERAN 0, 1 Y 2, Y EFECTUÓ 1.000 TIRADAS. EL VALOR 0 RESULTÓ EN 260 OCASIONES, UNA CARA EN 517, Y DOS CARAS EN 223 OCASIONES.

YA QUE x PUEDE TENER TODOS LOS VALORES DE X, n_x ES EL NÚMERO DE DATOS CON VALOR x. ENTONCES PODEMOS REESCRIBIR LA FÓRMULA COMO

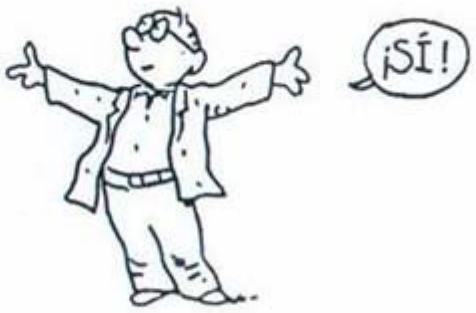
$$\bar{x} = \frac{1}{n}\sum_{\text{TODAS LAS } x} n_x x$$

O COMO

$$\bar{x} = \sum_{\text{TODAS LAS } x} x\frac{n_x}{n}$$

¡AH! PERO AHORA $\frac{n_x}{n}$ ES LA FRECUENCIA RELATIVA... LA «PROBABILIDAD APROXIMADA»... EL NÚMERO QUE SE ACERCA A $p(x)$... LUEGO, POR ANALOGÍA, FORMAMOS LA EXPRESIÓN

$$\sum_{\text{TODAS LAS } x} x\,p(x)$$



Y LA DEFINIMOS COMO MEDIA DE LA DISTRIBUCIÓN DE PROBABILIDAD.

DEFINICIÓN:
LA **media** DE LA VARIABLE ALEATORIA X SE DEFINE COMO

$$\mu = \sum_{\text{TODAS LAS } x} x p(x)$$

A ESTO SE LE LLAMA EL VALOR ESPERADO DE X, O E[X]. IMAGINA QUE ES LA SUMA DE TODOS LOS VALORES POSIBLES, CADA UNO PONDERADO POR SU PROBABILIDAD.

EL EXPERIMENTO DE LA LANZADORA LOCA DE MONEDAS NOS PERMITE COMPARAR SU MEDIA MUESTRAL $\bar{x}$ CON NUESTRA MEDIA POBLACIONAL μ:

MUESTRA

x	$\frac{n_x}{n}$	$x\frac{n_x}{n}$
0	0,26	0
1	0,517	0,517
2	0,223	0,446
		$0{,}963 = \bar{x}$

MODELO

x	$p(x)$	$xp(x)$
0	0,25	0
1	0,5	0,5
2	0,25	0,5
		$1 = \mu$

AHORA VAMOS A HACER LO MISMO CON LA VARIANZA. A LO MEJOR RECUERDAS LA FÓRMULA

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

MIDE (CASI) LA DISTANCIA CUADRÁTICA MEDIA ENTRE LOS DATOS Y LA MEDIA. IGUAL QUE ANTES, LA PODEMOS REFORMULAR:

$$s^2 = \sum_{\text{TODAS LAS } x}(x - \bar{x})^2 \frac{n_x}{n-1}$$

SALVO ESE MOLESTO DENOMINADOR $n-1$ EN LUGAR DE n, ESTA FÓRMULA TAMBIÉN PARECE UNA SUMA PONDERADA DE DISTANCIAS AL CUADRADO... ASÍ QUE FORMULAMOS OTRA DEFINICIÓN:

LA **varianza** DE UNA VARIABLE ALEATORIA X ES LA ESPERADA DEL CUADRADO DE LA DISTANCIA ENTRE LOS POSIBLES VALORES DE X Y LA MEDIA POBLACIONAL:

$$\sigma^2 = \sum_{\text{TODAS LAS } x} (x-\mu)^2 p(x)$$

LA **desviación típica** σ ES LA RAÍZ CUADRADA DE LA VARIANZA.

AHORA UTILIZAMOS LA TABLA DE LA PÁGINA ANTERIOR PARA ENCONTRAR LA VARIANZA DE UNA TIRADA DE DOS MONEDAS (EN LA QUE $\mu = 1$).

x	$p(x)$	$(x-\mu)^2 p(x)$
0	0,25	$(0-1)^2\ 0,25 = 0,25$
1	0,5	$(1-1)^2\ 0,50 = 0$
2	0,25	$(2-1)^2\ 0,25 = 0,25$
TOTAL		$0,50 = \sigma^2$

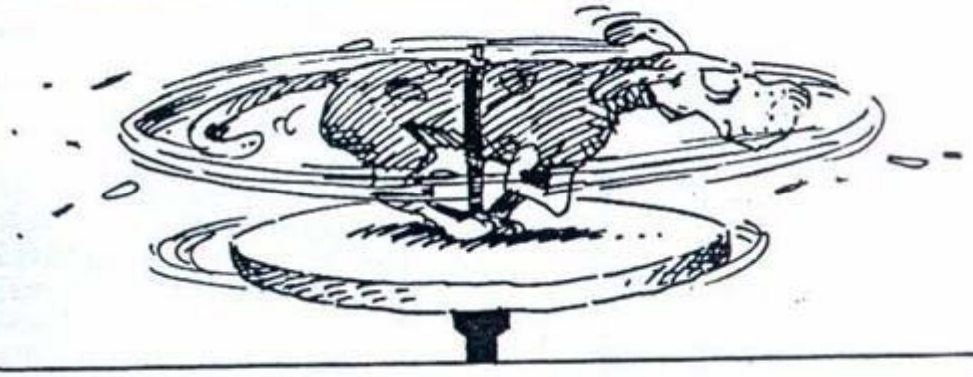
EN RESUMEN: μ Y σ, LA MEDIA Y LA DESVIACIÓN TÍPICA POBLACIONALES, SON PARÁMETROS QUE PODEMOS CALCULAR A PARTIR DE LAS DISTRIBUCIONES DE **PROBABILIDAD**. SON TOTALMENTE ANÁLOGAS A LA MEDIA MUESTRAL $\bar{x}$ Y A LA DESVIACIÓN TÍPICA s DE LOS DATOS MUESTRALES.

HASTA AHORA, NUESTROS EJEMPLOS HAN CONSISTIDO EN VARIABLES ALEATORIAS DISCRETAS. SUS RESULTADOS SON UN CONJUNTO DE VALORES AISLADOS («DISCRETOS»), COMO LOS QUE VIMOS EN EL CAPÍTULO 3. PERO TAMBIÉN HAY

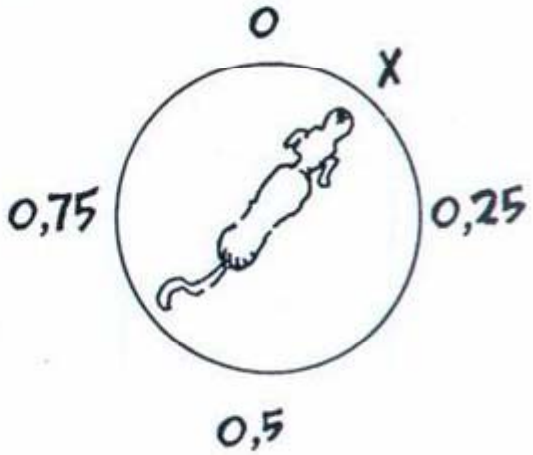
variables aleatorias continuas

IMAGINEMOS UN EXPERIMENTO ALEATORIO EN EL QUE TODOS LOS RESULTADOS TENGAN PROBABILIDAD CERO. ESO ES, $P(x) = 0$ PARA CUALQUIER x.

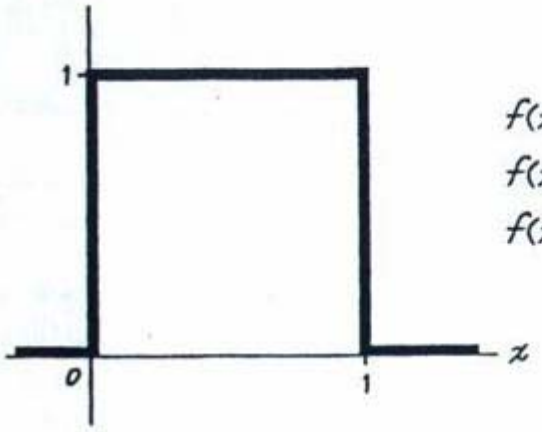
UN EJEMPLO MUY SIMPLE ES EL DE UN PERRO DE CAZA SOBRE UNA SUPERFICIE CIRCULAR QUE GIRA EN EQUILIBRIO. PUEDE PARAR EN CUALQUIER PUNTO DEL CÍRCULO. SI X REPRESENTA LA PROPORCIÓN DE TODA LA CIRCUNFERENCIA EN LA QUE SE ENCUENTRA, LA VARIABLE ALEATORIA X PUEDE TENER CUALQUIER VALOR ENTRE 0 Y 1: UNA SERIE INFINITA DE VALORES.



ALGUNAS PROBABILIDADES SON FÁCILES DE ENCONTRAR, COMO LA PROBABILIDAD DE QUE X ESTÉ DENTRO DE UNA REGIÓN: POR EJEMPLO, $P(0{,}25 \leq X \leq 0{,}75) = 0{,}5$, PORQUE ES LA MITAD DEL CÍRCULO. SIN EMBARGO, ¿QUÉ PASA CON $P(X = 0{,}5)$? YA QUE X PUEDE REPRESENTAR UN NÚMERO INFINITO DE VALORES, Y TODOS SON IGUAL DE POSIBLES, LA PROBABILIDAD DE QUE X SEA EXACTAMENTE 0,5 (O CUALQUIER OTRO VALOR EXACTAMENTE) ES PRECISAMENTE 0.



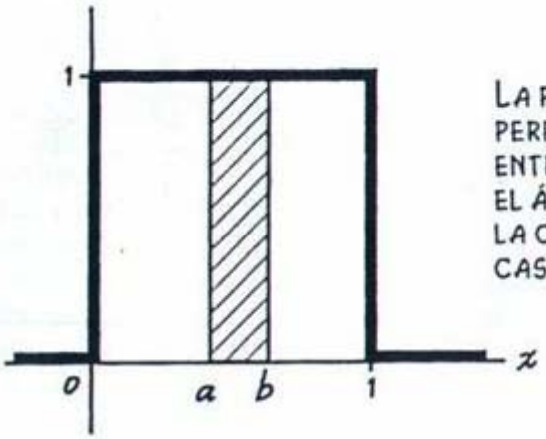
¿CÓMO PODEMOS REPRESENTARLO EN UN DIBUJO? POR ANALOGÍA CON EL CASO DE LAS PROBABILIDADES DISCRETAS, INTENTAMOS OBSERVAR LAS PROBABILIDADES CONTINUAS COMO **ÁREAS BAJO ALGO**. EN EL CASO DEL PERRO GIRATORIO, ESE «ALGO» TIENE ESTE ASPECTO:



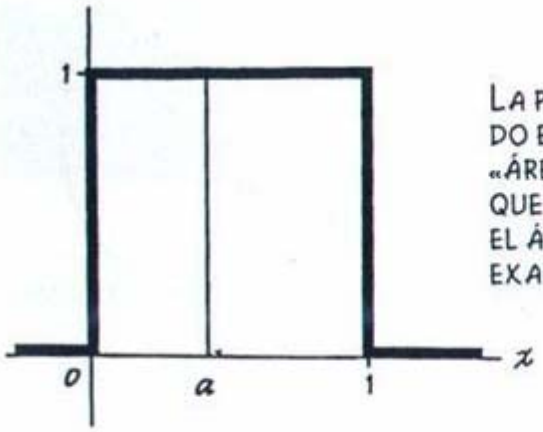
$f(x) = 0$ CUANDO $x < 0$

$f(x) = 1$ CUANDO $0 \leq x \leq 1$

$f(x) = 0$ CUANDO $x > 1$



LA PROBABILIDAD DE QUE EL PERRO SEÑALE CUALQUIER LUGAR ENTRE a Y b ES PRECISAMENTE EL ÁREA SOMBREADA BAJO LA CURVA ENTRE a Y b (EN ESTE CASO, $b - a$).

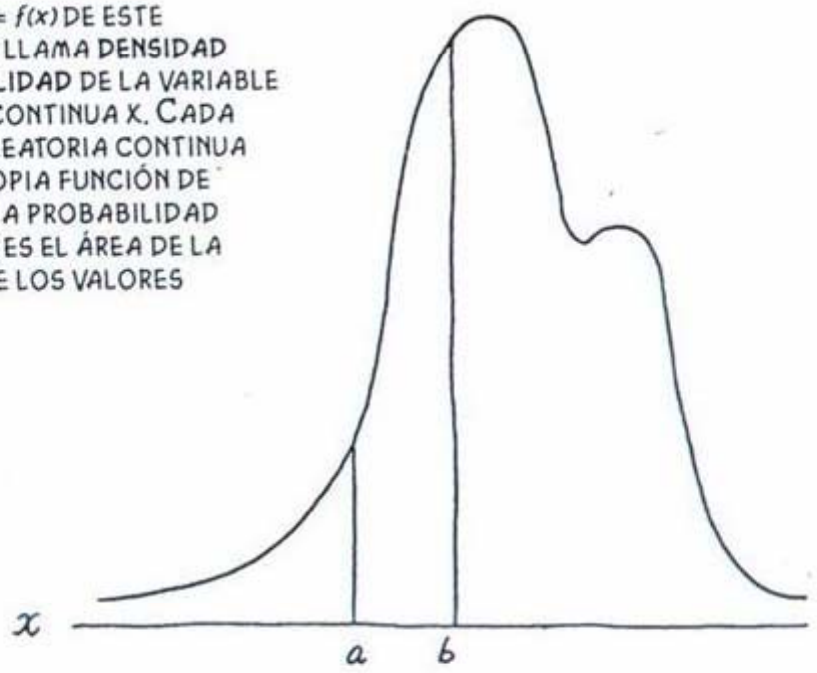


LA PROBABILIDAD DE UN RESULTADO EXACTO, SIN EMBARGO, ES EL «ÁREA» QUE HAY SOBRE UN PUNTO, QUE ES CERO. (TEN EN CUENTA QUE EL ÁREA TOTAL DE LA CURVA ES EXACTAMENTE 1.)

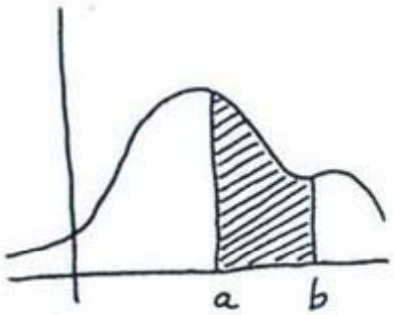
ESE MISMO DIBUJO DESCRIBE EL **GENERADOR DE NÚMEROS ALEATORIOS** QUE TIENEN CASI TODOS LOS ORDENADORES Y MUCHAS CALCULADORAS. SI APRIETAS UN BOTÓN, SALE UN NÚMERO ENTRE 0 Y 1; Y TODOS LOS NÚMEROS SON IGUAL DE PROBABLES, IGUAL QUE CON EL PERRO.

PERO, POR DESGRACIA, NO SON TOTALMENTE ALEATORIOS. ESTÁN GENERADOS POR ALGÚN ALGORITMO, ASÍ QUE, PARA SER PRECISOS, LOS LLAMAMOS NÚMEROS **PSEUDOALEATORIOS.**

LA CURVA $y = f(x)$ DE ESTE EJEMPLO SE LLAMA **DENSIDAD DE PROBABILIDAD** DE LA VARIABLE ALEATORIA CONTINUA X. CADA VARIABLE ALEATORIA CONTINUA TIENE SU PROPIA FUNCIÓN DE DENSIDAD. LA PROBABILIDAD $P(a \leq X \leq b)$ ES EL ÁREA DE LA CURVA ENTRE LOS VALORES DE x, a Y b.



EN GENERAL, LA DENSIDAD DE PROBABILIDAD NO ES TAN SIMPLE, Y A VECES, CALCULAR EL ÁREA NO TIENE NADA DE TRIVIAL.



$$\int_a^b f(x)\,dx$$

NOS VEMOS OBLIGADOS A UTILIZAR NOTACIÓN DE CÁLCULO PARA DESCRIBIR EL ÁREA DE LA CURVA $f(x)$. ESTE SÍMBOLO SE LEE «INTEGRAL DE f DESDE a HASTA b».

AL IGUAL QUE LAS PROBABILIDADES DISCRETAS, LAS DENSIDADES CONTINUAS TIENEN DOS PROPIEDADES QUE YA CONOCEMOS:

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

(INTENTA QUE NO TE ASUSTEN ESOS INFINITOS... SÓLO QUIEREN DECIR QUE OBSERVAMOS TODA EL ÁREA DE LA CURVA, DEL PRINCIPIO AL FINAL, ¡SÓLO QUE NO HAY NI PRINCIPIO NI FINAL!)

A PESAR DE QUE LA NOTACIÓN TE RESULTE EXTRAÑA, NO REPRESENTA MÁS QUE UN ÁREA... EL SIGNO DE LA INTEGRAL ES UNA «S» ALARGADA, DE «SUMA», QUE ES MÁS O MENOS LA FUNCIÓN QUE DESEMPEÑA LA INTEGRAL.

COMO ES ALGO PARECIDO A UNA SUMA, LA INTEGRAL SIRVE PARA DEFINIR LA

MEDIA Y LA VARIANZA de una variable aleatoria continua.

$$\mu = \int_{-\infty}^{\infty} x f(x)\,dx$$

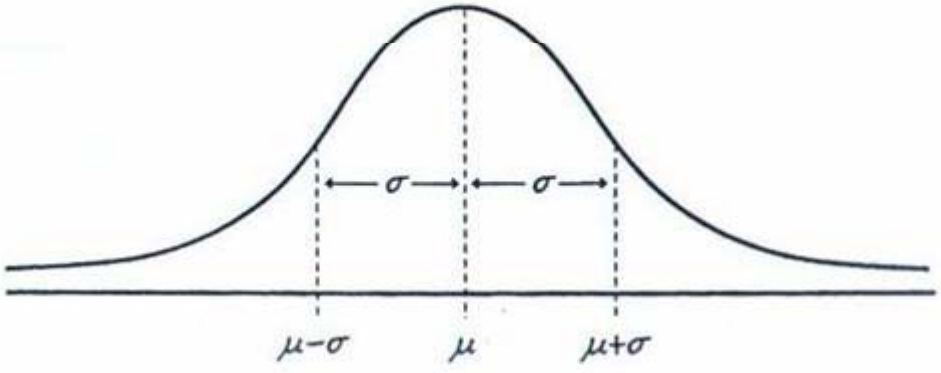
$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx$$

POR ANALOGÍA CON LAS FÓRMULAS DISCRETAS:

$$\mu = \sum_{\text{TODAS LAS } x} x p(x)$$

$$\sigma^2 = \sum_{\text{TODAS LAS } x} (x-\mu)^2 p(x)$$

AUNQUE NO RESULTE OBVIO AL OBSERVAR LAS FÓRMULAS, ESTAS DEFINICIONES DE MEDIA Y VARIANZA SON TOTALMENTE COHERENTES CON SU PAPEL DE CENTRO Y DISPERSIÓN MEDIA DE LAS PROBABILIDADES DADAS POR LA DENSIDAD $f(x)$. ÉSTE ES EL GRÁFICO QUE HAY QUE RECORDAR:



SUMA
de variables aleatorias

UNA VEZ CONOCIDA LA MEDIA Y LA VARIANZA DE UNA VARIABLE ALEATORIA, ¿QUÉ PODEMOS HACER CON ELLAS? BUENO, PARA EMPEZAR, SE PUEDEN BUSCAR LA MEDIA Y LA VARIANZA DE **OTRAS** VARIABLES ALEATORIAS...

POR EJEMPLO, VAMOS A TOMAR EL CASO DEL LANZAMIENTO DE UNA MONEDA. SI SALE CARA, $x = 1$, Y $x = 0$ SI SALE CRUZ.

x	0	1
$p(x)$	0,5	0,5

AHORA DEBERÍAS SER CAPAZ DE ENCONTRAR LA MEDIA

$$E[X] = 0 \cdot p(0) + 1 \cdot p(1)$$
$$= 0 + 0{,}5$$
$$= 0{,}5$$

Y LA VARIANZA

$$\sigma^2 = (0-0{,}5)^2 p(0) + (1-0{,}5)^2 p(1)$$
$$= 0{,}25$$

VAMOS A HACER UNA APUESTA: TE JUEGAS 6 DÓLARES Y YO LANZO UNA MONEDA: SI SALE CARA, GANAS 10 DÓLARES, Y CERO SI SALE CRUZ. ENTONCES, TUS GANANCIAS G SON

$$G = 10X - 6$$

¡UNA NUEVA VARIABLE ALEATORIA! ¿CUÁLES SON SU MEDIA Y SU VARIANZA?

SI LO PIENSAS UN POCO TE CONVENCERÁS DE QUE E[G] VIENE DADO POR

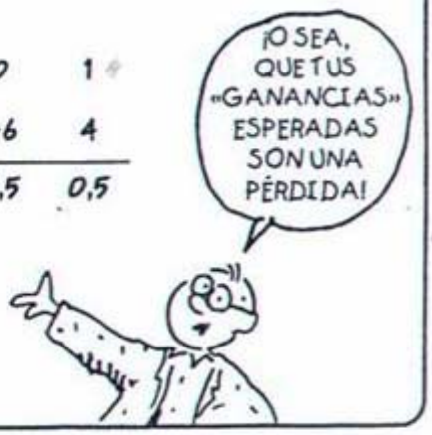
$$E[G] = E[10X - 6]$$
$$= 10E[X] - 6$$

QUE RESULTA EN

$$10(0,5) - 6 = -1$$

PUEDES COMPROBARLO CON ESTA TABLA:

x	0	1
g	−6	4
$p(g)$	0,5	0,5

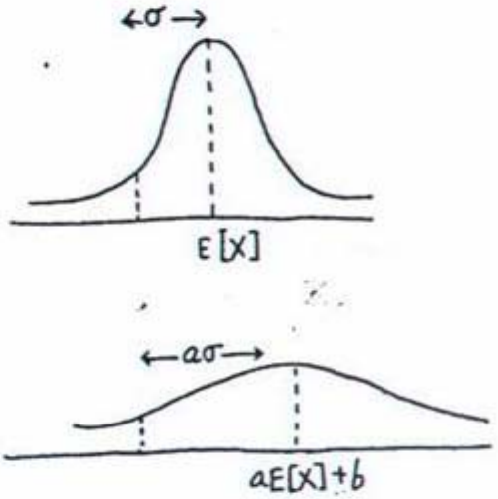


EN GENERAL, NO ES DIFÍCIL DEMOSTRAR QUE

$$E[aX+b] = aE[X] + b$$

CUANDO a Y b SON CUALQUIER NÚMERO Y X ES CUALQUIER VARIABLE ALEATORIA. EN CUANTO A LA VARIANZA, TAMBIÉN EXISTE UN RESULTADO GENERAL:

$$\sigma^2(aX+b) = a^2\sigma^2(X)$$



EN LA APUESTA ANTERIOR, LOS POSIBLES RESULTADOS SON −6 Y 4. ASÍ QUE ESTÁ CLARO QUE LA VARIANZA DE G TIENE QUE SER MAYOR QUE LA VARIANZA DE X. DE HECHO,

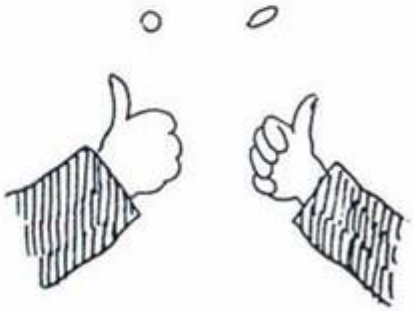
$$\sigma^2(G) = \sigma^2(10X-6)$$
$$= 100\sigma^2(X)$$
$$= 25$$

Y

$$\sigma(G) = 5$$

TAMBIÉN PUEDES SUMAR DOS VARIABLES ALEATORIAS. POR EJEMPLO, SUPÓN QUE LANZAMOS UNA MONEDA DOS VECES. EL NÚMERO DE CARAS DE LOS DOS LANZAMIENTOS ES $X_1 + X_2$, DONDE X_1 Y X_2 SON LAS VARIABLES ALEATORIAS DE LOS RESULTADOS DEL PRIMER Y SEGUNDO LANZAMIENTO.

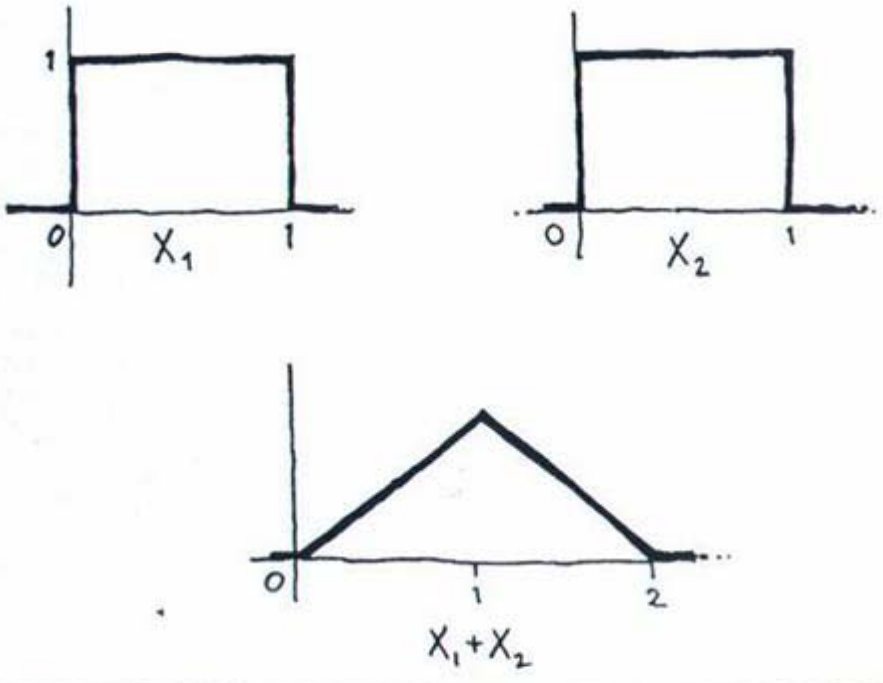
x_1+x_2	0	1	2
$p(x_1+x_2)$	0,25	0,5	0,25



DE NUEVO, ES MUY SENCILLO VER QUE

$$E[X_1+X_2] = E[X_1] + E[X_2]$$

(NO PREGUNTES POR LA DISTRIBUCIÓN DE PROBABILIDAD DE $X_1 + X_2$ PORQUE DEPENDE DE FORMA MUY COMPLICADA DE LAS DOS DISTRIBUCIONES ORIGINALES. POR EJEMPLO, SI TANTO X_1 COMO X_2 SON LA DISTRIBUCIÓN DEL PERRO GIRATORIO, LOS HISTOGRAMAS SE COMPORTARÍAN ASÍ:)



LA VARIANZA DE LA SUMA DE VARIABLES ALEATORIAS TIENE UNA FORMA MUY SIMPLE EN EL CASO ESPECIAL DE QUE X E Y SEAN **INDEPENDIENTES**. LA DEFINICIÓN TÉCNICA DE INDEPENDENCIA SE BASA EN LA PROPIEDAD DE LA PROBABILIDAD P(A Y B) = P(A)P(B). PERO, PARA NOSOTROS, LA INDEPENDENCIA SÓLO SIGNIFICA QUE X E Y ESTÁN GENERADAS POR **MECANISMOS INDEPENDIENTES** COMO EL LANZAMIENTO DE UNA MONEDA, UNA TIRADA DE DADOS, ETC.

CUANDO X E Y SON INDEPENDIENTES, SUS VARIANZAS SE SUMAN:

$$\sigma^2(X+Y) = \sigma^2(X)+\sigma^2(Y)$$

EN EL CASO DEL LANZAMIENTO DE DOS MONEDAS.

$$\sigma^2(X_1+X_2) = \sigma^2(X_1) + \sigma^2(X_2)$$
$$= 0{,}25 + 0{,}25$$
$$= 0{,}5$$

TODO ESTO SE PUEDE GENERALIZAR A LA SUMA DE MUCHAS VARIABLES ALEATORIAS:

$$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i]$$

Y CUANDO TODAS LAS X_i SON INDEPENDIENTES.

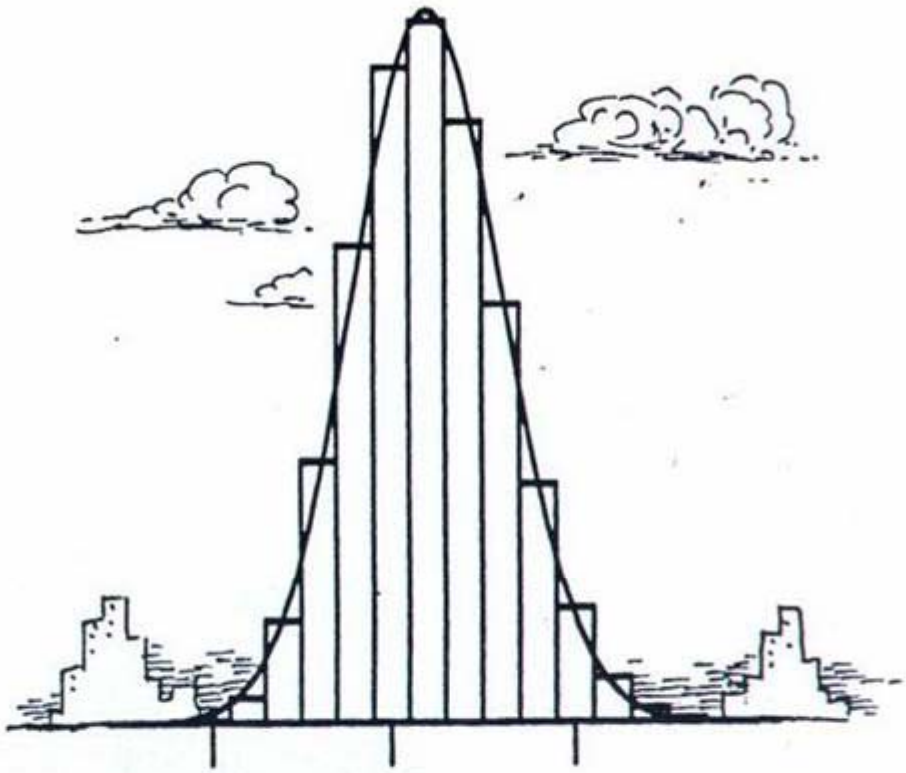
$$\sigma^2(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \sigma^2(X_i)$$

EN EL CAPÍTULO SIGUIENTE VEREMOS DOS IMPORTANTES EJEMPLOS DE VARIABLES ALEATORIAS: UNA, LA **BINOMIAL**, ES LA SUMA DE VARIAS VARIABLES ALEATORIAS INDEPENDIENTES. LA OTRA, LA **NORMAL**, ES UNA VARIABLE ALEATORIA CONTINUA QUE TIENE UNA SORPRENDENTE RELACIÓN CON LA BINOMIAL, Y TAMBIÉN CON CUALQUIER OTRA SUMA DE VARIABLES ALEATORIAS INDEPENDIENTES.

♦ Capítulo 5 ♦

HISTORIA DE DOS DISTRIBUCIONES

AHORA VEREMOS DOS IMPORTANTES EJEMPLOS DE VARIABLES ALEATORIAS, UNA DISCRETA Y OTRA CONTINUA.



EMPEZAREMOS POR LA DISCRETA, LA VARIABLE ALEATORIA BINOMIAL. IMAGINEMOS QUE TENEMOS UN PROCESO ALEATORIO CON TAN SÓLO DOS **POSIBLES RESULTADOS**: CARA O CRUZ, VICTORIA O DERROTA EN UN PARTIDO DE FÚTBOL, PASAR O NO PASAR LA INSPECCIÓN DE LA ITV. DE FORMA ARBITRARIA, A UNO DE ESTOS RESULTADOS LO LLAMAMOS **ÉXITO** Y AL OTRO, **FRACASO**.

LO QUE HACEMOS ES REPETIR EL EXPERIMENTO... EN FIN, REPETIDAS VECES. UN EXPERIMENTO DE ESTE TIPO SE LLAMA

Variable aleatoria de Bernoulli,

SIEMPRE QUE PRESENTE ESTAS PROPIEDADES CRÍTICAS:

1) EL RESULTADO DE CADA PRUEBA PUEDE SER ÉXITO O FRACASO.

2) LA PROBABILIDAD p DE ÉXITO ES LA MISMA EN TODAS LAS PRUEBAS.

3) LAS PRUEBAS SON INDEPENDIENTES: EL RESULTADO DE UNA NO AFECTA A LOS RESULTADOS POSTERIORES.

COMENZAREMOS POR UNA VARIABLE ALEATORIA DE BERNOULLI CON UNA PROBABILIDAD p DE ÉXITO. VAMOS A CONSTRUIR UNA NUEVA VARIABLE ALEATORIA REPITIENDO LA PRUEBA.

La variable aleatoria binomial

x ES EL **NÚMERO DE ÉXITOS** DE LAS PRUEBAS DE BERNOULLI REPETIDAS n VECES, CON UNA PROBABILIDAD p DE ÉXITO.

UN EJEMPLO DE VARIABLE ALEATORIA BINOMIAL ES EL NÚMERO DE CARAS (ÉXITOS) DE DOS LANZAMIENTOS DE UNA SOLA MONEDA. EN ESTE CASO $n = 2$ Y $p = 0{,}5$.

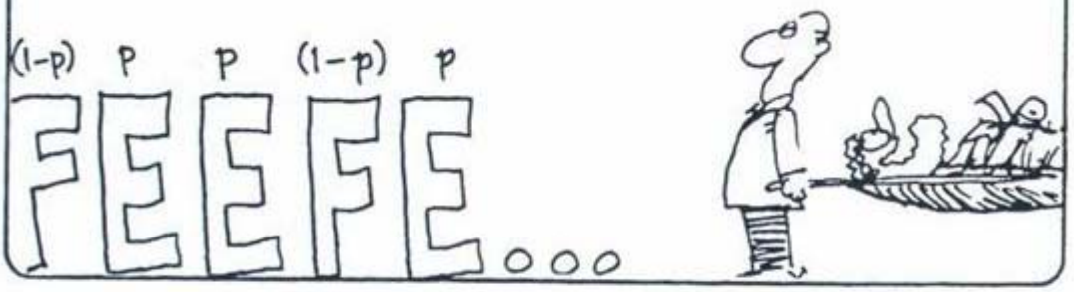
k = NÚMERO DE ÉXITOS	0	1	2
$PR(x = k)$	0,25	0,5	0,25

OTRO EJEMPLO ES LA PRIMERA PARTIDA DE DE MERE: TIRAR UN SOLO DADO CUATRO VECES SEGUIDAS. EL ÉXITO ES CONSEGUIR UN 6. LA DISTRIBUCIÓN ES:

EN GENERAL, ¿CUÁL ES LA DISTRIBUCIÓN DE PROBABILIDAD DE UNA BINOMIAL DE CUALQUIER PROBABILIDAD p Y NÚMERO n DE PRUEBAS? UN SIMPLE CÁLCULO DE LA PROBABILIDAD NOS DA LA RESPUESTA: LA PROBABILIDAD DE OBTENER k ÉXITOS EN n PRUEBAS, $Pr(X = k)$, ES

$$Pr(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

EN ESTE CASO $\binom{n}{k}$, QUE SE LEE «COMBINACIONES DE n ELEMENTOS TOMADOS DE k EN k», ES EL **COEFICIENTE BINOMIAL**. ÉSTE CUENTA TODAS LAS MANERAS POSIBLES DE OBTENER k ÉXITOS EN n PRUEBAS. CADA SECUENCIA INDIVIDUAL DE k ÉXITOS Y $n-k$ FRACASOS TIENE UNA PROBABILIDAD $p^k(1-p)^{n-k}$, SEGÚN LA REGLA DE MULTIPLICACIÓN. EL NÚMERO DE SECUENCIAS ES $\binom{n}{k}$.



LA FÓRMULA DE $\binom{n}{k}$ ES

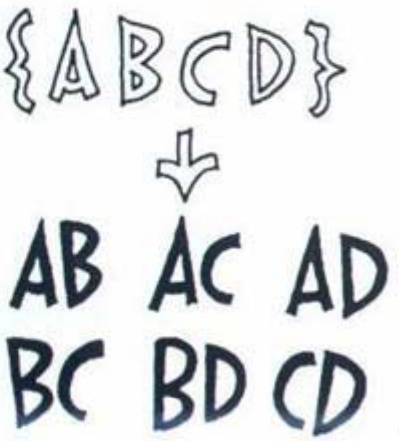
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

EN LA QUE

$$n! = n \times (n-1) \times (n-2) \times \ldots \times 1$$

Y $0!$ SE CONSIDERA 1. POR EJEMPLO, $\binom{4}{2}$, EL NÚMERO DE COMBINACIONES POSIBLES DE ELEGIR DOS LETRAS DE UN CONJUNTO DE CUATRO, ES

$$\binom{4}{2} = \frac{4!}{2!2!} = \frac{24}{4} = 6$$



OTRO PUNTO DE VISTA DE LOS COEFICIENTES BINOMIALES ES EL **TRIÁNGULO DE PASCAL**. CADA ENTRADA ES LA SUMA DE LOS DOS NÚMEROS QUE TIENE ENCIMA.

$\binom{3}{1}$ $\binom{7}{2}$ $\binom{5}{3}$ $\binom{8}{0}$ $\binom{10}{7}$ $\binom{10}{10}$

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1
1 6 15 20 15 6 1
1 7 21 35 35 21 7 1
1 8 28 56 70 56 28 8 1
1 9 36 84 126 126 84 36 9 1
1 10 45 120 210 252 210 120 45 10 1
1 11 55 165 330 462 462 330 165 55 11 1
1 12 66 220 495 792 924 792 495 220 66 12 1

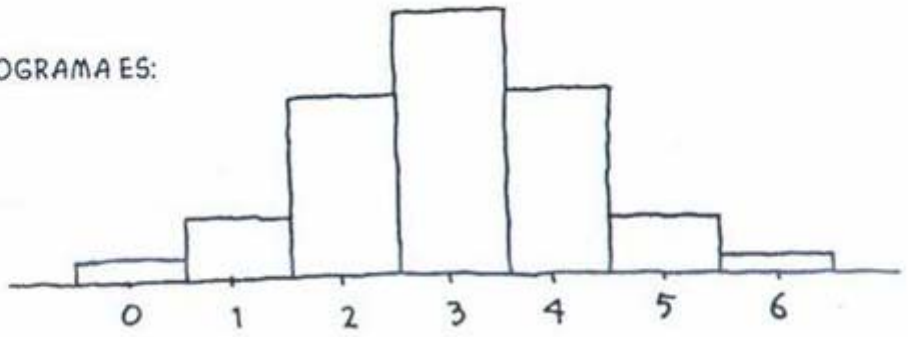
ETC.

PARA ENCONTRAR $\binom{n}{k}$ SÓLO HACE FALTA CONTAR HASTA LA FILA n Y HASTA LA ENTRADA k (SIN OLVIDAR QUÉ HAY QUE EMPEZAR CONTANDO EL CERO).

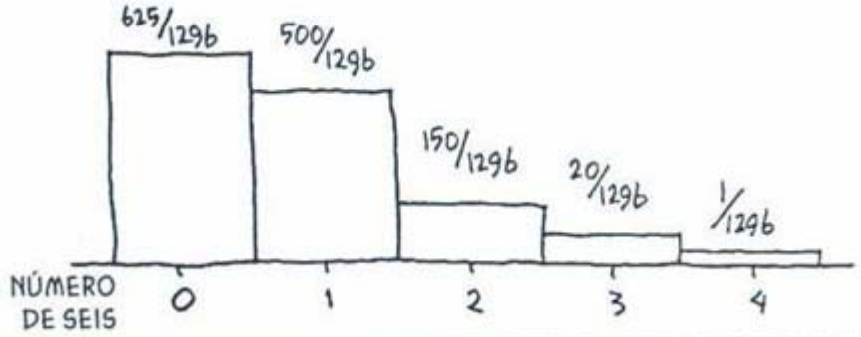
CUANDO $p = 0.5$, LA DISTRIBUCIÓN DE PROBABILIDAD DE LA BINOMIAL ES PERFECTAMENTE SIMÉTRICA. EN 6 LANZAMIENTOS DE UNA MONEDA, POR EJEMPLO, ES

k = # CARAS	0	1	2	3	4	5	6
$Pr(X=k)$	$\left(\frac{1}{2}\right)^6$	$\left(\frac{1}{2}\right)^6 \cdot 6$	$\left(\frac{1}{2}\right)^6 \cdot 15$	$\left(\frac{1}{2}\right)^6 \cdot 20$	$\left(\frac{1}{2}\right)^6 \cdot 15$	$\left(\frac{1}{2}\right)^6 \cdot 6$	$\left(\frac{1}{2}\right)^6$

Y EL HISTOGRAMA ES:



EN LA TIRADA DE CUATRO DADOS DE DE MERE, LA DISTRIBUCIÓN ES MÁS DESPROPORCIONADA:



LA MEDIA Y LA VARIANZA DE LA DISTRIBUCIÓN BINOMIAL SON

$$\mu = np$$

$$\sigma^2 = np(1-p)$$

OBSERVA QUE LA MEDIA, CON UN POCO DE INTUICIÓN, TIENE MUCHO SENTIDO: EN n PRUEBAS DE BERNOULLI, EL NÚMERO DE ÉXITOS QUE SE ESPERA DEBERÍA SER np. LA VARIANZA SE DERIVA DEL HECHO DE QUE LA BINOMIAL ES LA SUMA DE n PRUEBAS DE BERNOULLI INDEPENDIENTES CON UNA VARIANZA $p(1-p)$.

LOS PARÁMETROS DE LA DISTRIBUCIÓN BINOMIAL SON n Y p. TANTO LA DISTRIBUCIÓN COMO LA MEDIA Y LA VARIANZA DEPENDEN SÓLO DE ESOS DOS NÚMEROS. EN LA MAYORÍA DE LIBROS Y PROGRAMAS INFORMÁTICOS APARECEN TABLAS DE DISTRIBUCIÓN BINOMIAL. ÉSTA ES LA TABLA DE $n = 10$.

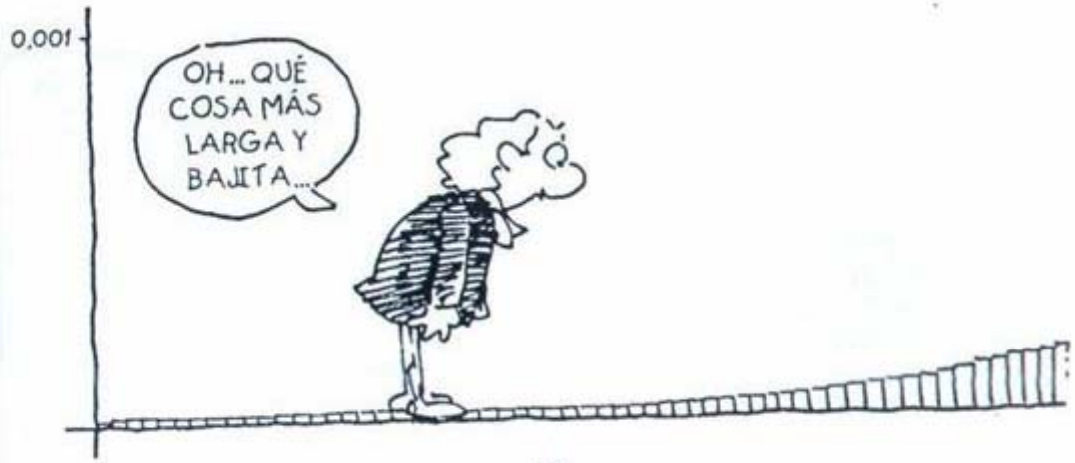
VALORES DE $Pr(X = k)$

	k										
	0	1	2	3	4	5	6	7	8	9	10
0,1	0,349	0,387	0,194	0,057	0,011	0,001	0,000	0,000	0,000	0,000	0,000
0,25	0,056	0,188	0,282	0,250	0,146	0,058	0,016	0,003	0,000	0,000	0,000
p 0,50	0,001	0,010	0,044	0,117	0,205	0,246	0,205	0,117	0,044	0,010	0,001
0,75	0,000	0,000	0,000	0,003	0,016	0,058	0,146	0,250	0,282	0,188	0,056
0,9	0,000	0,000	0,000	0,000	0,000	0,001	0,011	0,057	0,194	0,387	0,349

SIN EMBARGO, HACER ESTOS CÁLCULOS CON VALORES GRANDES DE n PUEDE CONVERTIRSE EN UNA TORTURA... O, AL MENOS LO ERA EN EL SIGLO XVIII, CUANDO **JAMES BERNOULLI** Y **ABRAHAM DE MOIVRE** INTENTABAN HACERLO SIN LA AYUDA DE UN ORDENADOR.

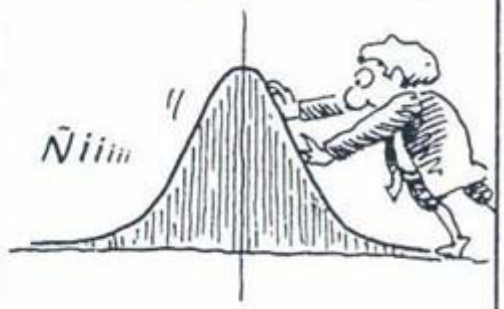
CON UN ARMA DE RECIENTE INVENCIÓN, EL **CÁLCULO**. DE MOIVRE DEMOSTRÓ QUE CUANDO $p = 0{,}5$, LA DISTRIBUCIÓN BINOMIAL SE PODÍA OBTENER APROXIMADAMENTE MEDIANTE UNA FUNCIÓN DE **DENSIDAD CONTINUA**, MUY FÁCIL DE DESCRIBIR.

PARA VER SU FUNCIONAMIENTO, IMAGINEMOS UNA DISTRIBUCIÓN BINOMIAL CON $p = 0{,}5$ Y UN NÚMERO n MUY ELEVADO, POR EJEMPLO, UN MILLÓN...



AHORA, DECÍA DE MOIVRE, SE DESPLAZA EL GRÁFICO HASTA QUE LA MEDIA SEA CERO.

SE COMPRIME LA CURVA A LO LARGO DEL EJE *x* HASTA QUE LA DESVIACIÓN TÍPICA SEA *1*. A LA VEZ QUE SE ESTIRA A LO LARGO DEL EJE *y* PARA QUE EL ÁREA DE LA CURVA SEA IGUAL A *1*.



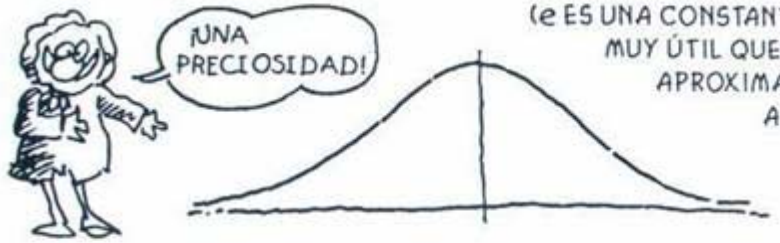
EL RESULTADO SE PARECE MUCHO A UNA CURVA SUAVIZADA, EN FORMA DE **CAMPANA, SIMÉTRICA**, Y DE MOIVRE DEMOSTRÓ QUE VIENE DADA POR UNA FÓRMULA MUY SIMPLE:

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

ESTA FUNCIÓN RECIBE EL NOMBRE DE

distribución normal tipificada.

(*e* ES UNA CONSTANTE MATEMÁTICA MUY ÚTIL QUE EQUIVALE APROXIMADAMENTE A 2,718.)



(CONVÉNCETE DE QUE ESTA FUNCIÓN TIENE UN GRÁFICO EN FORMA DE CAMPANA. PARA VALORES DE *z* ALEJADOS DE CERO, $f(z)$ ES PRÁCTICAMENTE CERO. TIENE UN DENOMINADOR MUY ELEVADO; Y ES SIMÉTRICO, YA QUE $f(z) = f(-z)$. Y TIENE UN MÁXIMO DE $z = 0$.)

ESTA DISTRIBUCIÓN SE LLAMA NORMAL TIPIFICADO* PORQUE TODA ESA COMPRESIÓN Y EXTENSIÓN A LO LARGO DE LOS EJES ESTÁ PENSADA PARA DARLES ESTAS SIMPLES PROPIEDADES, QUE AHORA NOSOTROS PRESENTAMOS SIN PRUEBA ALGUNA:

$$\mu = 0$$

$$\sigma = 1$$

* TAMBIÉN SE LLAMA DISTRIBUCIÓN NORMAL CENTRADA Y REDUCIDA. [N. T.]

PARA RESUMIR LA TEORÍA DE DE MOIVRE, SI «NORMALIZAMOS» LA DISTRIBUCIÓN BINOMIAL CON $p = 1/2$ (O SEA, HACIENDO QUE SU CENTRO SEA CERO Y SU DESVIACIÓN TÍPICA = 1) ENTONCES SE APROXIMA MUCHO A LA DISTRIBUCIÓN NORMAL TIPIFICADA.

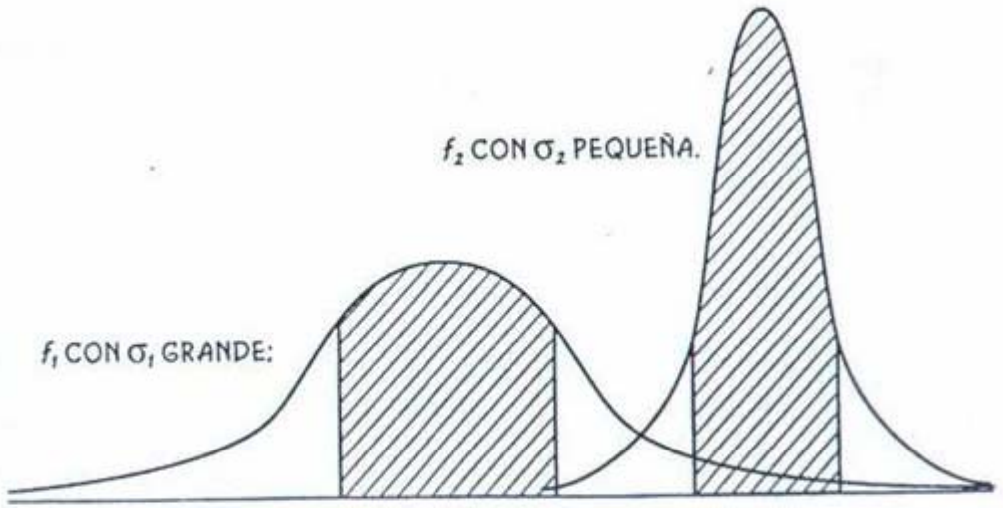
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

OTRAS NORMALES, CON DISTINTAS MEDIAS Y VARIANZAS, SE OBTIENEN EXTENDIENDO Y DESPLAZANDO LA NORMAL TIPIFICADA. EN GENERAL, PODEMOS ESCRIBIR LA FÓRMULA

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

ESTO NOS DA UNA DISTRIBUCIÓN SIMÉTRICA Y CAMPANIFORME CON EL CENTRO EN LA MEDIA μ Y LA DESVIACIÓN TÍPICA σ.

AQUÍ TIENES DOS NORMALES DIFERENTES CON LA ZONA DE LA DESVIACIÓN TÍPICA SOMBREADA.

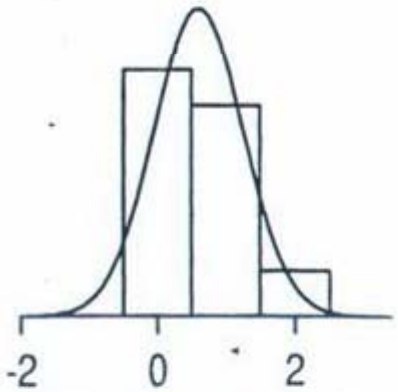


DE MOIVRE DEMOSTRÓ QUE LA NORMAL TIPIFICADA SE CORRESPONDE CON LA BINOMIAL (NORMALIZADA) DE $p = 0{,}5$. PERO LO CIERTO ES QUE FUNCIONA CON CUALQUIER VALOR DE p.

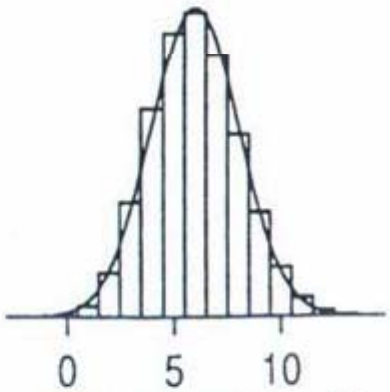
EN GENERAL: PARA CUALQUIER VALOR DE p, LA DISTRIBUCIÓN BINOMIAL DE n PRUEBAS CON PROBABILIDAD p SE APROXIMA A LA CURVA NORMAL CON $\mu = np$ Y $\sigma = np(1-p)$.

LA VERDAD ES QUE ESTO ES UN POCO EXTRAÑO. TODAS LAS NORMALES SON SIMÉTRICAS Y CAMPANIFORMES, PERO —COMO YA HEMOS VISTO— LAS DISTRIBUCIONES BINOMIALES NO SON SIMÉTRICAS CUANDO $p \neq 0{,}5$.

SIN EMBARGO, RESULTA QUE A MEDIDA QUE n CRECE, LA ASIMETRÍA DE LA BINOMIAL SE COMPENSA, COMO PUEDES VER EN ESTE EJEMPLO:



Binomial: $n = 2$ y $p = 0{,}3$



Binomial: $n = 20$ y $p = 0{,}3$

DE HECHO, EL DESCUBRIMIENTO DE DE MOIVRE SOBRE LA BINOMIAL ES UN CASO ESPECIAL DE UN RESULTADO AÚN MÁS GENERAL, QUE NOS AYUDA A EXPLICAR POR QUÉ LA NORMAL ES TAN IMPORTANTE Y DE NATURALEZA TAN EXTENDIDA. SE TRATA DEL SIGUIENTE:

«Teorema central del límite»:

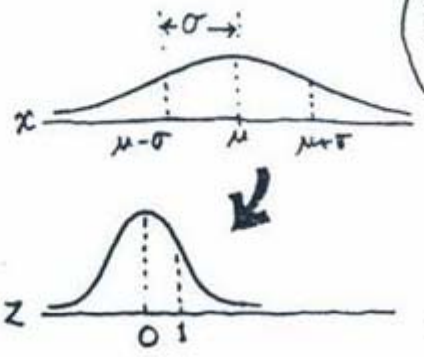
LOS DATOS INFLUIDOS POR MUCHOS PEQUEÑOS EFECTOS ALEATORIOS INDEPENDIENTES TIENEN, MÁS O MENOS, UNA DISTRIBUCIÓN NORMAL.

ASÍ SE EXPLICA QUE LA NORMAL **ESTÉ EN TODAS PARTES**: LAS FLUCTUACIONES DE LA BOLSA, LOS PESOS DE LOS ESTUDIANTES, LA MEDIA ANUAL DE TEMPERATURAS, LAS NOTAS DE SELECTIVIDAD: TODOS SON RESULTADOS DE MÚLTIPLES EFECTOS DIFERENTES. POR EJEMPLO, EL PESO DE UN ESTUDIANTE ES EL RESULTADO DE LA GENÉTICA, LA NUTRICIÓN, LAS ENFERMEDADES Y LA CERVEZA DE LA FIESTA DE LA NOCHE ANTERIOR. CUANDO LOS JUNTAMOS TODOS, ¡OBTENEMOS LA NORMAL! (RECUERDA QUE LA BINOMIAL ES EL RESULTADO DE n PRUEBAS DE BERNOULLI INDEPENDIENTES.)

LA TRANSFORMACIÓN z

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

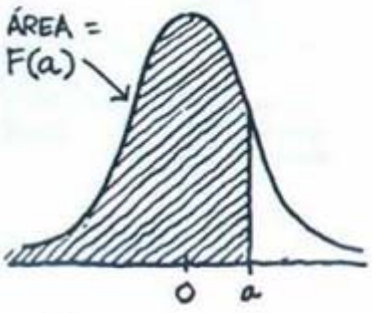
CONVIERTE UNA VARIABLE ALEATORIA NORMAL DE MEDIA μ Y DESVIACIÓN TÍPICA σ EN UNA VARIABLE ALEATORIA NORMAL TIPIFICADA CON MEDIA 0 Y DESVIACIÓN TÍPICA 1.



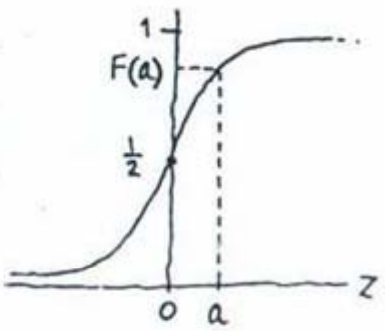
ENTONCES, TODO LO QUE NECESITAMOS PARA ENCONTRAR CUALQUIER DISTRIBUCIÓN NORMAL ES UNA SOLA TABLA DE LA NORMAL TIFIFICADA $f(z)$.

z	-2,5	-2,4	-2,3	-2,2	-2,1	-2,0	-1,9	-1,8	-1,7	-1,6
F(z)	0,006	0,008	0,011	0,014	0,018	0,023	0,029	0,036	0,045	0,055
z	-1,5	-1,4	-1,3	-1,2	-1,1	-1,0	-0,9	-0,8	-0,7	-0,6
F(z)	0,067	0,081	0,097	0,115	0,136	0,159	0,184	0,212	0,242	0,274
z	-0,5	-0,4	-0,3	-0,2	- 0,1	0,0	0,1	0,2	0,3	0,4
F(z)	0,309	0,345	0,382	0,421	0,460	0,500	0,540	0,579	0,618	0,655
z	0,5	0,6	0,7	0,8	0,9	1,0	1,1	1,2	1,3	1,4
F(z)	0,691	0,726	0,758	0,788	0,816	0,841	0,864	0,885	0,903	0,919
z	1,5	1,6	1,7	1,8	1,9	2,0	2,1	2,2	2,3	2,4
F(z)	0,933	0,945	0,955	0,964	0,971	0,977	0,982	0,986	0,989	0,992
z	2,5									
F(z)	0,994									

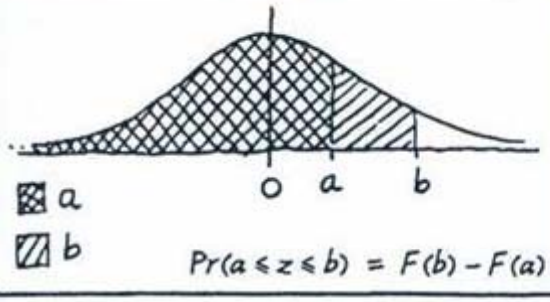
AQUÍ $F(a) = Pr(z \leq a)$, EL ÁREA DE LA CURVA DE DENSIDAD A LA IZQUIERDA DE $z = a$.



(TAMBIÉN PODEMOS CONFECCIONAR UN GRÁFICO DE $y = F(z)$. LA PROBABILIDAD ACUMULADA. TIENE ESTE ASPECTO.)



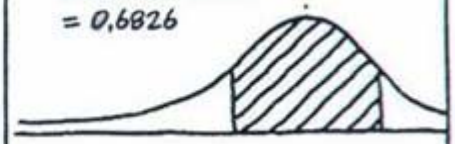
LA TABLA NOS PERMITE ENCONTRAR LA PROBABILIDAD DE QUE Z ESTÉ EN UN INTERVALO $a \leq z \leq b$. TAN SÓLO ES LA DIFERENCIA ENTRE LAS ÁREAS $F(b)$ Y $F(a)$.



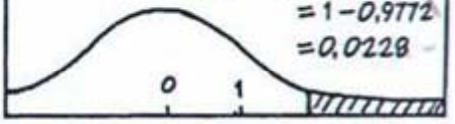
$$Pr(a \leq z \leq b) = F(b) - F(a)$$

DE ESTE MODO, POR EJEMPLO.

$$Pr(-1 < z < 1) = F(1) - F(-1)$$
$$= 0{,}8413 - 0{,}1587$$
$$= 0{,}6826$$

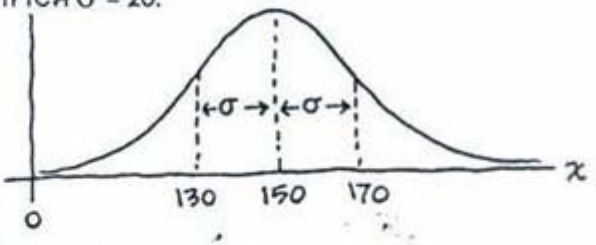


$$Pr(z \geq 2) = 1 - F(2)$$
$$= 1 - 0{,}9772$$
$$= 0{,}0228$$



SI UTILIZAMOS LA SUSTITUCIÓN $z = \frac{x-\mu}{\sigma}$, TAMBIÉN PODEMOS USAR LA MISMA TABLA PARA ENCONTRAR LAS PROBABILIDADES DE OTRAS DISTRIBUCIONES NORMALES.

POR EJEMPLO, SUPONGAMOS QUE LOS PESOS DE LOS ESTUDIANTES TIENEN UNA DISTRIBUCIÓN NORMAL CON MEDIA $\mu = 150$ LIBRAS Y UNA DESVIACIÓN TÍPICA $\sigma = 20$:

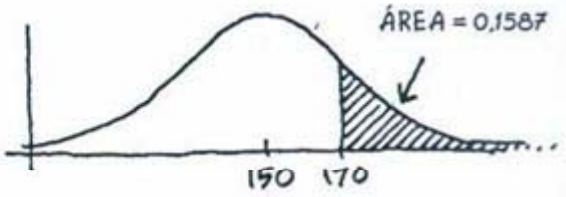


ENTONCES, ¿CUÁL ES LA PROBABILIDAD DE PESAR MÁS DE 170 LIBRAS?

AHORA SE TRATA «SÓLO» DE ÁLGEBRA.

$$Pr(X > 170) =$$
$$Pr\left(\frac{X-\mu}{\sigma} > \frac{170-150}{20}\right) =$$
$$Pr\left(Z > \frac{20}{20}\right) =$$
$$Pr(Z > 1)$$

ESO ES $1 - F(1)$, QUE COMO PODEMOS VER EN LA TABLA ES $1 - 0{,}8413 = 0{,}1587$.

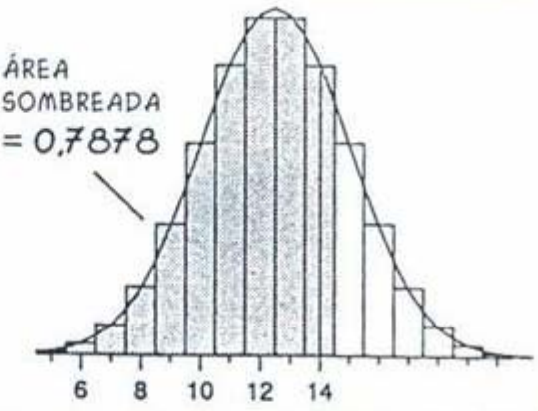


ALGO MENOS DE UN ESTUDIANTE DE CADA SEIS PESA MÁS DE 170 LIBRAS.

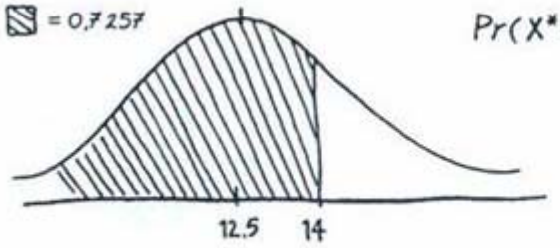
ENTONCES, LA REGLA GENERAL PARA CALCULAR LAS PROBABILIDADES ASOCIADAS A LA DISTRIBUCIÓN NORMAL ES:

$$Pr(a \leq X \leq b) = F\left(\frac{b-\mu}{\sigma}\right) - F\left(\frac{a-\mu}{\sigma}\right)$$

Y AHORA, VOLVIENDO A DE MOIVRE Y SU APROXIMACIÓN BINOMIAL... VAMOS A VER UNA DISTRIBUCIÓN BINOMIAL CON $n = 25$ PRUEBAS Y $p = 0{,}5$ (25 LANZAMIENTOS DE UNA MONEDA, POR EJEMPLO). PODEMOS CALCULAR (O CONSULTAR EN LA TABLA) CUALQUIER PROBABILIDAD, POR EJEMPLO $Pr(x \leq 14)$. Y ES EXACTAMENTE 0,7878.



AHORA CALCULAMOS UNA VARIABLE ALEATORIA NORMAL X^* CON LA MISMA MEDIA $\mu = np = (25)(0{,}5) = 12{,}5$ Y DESVIACIÓN TÍPICA $\sigma = np(1-p) = 2{,}5$.



$$Pr(X^* \leq 14) = Pr\left(Z \leq \frac{14 - 12{,}5}{2{,}5}\right)$$
$$= Pr(Z \leq 0{,}6)$$
$$= 0{,}7257$$

¡AH! ¡PERO AÚN PODEMOS MEJORARLO! SI OBSERVAS EL HISTOGRAMA CON ATENCIÓN, VERÁS QUE LAS BARRAS TIENEN UN NÚMERO EN EL CENTRO. ESTO SIGNIFICA QUE $Pr(X^* \leq 14)$ ES EN REALIDAD EL ÁREA DE LAS BARRAS MENORES A $x = 14{,}5$. DEBEMOS TENER EN CUENTA ESE 0,5, Y DE HECHO,

$$Pr(X^* \leq 14{,}5) = Pr(Z \leq 0{,}8)$$
$$= 0{,}7881$$

¡UNA APROXIMACIÓN MAGNÍFICA A 0,7878!

ESE OTRO 0,5 QUE HEMOS AÑADIDO SE LLAMA

corrección de continuidad.

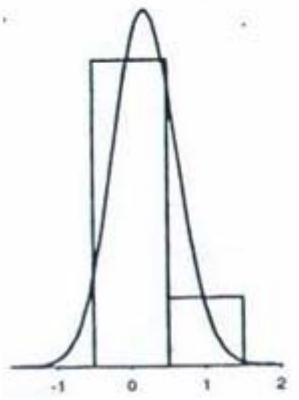
TENEMOS QUE INCLUIRLO PARA OBTENER UNA BUENA APROXIMACIÓN CONTINUA A NUESTRA VARIABLE ALEATORIA BINOMIAL DISCRETA X. TODO SE RESUME EN ESTA HORRIBLE FÓRMULA:

$$Pr(a \leq X \leq b) \simeq Pr\left(\frac{a-\frac{1}{2}-np}{\sqrt{np(1-p)}} \leq Z \leq \frac{b+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$$

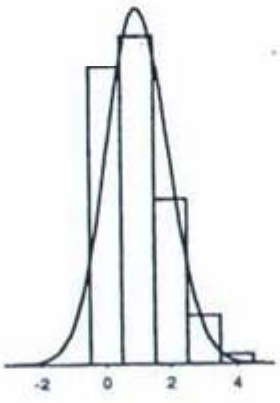
¿CUÁNDO ES LA APROXIMACIÓN «LO SUFICIENTEMENTE BUENA»? PARA LOS ESTADÍSTICOS, LA REGLA EMPÍRICA ES LA SIGUIENTE: SIEMPRE QUE n SEA LO BASTANTE GRANDE PARA QUE TANTO EL NÚMERO DE ÉXITOS COMO EL DE FRACASOS SEA MAYOR QUE CINCO:

$$np \geq 5 \quad \text{Y} \quad n(1-p) \geq 5$$

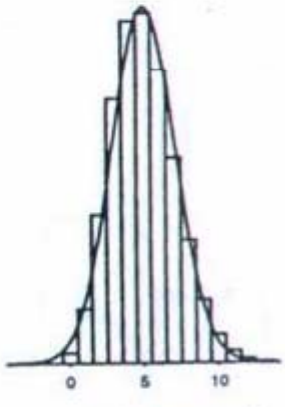
EN ESTOS HISTOGRAMAS PUEDES VER QUE CUANDO $p = 0{,}1$ LA EQUIVALENCIA ES BASTANTE MEDIOCRE, O INCLUSO MUY MALA, HASTA QUE n LLEGA A 50 Y HACE QUE $np = 5$.



$n=2,\ p=0{,}1$

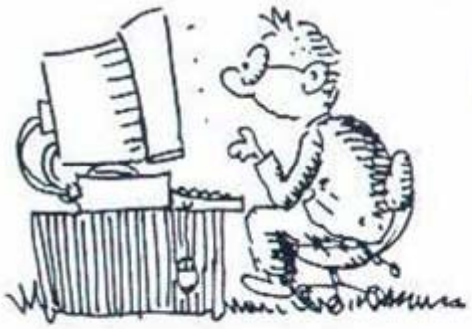


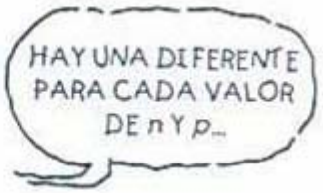
$n=10,\ p=0{,}1$



$n=50,\ p=0{,}1$

¿QUÉ TIENE DE MARAVILLOSO ESTA APROXIMACIÓN NORMAL? LA DISTRIBUCIÓN BINOMIAL SE DA MUY A MENUDO EN LA NATURALEZA, Y NO ES DIFÍCIL DE COMPRENDER, PERO CALCULARLA PUEDE SER AGOTADOR.





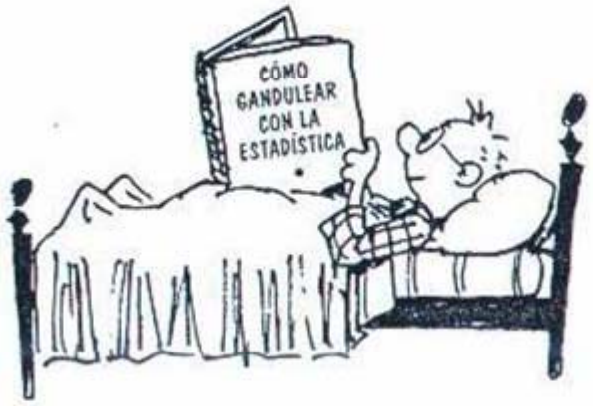
LA NORMAL QUE SE LE APROXIMA ES QUIZÁ MENOS INTUITIVA, PERO MUY FÁCIL DE USAR. LA TRANSFORMACIÓN z CONVIERTE CUALQUIER NORMAL A LA NORMAL TIPIFICADA, Y ESO NOS PERMITE LEER LAS PROBABILIDADES DIRECTAMENTE DE UNA SIMPLE TABLA NUMÉRICA.

Y ADEMÁS, ¡LA NORMAL ES LA MADRE DE TODAS LAS DISTRIBUCIONES!

♦ Capítulo 6 ♦
MUESTREO

A ESTAS ALTURAS, TRAS UNA DIETA REGULAR DE MONEDAS, DATOS E IDEAS ABSTRACTAS, A LO MEJOR TE PREGUNTAS QUÉ TIENE QUE VER TODO ESTE MATERIAL ESTADÍSTICO QUE HEMOS DESARROLLADO CON EL **MUNDO REAL**. BUENO, POR FIN LO VAS A DESCUBRIR...

EN ESTE CAPÍTULO EMPEZAMOS A VER LA TAREA **REAL** DE LA ESTADÍSTICA, QUE AL FIN Y AL CABO ES AHORRARNOS **TIEMPO** Y DINERO. LA GENTE ODIA PERDER EL TIEMPO EN **TRABAJOS INNECESARIOS**, Y SI HAY ALGO QUE LA ESTADÍSTICA PUEDE HACER ES DECIRNOS EXACTAMENTE CUÁNTA HOLGAZANERÍA NOS PODEMOS PERMITIR:



EL PROBLEMA QUE TIENE EL MUNDO REAL ES QUE LOS CONJUNTOS DE COSAS SON TAN GRANDES QUE RESULTA MUY DIFÍCIL CONSEGUIR LA INFORMACIÓN QUE QUEREMOS:

EL PROCEDIMIENTO COMPLETO, LABORIOSO, CONCIENZUDO, COMO LO HARÍA UN CASTOR, DE CONTESTAR A TODAS ESTAS PREGUNTAS SERÍA MEDIR TODOS Y CADA UNO DE LOS PEPINILLOS DEL MUNDO (POR EJEMPLO) Y HACER LOS CÁLCULOS.

PERO NOSOTROS NO SOMOS CASTORES. ¡SOMOS ESTADÍSTICOS!
BUSCAMOS LA FORMA MÁS SENCILLA...

NUESTRO MÉTODO ES TOMAR UNA MUESTRA... UN SUBCONJUNTO RELATIVAMENTE PEQUEÑO DE LA POBLACIÓN TOTAL. IGUAL QUE CUANDO SE HACE UN SONDEO DE OPINIÓN DURANTE UNAS ELECCIONES.

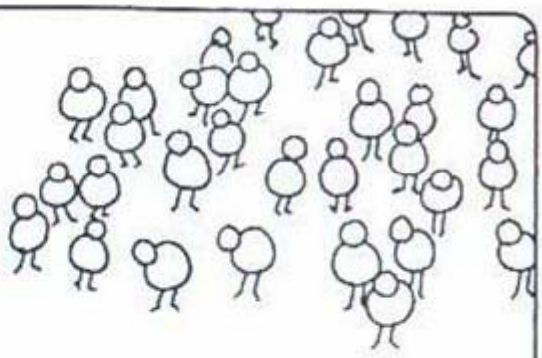
UNA PREGUNTA OBVIA ES: ¿CUÁNTOS ELEMENTOS DEBE TENER LA MUESTRA PARA OBTENER RESULTADOS SIGNIFICATIVOS?

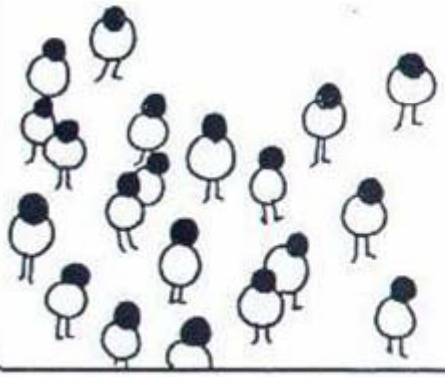
Y LA RESPUESTA, QUE DEBE QUEDARTE GRABADA EN EL CEREBRO PARA SIEMPRE JAMÁS, ES: SI n ES EL NÚMERO DE ELEMENTOS DE LA MUESTRA, ENTONCES TODO ESTÁ GOBERNADO POR

$$\frac{1}{\sqrt{n}}.$$

DISEÑO DEL MUESTREO

ANTES DE EMPEZAR CON LOS NÚMEROS, DEBERÍAMOS SEÑALAR QUE LA **CALIDAD** DE LA MUESTRA ES TAN IMPORTANTE COMO SU **TAMAÑO**. ¿CÓMO PODEMOS ESTAR SEGUROS DE QUE ESCOGEMOS UNA MUESTRA REPRESENTATIVA?



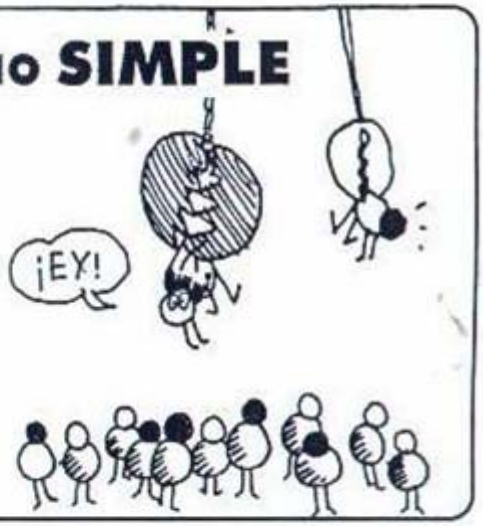


EL MISMO PROCESO DE SELECCIÓN ES DE VITAL IMPORTANCIA. POR EJEMPLO, UNA ENCUESTA DE VOTANTES QUE EXCLUYA SISTEMÁTICAMENTE A LOS NEGROS NO TENDRÍA NINGÚN VALOR, Y HAY MILES DE FORMAS MÁS DE ESTROPEAR, O SESGAR, UNA MUESTRA.

PARA NO PROLONGAR EL MISTERIO, LA FORMA DE OBTENER RESULTADOS ESTADÍSTICOS FIABLES ES ESCOGER LA MUESTRA **al azar.**

EL **MUESTREO ALEATORIO SIMPLE**

SUPONGAMOS QUE TENEMOS UNA GRAN POBLACIÓN DE OBJETOS Y UN PROCEDIMIENTO PARA ESCOGER *n* DE ELLOS. SI ESE PROCEDIMIENTO ASEGURA QUE **TODAS LAS MUESTRAS POSIBLES DE *n* OBJETOS TIENEN LA MISMA PROBABILIDAD**, ENTONCES ESE PROCEDIMIENTO RECIBE EL NOMBRE DE **muestreo aleatorio simple.**



EL MUESTREO ALEATORIO SIMPLE PRESENTA DOS PROPIEDADES QUE LO CONVIERTEN EN UN ESTÁNDAR FRENTE AL QUE MEDIMOS TODOS LOS OTROS MÉTODOS:

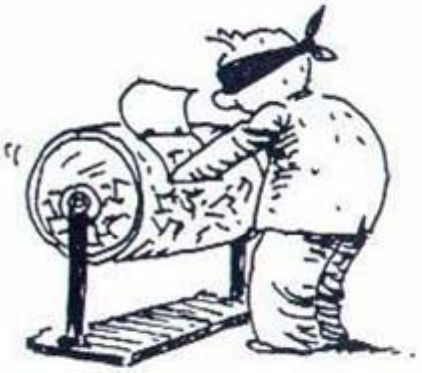
1) REPRESENTATIVA: CADA UNIDAD TIENE LAS MISMAS POSIBILIDADES DE SER ESCOGIDA.*

2) INDEPENDENCIA: LA SELECCIÓN DE UNA UNIDAD NO INFLUYE EN LA SELECCIÓN DE OTRAS UNIDADES.

* UN CONCEPTO ESTADÍSTICO MÁS FORMAL ES LA AUSENCIA DE SESGO. [N.T.]

POR DESGRACIA, EN EL MUNDO REAL ES MUY DIFÍCIL ENCONTRAR MUESTRAS COMPLETAMENTE INDEPENDIENTES Y REPRESENTATIVAS. POR EJEMPLO, HACER UNA ENCUESTA A LOS VOTANTES MARCANDO NÚMEROS DE TELÉFONO AL AZAR ES UN MÉTODO NO REPRESENTATIVO: NO TIENE EN CUENTA A LOS VOTANTES QUE NO DISPONEN DE TELÉFONO Y CUENTA VARIAS VECES A LOS QUE TIENEN VARIOS NÚMEROS.

TEÓRICAMENTE, ES POSIBLE OBTENER UNA MUESTRA AL AZAR CONSTRUYENDO UN **MARCO DE MUESTREO**: UNA LISTA CON TODAS LAS UNIDADES DE LA POBLACIÓN. UTILIZANDO UN GENERADOR DE NÚMERO ALEATORIO, ESCOGEMOS *n* OBJETOS AL AZAR.



DE IGUAL FORMA, PODEMOS ESCRIBIR TODOS LOS NOMBRES EN TARJETAS Y EXTRAER *n* DE ELLOS DE UN BOMBO.

SIN EMBARGO, NO SIEMPRE ES TAN SENCILLO. EL MARCO PUEDE RESULTAR PROHIBITIVO, CARO, POLÉMICO E, INCLUSO, IMPOSIBLE DE ESTABLECER. POR EJEMPLO, UN ESTUDIO SOBRE LA CALIDAD DEL AGUA DE LA AGENCIA PARA LA PROTECCIÓN DEL MEDIO AMBIENTE DE ESTADOS UNIDOS NECESITABA UN MARCO DE MUESTRA DE LOS LAGOS DEL PAÍS. ASÍ QUE ALGUIEN TENÍA QUE DECIDIR:

¿EXISTE ALGÚN OTRO MÉTODO MÁS EFICAZ Y RENTABLE QUE UNA MUESTRA ALEATORIA SIMPLE? SÍ, SI ES QUE YA SABES ALGO DE LA POBLACIÓN. POR EJEMPLO...

EL MUESTREO

estratificado

DIVIDE LAS UNIDADES DE POBLACIÓN EN GRUPOS HOMOGÉNEOS (ESTRATOS) Y LUEGO LLEVA A CABO MUESTREO ALEATORIO SIMPLE DE CADA GRUPO.

POR EJEMPLO, LA POBLACIÓN DE TODAS LAS CONSERVAS EN VINAGRE SE PUEDE ESTRATIFICAR POR EL TIPO DE CONSERVA. DENTRO DE CADA TIPO, O ESTRATO, EL TAMAÑO SERÁ MENOS VARIABLE.

EL MUESTREO POR

conglomerados

AGRUPA LA POBLACIÓN EN PEQUEÑOS CONGLOMERADOS, REALIZA UNA MUESTRA ALEATORIA SIMPLE DE ELLOS Y TIENE EN CUENTA ABSOLUTAMENTE TODO DENTRO DE CADA CONGLOMERADO MUESTREADO. ESTO PUEDE RESULTAR RENTABLE SI LOS COSTES DE TRANSPORTE ENTRE LAS UNIDADES DE MUESTRA ALEATORIA SON ELEVADOS.

UN BUEN EJEMPLO ES EL DE UNA ENCUESTA SOBRE LA VIVIENDA, QUE DIVIDE LA CIUDAD EN BLOQUES Y ESTUDIA CADA UNIDAD DE VIVIENDAS DE CADA BLOQUE DE LA MUESTRA.

EL MUESTREO **sistemático** EMPIEZA CON UNA UNIDAD ESCOGIDA AL AZAR Y LUEGO SELECCIONA CADA UNIDAD QUE SE ENCUENTRE A k UNIDADES DE AQUELLA. POR EJEMPLO, UN ESTUDIO DEL TRÁFICO EN AUTOPISTAS PODRÍA ESTUDIAR UNO DE CADA CIEN COCHES QUE PASARA POR EL PEAJE. ESTE PLAN ES FÁCIL DE APLICAR Y PUEDE SER MÁS EFICAZ SI LOS PATRONES DEL TRÁFICO VARÍAN CON EL PASO DE LAS HORAS.

Nota de advertencia número 1:

LA MAYORÍA DE LOS MÉTODOS ESTADÍSTICOS SE BASAN EN LA INDEPENDENCIA Y LA REPRESENTATIVIDAD DEL MUESTREO ALEATORIO SIMPLE. LOS RESULTADOS POSTERIORES RESPONDEN ÚNICAMENTE AL MUESTREO ALEATORIO SIMPLE. EN OTROS PROCEDIMIENTOS DE MUESTREO, LOS RESULTADOS DEBEN MODIFICARSE. LOS DETALLES APARECEN EN LIBROS DE TEXTO ESPECIALIZADOS EN MUESTREO Y EN ALGORITMOS COMPUTACIONALES.

Nota de advertencia número 2:

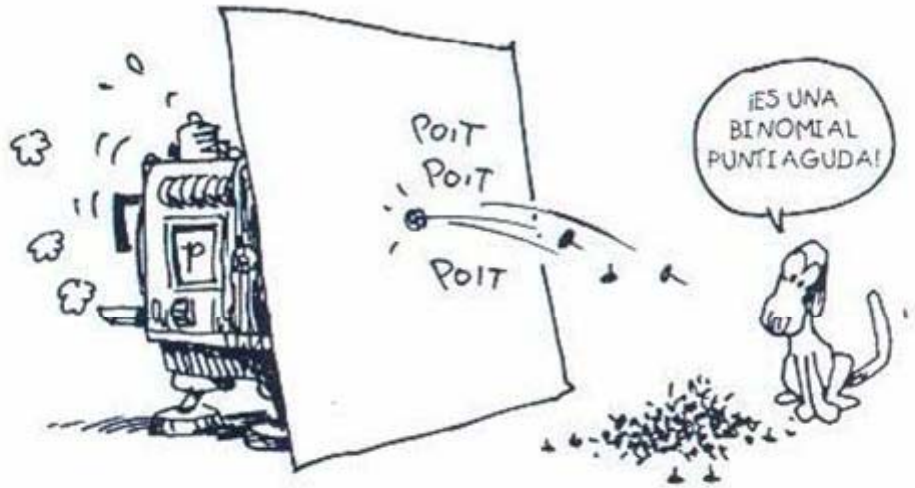
NO EXISTE ANÁLISIS ESTADÍSTICO FIABLE SIN UN DISEÑO ALEATORIZADO, NO IMPORTA CUÁNTO SE MODIFIQUE DESPUÉS. LA BELLEZA DEL MUESTREO ALEATORIO RESIDE EN QUE «GARANTIZA ESTADÍSTICAMENTE» LA EXACTITUD DEL ESTUDIO.

UNO DE LOS MÉTODOS MÁS COMUNES TIENDE A MENUDO A LA PARCIALIDAD: SE TRATA DEL MUESTREO **oportunista.** ESTE MÉTODO EVITA TODA LA PROBLEMÁTICA DE DISEÑAR UN PROCEDIMIENTO Y SE LIMITA A TOMAR LAS PRIMERAS n UNIDADES DE LA POBLACIÓN QUE SE PRESENTEN.

UN CLÁSICO EJEMPLO ES EL LIBRO DE SHERE HITE *MUJERES Y AMOR*. SE ENVIARON 100.000 CUESTIONARIOS A ORGANIZACIONES DE MUJERES (UN MUESTREO OPORTUNISTA), Y SÓLO UN 4,5% SE RELLENARON Y ENTREGARON (RESPUESTA PARCIAL). ASÍ QUE SUS «RESULTADOS» ESTABAN BASADOS EN UNA MUESTRA DE MUJERES QUE, POR UNA RAZÓN U OTRA, TENÍAN UNA GRAN MOTIVACIÓN PARA CONTESTAR LAS PREGUNTAS DE LA ENCUESTA.

El TAMAÑO MUESTRAL y el error típico

EL ASTUTO LECTOR SE DARÁ CUENTA EN SEGUIDA DE QUE SE TRATA DE UN **SISTEMA DE BERNOULLI**: CADA NUEVO CLAVO ES EL RESULTADO DE UNA PRUEBA DE BERNOULLI CON PROBABILIDAD p DE ÉXITO (EN ESTE CASO, NO SER DEFECTUOSO) Y PROBABILIDAD $1-p$ DE FRACASO (SER DEFECTUOSO).



PENSAMOS EN ESTA SITUACIÓN COMO SI HUBIESE UNA «MÁQUINA DE BERNOULLI» **REAL AUNQUE ESCONDIDA** CUYA PROBABILIDAD p RIGE LOS RESULTADOS QUE OBSERVAMOS EN EL LLAMADO «MUNDO REAL».

COMO LA MÁQUINA DE BERNOULLI ES INVISIBLE, NO SABEMOS CUÁL ES LA PROBABILIDAD p. PERO NOS GUSTARÍA DESCUBRIRLO. ASÍ QUE TOMAMOS UNA **MUESTRA ALEATORIA** DE N CLAVOS Y VEMOS QUE, DE TODOS ELLOS, x NO TIENEN NINGÚN DEFECTO.

BIEN, LA PROPORCIÓN DE ÉXITOS NO DEBERÍA DIFERIR MUCHO DE p... ASÍ QUE LA LLAMAMOS $\hat{p}$. «PE CON CIRCUNFLEJO».

$$\hat{p} = \frac{x}{n}$$

$\hat{p}$ ES EL NÚMERO DE ÉXITOS x DE LA MUESTRA, DIVIDIDO ENTRE EL TAMAÑO n DE ÉSTA. POR EJEMPLO, SI p FUERA 0.85, Y HUBIÉSEMOS TOMADO $n = 1.000$ TORNILLOS COMO MUESTRA, QUIZÁ ALREDEDOR DE $x = 832$ ESTARÍAN BIEN Y ENTONCES $\hat{p} = 0.832$.

NOS PREGUNTAMOS: ¿ES BUENA ESTA ESTIMACIÓN?

Y CONTESTAMOS CON OTRO INTERROGANTE: ¿QUÉ SIGNIFICA LA PRIMERA PREGUNTA?

NO PODEMOS SABER LA DIFERENCIA EXACTA ENTRE $\hat{p}$ Y p, PORQUE NO CONOCEMOS EL VALOR p. LA AUTÉNTICA PREGUNTA ES LA SIGUIENTE: SI TOMÁRAMOS MUCHAS MUESTRAS DE 1.000 CLAVOS Y OBSERVÁRAMOS EL NÚMERO $\hat{p}$ DE CADA MUESTRA, ¿CUÁL SERÍA LA DISTRIBUCIÓN DE ESOS VALORES DE $\hat{p}$ ALREDEDOR DE p?

DE HECHO, ESTOS VALORES DE $\hat{p}$ CADA VEZ SE PARECEN MÁS A UNA VARIABLE ALEATORIA: LA SELECCIÓN DE UNA MUESTRA DE n UNIDADES ES UN EXPERIMENTO ALEATORIO, ¡Y LA OBSERVACIÓN $\hat{p}$ ES UN RESULTADO NUMÉRICO!

PARA SER EXACTOS, SI X ES EL NÚMERO DE ÉXITOS DE LA MUESTRA, ENTONCES X NO ES MÁS QUE NUESTRA VIEJA AMIGA LA VARIABLE ALEATORIA BINOMIAL (n PRUEBAS, PROBABILIDAD p)... Y DEFINIMOS LA PROPORCIÓN OBSERVADA COMO LA VARIABLE ALEATORIA

$$\hat{P} = \frac{X}{n}$$

COMO LO SABEMOS TODO SOBRE X, PODEMOS DEDUCIR SIN PROBLEMAS UNOS CUANTOS HECHOS SOBRE $\hat{P}$:

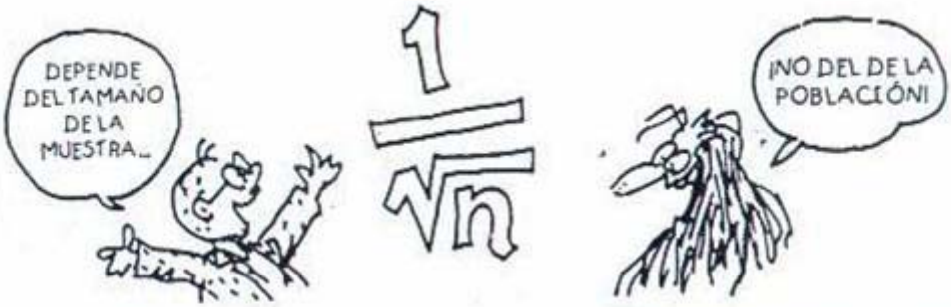
1) LA MEDIA DE $\hat{P}$ ES $E[\hat{P}] = p$

2) LA DESVIACIÓN TÍPICA DE $\hat{P}$ ES

$$\sigma(\hat{P}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

3) PARA UNA n MUY GRANDE, $\hat{P}$ ES APROXIMADAMENTE NORMAL.

¡Y ESO ES TODO! LOS VALORES OBSERVADOS DE $\hat{P}$ SE CENTRARÁN EN p (EVIDENTEMENTE), Y SU DESVIACIÓN TÍPICA, O DISPERSIÓN, SERÁ PROPORCIONAL AL NÚMERO MÁGICO QUE HABÍAMOS MENCIONADO AL PRINCIPIO DEL CAPÍTULO:



Y COMO $\hat{P}$ ES BASTANTE NORMAL, PODEMOS USAR LA REGLA EMPÍRICA PARA CONCLUIR QUE APROXIMADAMENTE UN 68% DE TODAS LAS ESTIMACIONES QUEDARÁN A MENOS DE UNA DESVIACIÓN TÍPICA DEL VALOR REAL p.

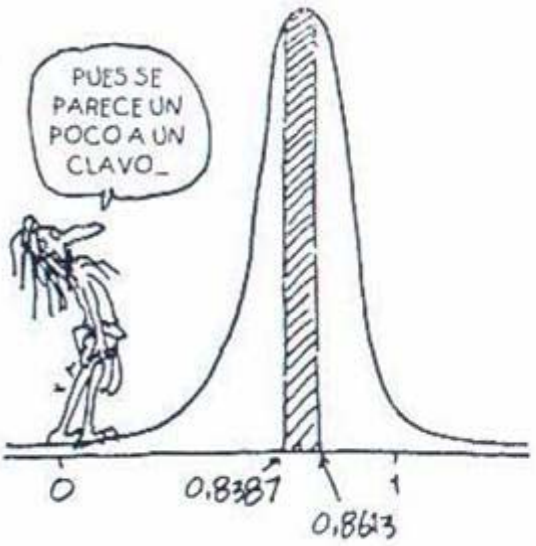
Y VOLVIENDO A LOS CLAVOS, CON $n = 1.000$ Y $p = 0,85$, LA DESVIACIÓN TÍPICA DE $\hat{p}$ ES

$$\sigma(\hat{P}) = \sqrt{\frac{(0,85)(0,15)}{1.000}}$$

$$= 0,0113$$

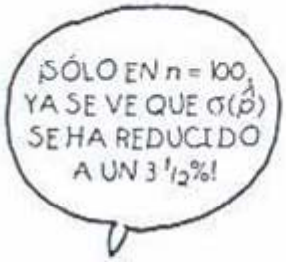
ASÍ QUE ESPERAMOS QUE ALREDEDOR DE UN 68% DE NUESTRAS ESTIMACIONES QUEDEN DENTRO DEL PEQUEÑO INTERVALO

$$0,8387 \leq \hat{p} \leq 0,8613$$



LA DESVIACIÓN TÍPICA DE $\hat{p}$ ES UNA MEDIDA DEL **error muestral.**

COMO YA HEMOS VISTO, PARA LA BINOMIAL ESTE ERROR MUESTRAL ES INVERSAMENTE PROPORCIONAL A $\sqrt{n}$. SI SE AUMENTA EL TAMAÑO MUESTRAL EN UN FACTOR 4, LA DISPERSIÓN $\sigma(\hat{p})$ SE REDUCE EN UN FACTOR 2.



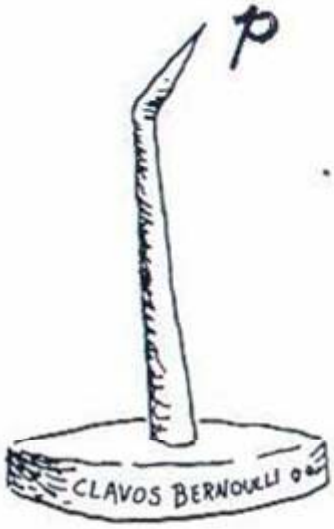
TAMAÑOS MUESTRALES DE CLAVOS, $p = 0,85$

n	1	4	16	25	100	10.000
$\sqrt{n}$	1	2	4	5	10	100
$\sigma(\hat{p})$	0,357	0,1785	0,089	0,071	0,0357	0,0036

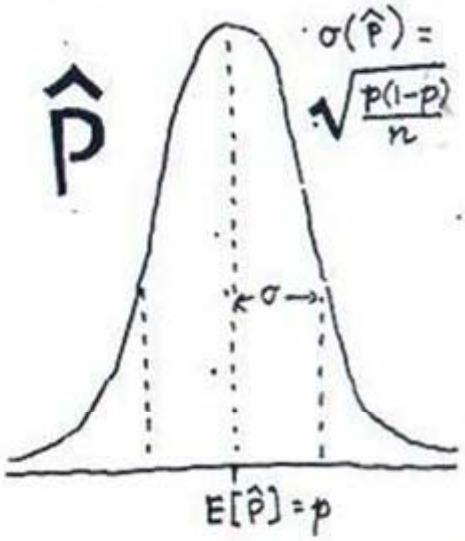
NOTA LINGÜÍSTICA: UNA **ESTIMACIÓN** ES UNA SOLA MEDIDA U OBSERVACIÓN. UN **ESTIMADOR** ES UNA REGLA PARA OBTENER ESTIMACIONES. EN ESTE CASO EL ESTIMADOR ES LA VARIABLE ALEATORIA $\hat{P} = \frac{X}{n}$.

CASI TODA LA ESTADÍSTICA IMPLICA UN PROCESO DE CUATRO ETAPAS POR EL QUE ACABAMOS DE PASAR:

DEFINIR LA POBLACIÓN CON UN PARÁMETRO DESCONOCIDO.



ENCONTRAR UN ESTIMADOR, SU DISTRIBUCIÓN MUESTRAL TEÓRICA Y SU DESVIACIÓN TÍPICA.



EXTRAER UNA MUESTRA ALEATORIA Y ENCONTRAR LA ESTIMACIÓN.

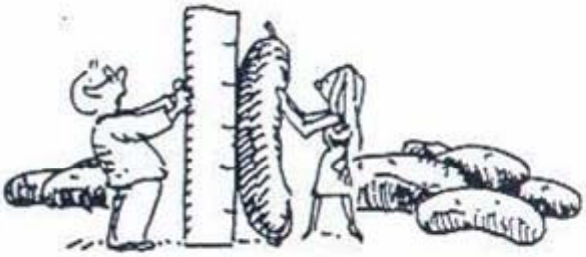
HACER UN INFORME CON LOS RESULTADOS Y SU ERROR MUESTRAL O ESTADÍSTICO.

Distribución muestral de la **MEDIA**

Y AHORA PASAMOS DE LOS CLAVOS A LOS PEPINILLOS EN VINAGRE...

A LOS FABRICANTES DE BOTES LES GUSTARÍA SABER EL TAMAÑO MEDIO DE UN PEPINILLO SIN TENER QUE EXAMINAR TODOS LOS PEPINOS DEL CONTINENTE. SELECCIONAN n PEPINILLOS AL AZAR Y LOS MIDEN, $x_1, x_2, \ldots, x_n$.

AHORA QUIZÁ YA TE HAYAS ACOSTUMBRADO A QUE CADA X_i ES UNA **VARIABLE ALEATORIA**: EL RESULTADO NUMÉRICO DE UN EXPERIMENTO ALEATORIO.



SI μ ES EL TAMAÑO MEDIO (DESCONOCIDO) DE UN PEPINILLO, Y σ ES LA DESVIACIÓN TÍPICA DE LA **DISTRIBUCIÓN DEL TAMAÑO DEL PEPINILLO**, ENTONCES

$$E[X_i] = \mu$$
$$\sigma(X_i) = \sigma$$

PARA CADA i (YA QUE X_i PODRÍA HABER SIDO EL TAMAÑO DE CUALQUIER PEPINILLO).

A CONTINUACIÓN OBSERVAMOS LA MEDIA MUESTRAL: EL TAMAÑO MEDIO DE LOS PEPINILLOS ESCOGIDOS. ES UNA NUEVA VARIABLE ALEATORIA QUE VIENE DADA POR:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

IGUAL QUE ANTES, NOS GUSTARÍA SABER LO «CERCA» QUE SE ENCUENTRA DE μ. ES DECIR, SI REALIZÁRAMOS ESTE MUESTREO REPETIDAS VECES, ¿CUÁL SERÍA LA DISTRIBUCIÓN DE $\bar{X}$? COMO TENEMOS DATOS DE $x_1, x_2, \ldots$ Y x_n, TAMBIÉN SABEMOS QUE

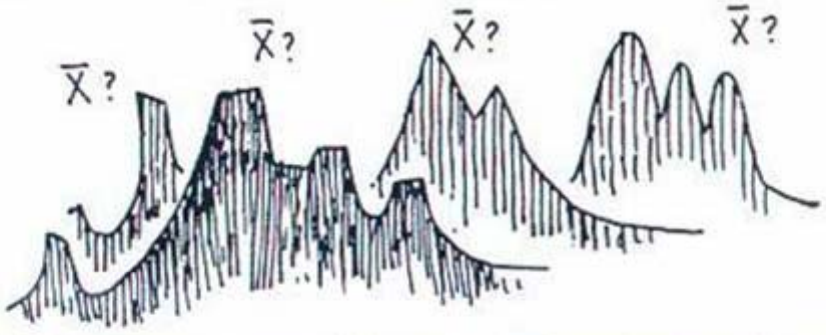
$$E[\bar{X}] = \mu$$

$$\sigma(\bar{X}) = \sigma/\sqrt{n}$$

¡DE NUEVO NOS ENCONTRAMOS CON ESE DENOMINADOR MÁGICO! LA DISPERSIÓN DE LAS MEDIAS MUESTRALES QUE HEMOS OBSERVADO ES PROPORCIONAL A

$$\frac{1}{\sqrt{n}}.$$

SIN EMBARGO, DESCONOCEMOS LA FORMA DE LA DISTRIBUCIÓN DE $\bar{X}$. LA DISTRIBUCIÓN DE PROBABILIDAD MUESTRAL DE $\hat{p}$ ERA CASI NORMAL PORQUE ESTABA BASADA EN UNA VARIABLE ALEATORIA BINOMIAL. PERO, ¿¿QUÉ PASA CON $\bar{X}$, EL ESTIMADOR DE LA MEDIA MUESTRAL??



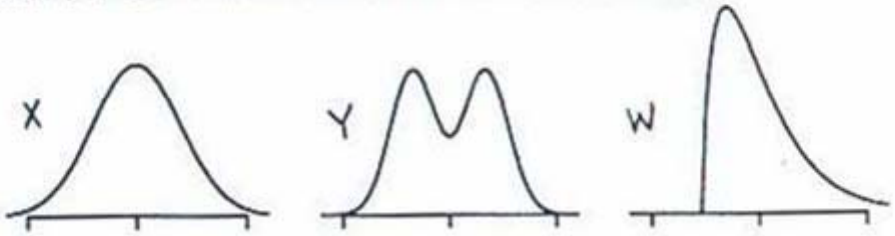
¡RESULTA QUE $\bar{X}$ TAMBIÉN ES APROXIMADAMENTE NORMAL! ESTE FAMOSO RESULTADO SE LLAMA TAMBIÉN

TEOREMA CENTRAL DEL LÍMITE

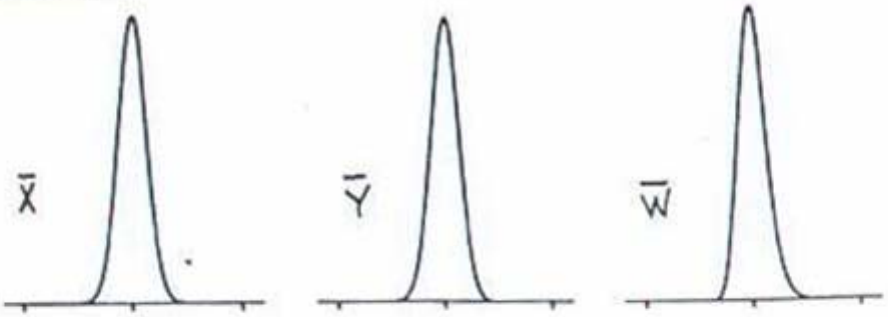
Y DICE ASÍ: SI TOMAMOS MUESTRAS ALEATORIAS DE TAMAÑO n DE UNA POBLACIÓN DE MEDIA μ Y DESVIACIÓN TÍPICA σ, ENTONCES, A MEDIDA QUE n SE HACE MAYOR, $\bar{X}$ SE ACERCA A LA **DISTRIBUCIÓN NORMAL** CON MEDIA μ Y DESVIACIÓN TÍPICA $\sigma/\sqrt{n}$. ENTONCES,

$$Pr(a \leq \bar{X} \leq b) \approx Pr\left(\frac{a-\mu}{\sigma/\sqrt{n}} \leq Z \leq \frac{b-\mu}{\sigma/\sqrt{n}}\right)$$

¿QUÉ TIENE ESTO DE EXTRAORDINARIO? NOS DICE QUE, SIN IMPORTAR LA FORMA QUE TENGA LA DISTRIBUCIÓN ORIGINAL (EN ESTE CASO, DEL TAMAÑO DE LOS PEPINILLOS), LA DISTRIBUCIÓN DE LA **MEDIA MUESTRAL** CONVERGE A UNA **NORMAL**. PARA ENCONTRAR LA DISTRIBUCIÓN DE $\bar{X}$, TAN SÓLO NECESITAMOS SABER LA MEDIA Y LA DESVIACIÓN TÍPICA POBLACIONALES.



ESTAS TRES DENSIDADES DE PROBABILIDAD DE AQUÍ ARRIBA TIENEN LA MISMA MEDIA Y DESVIACIÓN TÍPICA. A PESAR DE QUE TIENEN FORMAS DIFERENTES, CUANDO $n = 10$, LAS DISTRIBUCIONES MUESTRALES DE LA MEDIA, $\bar{X}$, SON CASI IDÉNTICAS.



La distribución t

POR MUY ASOMBROSO QUE SEA EL TEOREMA CENTRAL DEL LÍMITE, PRESENTA COMO MÍNIMO DOS PROBLEMAS:

UNO: DEPENDE DE UN TAMAÑO MUESTRAL MUY GRANDE.

DOS: PARA UTILIZARLO, NECESITAMOS CONOCER σ, LA DESVIACIÓN TÍPICA.

PERO, A MENUDO, LAS MUESTRAS SON PEQUEÑAS, Y NORMALMENTE SE DESCONOCE σ. SIN DUDA, EN EL CASO DE LOS PEPINILLOS NO TENEMOS LA MENOR IDEA DE CUÁNTO DISTA DE LA MEDIA EL TAMAÑO DE CADA UNO.

LO QUE PODEMOS HACER EN ESTE CASO ES ESTIMAR σ UTILIZANDO LA DESVIACIÓN TÍPICA DE LA MUESTRA, QUE, COMO RECORDARÁS, VIENE DADA POR LA FÓRMULA

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

ENTONCES, EN EL LUGAR DE LA VARIABLE ALEATORIA

$$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$$

SUSTITUIMOS σ POR s, Y DEFINIMOS UNA NUEVA VARIABLE ALEATORIA t

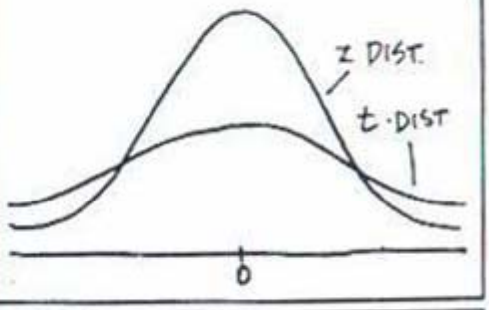
$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}}$$

PUEDES PENSAR EN LA VARIABLE ALEATORIA t COMO EN LO MEJOR QUE SE PUEDE HACER DADAS LAS CIRCUNSTANCIAS. SU DISTRIBUCIÓN RECIBE EL NOMBRE DE t DE STUDENT, PORQUE SU INVENTOR, WILLIAM GOSSET, LA PUBLICÓ CON EL SEUDÓNIMO DE «STUDENT».

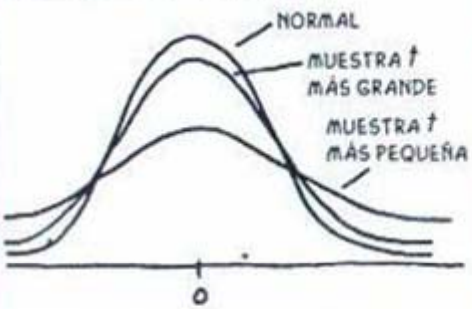
(GOSSET ERA EMPLEADO DE LA FÁBRICA DE CERVEZA GUINNESS, Y LE OBLIGARON, POR ALGÚN MOTIVO, A UTILIZAR SEUDÓNIMO.)

CON LA PRESUNCIÓN DE QUE LA DISTRIBUCIÓN POBLACIONAL ORIGINAL ERA NORMAL, O CASI NORMAL, «STUDENT» PUDO LLEGAR A UNA CONCLUSIÓN:

t TIENE MÁS DISPERSIÓN QUE z. ES MÁS «PLANA» QUE LA NORMAL. ESO ES PORQUE EL USO DE s INTRODUCE MAYOR INCERTIDUMBRE Y HACE QUE t SEA MENOS PRONUNCIADA QUE z.



LA CANTIDAD DE DISPERSIÓN DEPENDE DEL TAMAÑO MUESTRAL. CUANTO MAYOR SEA LA MUESTRA, MÁS SEGUROS PODEMOS ESTAR DE QUE s SE ACERCA A σ, Y t SE ACERCA MÁS A z, LA NORMAL.



GOSSET CONSIGUIÓ CALCULAR TABLAS DE t PARA VARIOS TAMAÑOS MUESTRALES. VEREMOS CÓMO USARLAS EN EL PRÓXIMO CAPÍTULO.

EN ESTE CAPÍTULO HEMOS TRATADO UN PROBLEMA CLAVE DE LA **ESTADÍSTICA DEL MUNDO REAL**: CÓMO SELECCIONAR UNA **MUESTRA** DE UNA POBLACIÓN GRANDE PARA QUE EL ANÁLISIS ESTADÍSTICO SEA VÁLIDO. ADEMÁS DEL «ESTÁNDAR DORADO» DE LA MUESTRA ALEATORIA SIMPLE, TAMBIÉN HEMOS DESCRITO OTROS ESQUEMAS MUESTRALES QUE SE UTILIZAN POR SU EFICACIA, PRECIO Y ASPECTO PRÁCTICO.

A CONTINUACIÓN, DANDO POR SUPUESTO EL MUESTREO ALEATORIO SIMPLE, HEMOS VISTO LA **DISTRIBUCIÓN** DE VARIOS ESTADÍSTICOS MUESTRALES. ES DECIR, HEMOS CONTEMPLADO LA MUESTRA COMO **EXPERIMENTO ALEATORIO** Y ASÍ SUS ESTADÍSTICOS SE HAN CONVERTIDO EN **VARIABLES ALEATORIAS**.

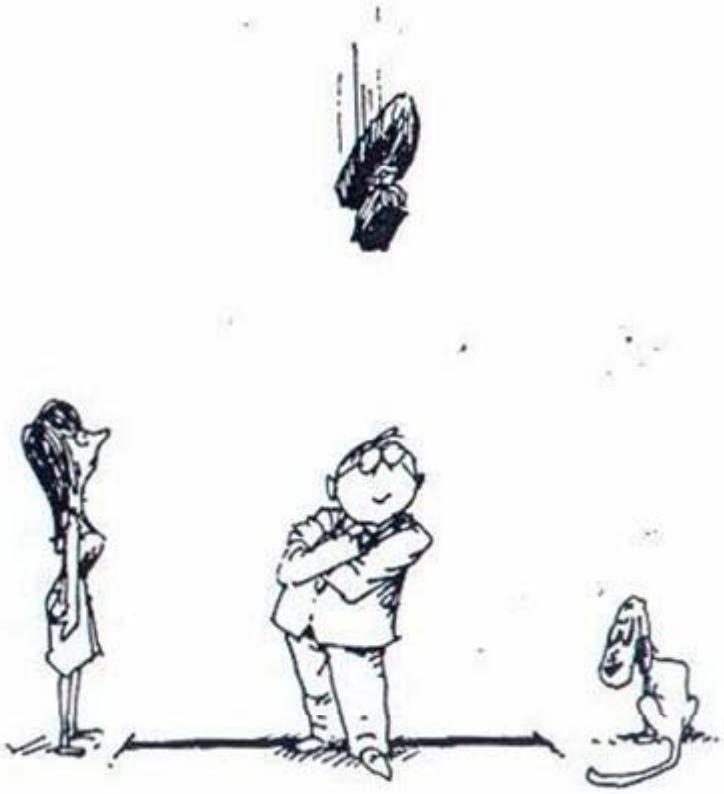
HEMOS DESCUBIERTO QUE LAS **PROPORCIONES** MUESTRALES $\hat{p}$ TENÍAN UNA DISTRIBUCIÓN MÁS O MENOS NORMAL, MIENTRAS QUE LAS DE LA **MEDIA** MUESTRAL $\bar{X}$ DEPENDÍAN DEL TAMAÑO DE LA MUESTRA. EN LAS MUESTRAS DE MAYOR TAMAÑO, LA DISTRIBUCIÓN ERA APROXIMADAMENTE NORMAL, MIENTRAS QUE EN LAS DE MENOR TAMAÑO, UTILIZAMOS LA DISTRIBUCIÓN *t* DE STUDENT.

* CON TÉ, ¡POR SUPUESTO!

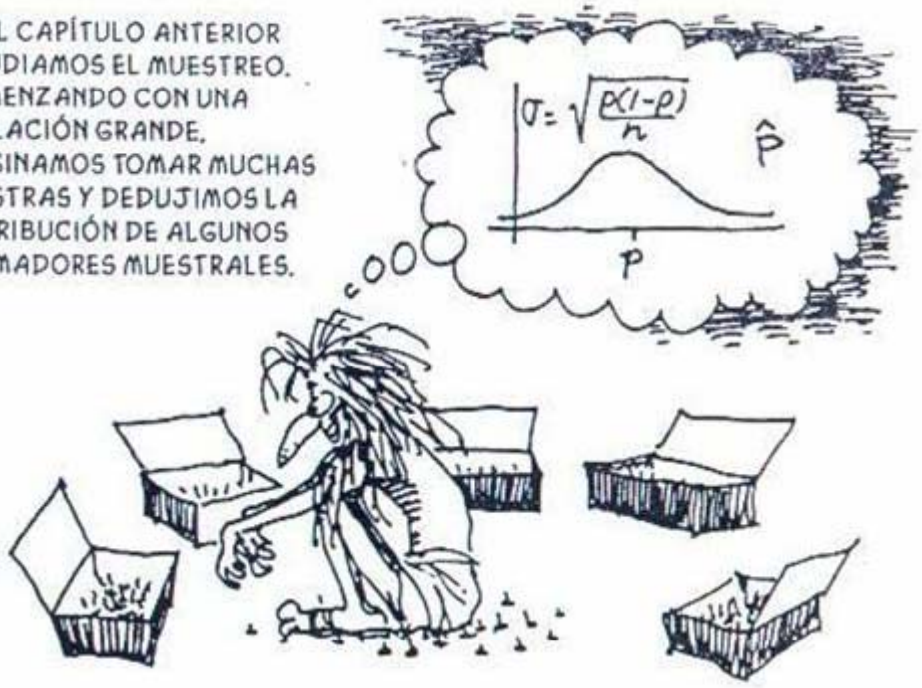
EN LOS DOS CAPÍTULOS SIGUIENTES, VEREMOS CÓMO UTILIZAR ESTAS DISTRIBUCIONES PARA HACER INFERENCIAS **ESTADÍSTICAS**: DADA UNA SOLA OBSERVACIÓN, COMO UN SONDEO DE OPINIÓN, ¿CÓMO USAMOS NUESTRO CONOCIMIENTO DE $\hat{p}$ Y $\bar{x}$ PARA EVALUARLO?

♦ Capítulo 7 ♦

INTERVALOS DE CONFIANZA



EN EL CAPÍTULO ANTERIOR ESTUDIAMOS EL MUESTREO. COMENZANDO CON UNA POBLACIÓN GRANDE, IMAGINAMOS TOMAR MUCHAS MUESTRAS Y DEDUJIMOS LA DISTRIBUCIÓN DE ALGUNOS ESTIMADORES MUESTRALES.



EN ESTE CAPÍTULO, HAREMOS LO CONTRARIO. CON UNA MUESTRA, NOS PLANTEAMOS LA SIGUIENTE PREGUNTA: ¿QUÉ SISTEMA ALEATORIO HA GENERADO SUS ESTADÍSTICOS?

ESTO REPRESENTA UN CAMBIO EN NUESTRA FORMA DE PENSAR: DEL RAZONAMIENTO DEDUCTIVO A LA INDUCCIÓN.

¡IGUAL QUE UNA INVESTIGACIÓN CRIMINAL, WATSON!

EN EL RAZONAMIENTO DEDUCTIVO VAMOS DE UNA HIPÓTESIS A UNA CONCLUSIÓN: «SI LORD FASTBACK COMETIERA UN ASESINATO, LIMPIARÍA LAS HUELLAS DACTILARES DE LA PISTOLA.»

EL RAZONAMIENTO INDUCTIVO, POR EL CONTRARIO, DISCURRE HACIA ATRÁS, DESDE UN CONJUNTO DE OBSERVACIONES A UNA HIPÓTESIS RAZONABLE:

MM... LA INSIGNIA DE LORD FASTBACK EN EL PAÑUELO Y ESA PISTOLA... FASTBACK ES EL ASESINO, WATSON. ¡ESTOY UN 95% SEGURO!

¡BRILLANTE INDUCCIÓN, HOLMES!

LA CIENCIA, TAMBIÉN LA ESTADÍSTICA, ES DE ALGÚN MODO UN TRABAJO DETECTIVESCO. EMPEZAMOS CON UN CONJUNTO DE OBSERVACIONES, Y NOS PREGUNTAMOS QUÉ SE PUEDE DECIR DE LOS SISTEMAS QUE LAS GENERARON.

LA ESTIMACIÓN

CON INTERVALOS DE CONFIANZA

ES UNA DE LAS FORMAS MÁS EFECTIVAS DE INFERENCIA ESTADÍSTICA, Y SE PUEDE VER A DIARIO ANTES DE UNAS ELECCIONES...

EN UNAS ELECCIONES RECIENTES, EN ALGÚN LUGAR, EL SENADOR ASTUTO ENCARGA UN SONDEO DE OPINIÓN A LA COMPAÑÍA GRANDES INVESTIGACIONES HOLMES. EL ENCUESTADOR HOLMES TOMA UNA MUESTRA ALEATORIA SIMPLE DE 1.000 VOTANTES Y LES PREGUNTA QUÉ OPINIÓN LES MERECE ASTUTO.

DESPUÉS DE CENSURAR LOS COMENTARIOS DE UNAS CUANTAS OBSERVACIONES EXTREMAS GRUÑONAS, HOLMES CREE QUE 550 VOTANTES ESTÁN A FAVOR DE SU CLIENTE, EL SENADOR ASTUTO.


¿SÓLO HA PREGUNTADO A 1.000 PERSONAS? PERO ¡¡SI HAY UN MILLÓN DE VOTANTES EN ESTE ESTADO!!
ME DA IGUAL QUE SEAN UN MILLÓN O UN BILLÓN...
¡HE HECHO UN MUESTREO ALEATORIO SIMPLE! ¡ASÍ QUE LE PUEDO OFRECER UNA FIRME GARANTÍA!
Y ESA GARANTÍA... ES DEL TIPO TOTAL Y CON REEMBOLSO, ¿VERDAD?
¿EH?
EH... BUENO... LO CIERTO ES QUE ES MÁS BIEN ESTADÍSTICA...
SÍ, PUEDO AFIRMAR CON UN 95% DE SEGURIDAD QUE LA VERDADERA PROPORCIÓN DE LA POBLACIÓN, p, ESTÁ ENTRE 0,519 Y 0,581.
¡¡¿¿UN 95% DE SEGURIDAD??!! ¡¡CIELOS! ¿Y QUÉ SUPONE QUE PASARÍA SI YO ME PRESENTARA A LAS ELECCIONES CON UN 95% DE HONESTIDAD?
NO LO SÉ. NUNCA ME HAN JUZGADO...
CUANDO ASTUTO SE CALMA, HOLMES LE EXPLICA LO QUE QUIERE DECIR CON UN 95% DE SEGURIDAD: SABE QUE SU PROCEDIMIENTO DE ESTIMACIÓN TIENE UNA PROBABILIDAD DE UN 95% DE PRODUCIR UN INTERVALO EN EL QUE SE ENCUENTRE p. ES DECIR, EN SU AÑOS COMO ENCUESTADOR, p ESTABA DENTRO DEL INTERVALO DE CONFIANZA QUE RODEA EL VALOR OBSERVADO, $\hat{p}$, EN UN 95% DE LAS OCASIONES.

EL SENADOR ASTUTO AÚN ESTÁ ALGO CONFUSO, ASÍ QUE HOLMES LE DA UNA CLASE DE **tiro al arco**.

¡DISPARA! LO QUE SEA QUE ME QUITE DE LA CABEZA LA ESTADÍSTICA DE LAS NARICES

VAMOS A CONSIDERAR A UNA ARQUERA QUE DISPARA A UNA DIANA. SUPONGAMOS QUE DA EN EL BLANCO DE 10 CENTÍMETROS UN 95% DE LAS VECES QUE DISPARA. ES DECIR, SÓLO UNA FLECHA DE CADA 20 NO DA EN EL BLANCO.

Y DETRÁS DE LA DIANA ENCONTRAMOS A UN VALEROSO DETECTIVE, QUE NO VE EL BLANCO. LA ARQUERA DISPARA UNA SOLA FLECHA.

EL DETECTIVE CONOCE EL NIVEL DE HABILIDAD DE LA ARQUERA Y DIBUJA UN CÍRCULO CON RADIO DE 10 CENTÍMETROS ALREDEDOR DE LA FLECHA. ¡TIENE UNA **SEGURIDAD DE UN 95%** DE QUE EN ESE CÍRCULO SE ENCUENTRA EL CENTRO DE LA DIANA!

10 CM

HA RAZONADO QUE SI DIBUJABA CÍRCULOS DE 10 CENTÍMETROS DE RADIO ALREDEDOR DE **MUCHAS FLECHAS**, EL BLANCO SE ENCONTRARÍA DENTRO DE ESOS CÍRCULOS UN 95% DE LAS VECES.

CENTRO AUTÉNTICO

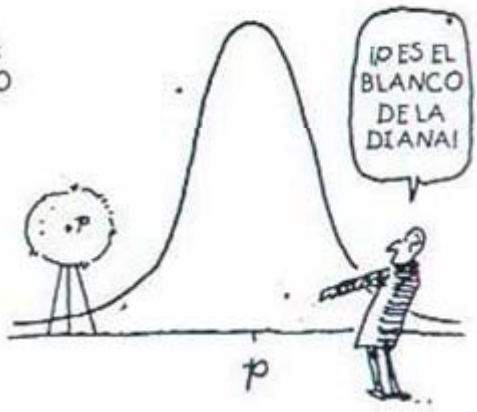
X = FLECHA

(LOS PROBABILISTAS UTILIZAN EL TÉRMINO ESTOCÁSTICO PARA DESCRIBIR LOS MODELOS ALEATORIOS. PROVIENE DEL GRIEGO **STOCHAZESTHAI**, QUE SIGNIFICA APUNTAR A UN OBJETIVO, O ADIVINAR, DE STOCHOS, OBJETIVO O DIANA.)

AHORA, HOLMES TRADUCE LA LECCIÓN DE TIRO AL ARCO AL LENGUAJE QUE DESARROLLAMOS EN EL CAPÍTULO ANTERIOR.

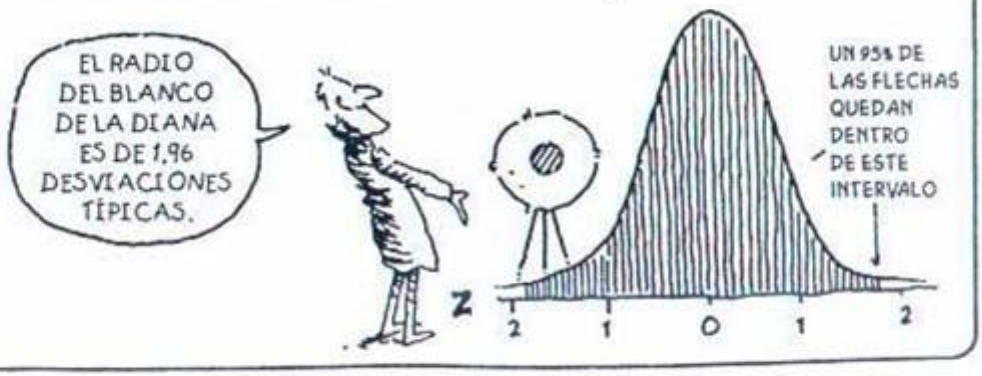
Primer paso: DISPARAR MUCHAS FLECHAS. CON UN CÁLCULO DE PROBABILIDAD SE DESCUBRE EL TAMAÑO DEL BLANCO DE LA «DIANA». LAS ESTIMACIONES DE $\hat{p}$ SON NUESTRAS FLECHAS. YA HEMOS VISTO QUE LA DISTRIBUCIÓN MUESTRAL DE $\hat{p}$ ES CASI **NORMAL**, CON MEDIA p Y DESVIACIÓN TÍPICA

$$\sigma(\hat{p}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$



COMO LA CURVA ES NORMAL, UTILIZAMOS LA TRANSFORMACIÓN z Y UNA TABLA ESTÁNDAR PARA ENCONTRAR LA AMPLITUD DEL INTERVALO EN EL QUE ESTÁN UN 95% DE LAS «FLECHAS». (DENTRO DE UNAS PÁGINAS, VEREMOS CÓMO HACERLO CON EXACTITUD.) LOS CÁLCULOS NOS DICEN QUE LA AMPLITUD ES DE 1,96 DESVIACIONES TÍPICAS:

$$0{,}95 = Pr\,(-1{,}96 \leq Z \leq 1{,}96)$$



AHORA TOCA UN POCO DE ÁLGEBRA. SEGÚN LA DEFINICIÓN DE LA TRANSFORMACIÓN z.

$$0{,}95 \simeq \Pr\left(-1{,}96 \leq \frac{\hat{p}-p}{\sigma(p)} \leq 1{,}96\right)$$

QUE SE CONVIERTE EN

$$0{,}95 \simeq \Pr\left(p-1{,}96\sigma(p) \leq \hat{p} \leq p+1{,}96\sigma(p)\right)$$

QUE ES SÓLO OTRA FORMA DE DECIR QUE EL 95% DE LAS «FLECHAS» $\hat{p}$ QUEDAN ENTRE $p-1{,}96\sigma(p)$ Y $p+1{,}96\sigma(p)$.

DESDE NUESTRA POSICIÓN VEMOS LA DIANA POR DETRÁS. OTRA VUELTA DE TUERCA DEL ÁLGEBRA LO CONVIERTE EN

$$0{,}95 \simeq \Pr\left(\hat{p}-1{,}96\sigma(p) \leq p \leq \hat{p}+1{,}96\sigma(p)\right)$$

AQUÍ ESTAMOS DIBUJANDO CÍRCULOS ALREDEDOR DE MUCHAS FLECHAS (ES DECIR, ESTABLECIENDO INTERVALOS ALREDEDOR DE $\hat{p}$) Y DECIMOS QUE UN 95% DE ELLOS INCLUYE p.

PERO NOS ENCONTRAMOS CON UN PROBLEMILLA... NO SABEMOS CUÁNTO MIDE EN REALIDAD EL BLANCO DE LA DIANA, PORQUE NO CONOCEMOS EL VALOR DE p, Y LA AMPLITUD DEBE SER UN MÚLTIPLO DE $\sigma(p)$.

ASÍ QUE NOS LO AMAÑAMOS UN POCO Y UTILIZAMOS EL ERROR ESTÁNDAR O TÍPICO (SE)* DE $\hat{p}$.

$$SE(\hat{p}) = \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

EN SU LUGAR... SE ACERCA BASTANTE... ES TODO LO QUE PODEMOS HACER... ¡Y HASTA TIENE UNA JUSTIFICACIÓN TEÓRICA!

* STANDARD ERROR = SE [N.T.]

AHORA LA FÓRMULA ES

$$0{,}95 \approx \Pr\left(\hat{p} - 1{,}96\,SE(\hat{p}) \le p \le \hat{p} + 1{,}96\,SE(\hat{p})\right)$$

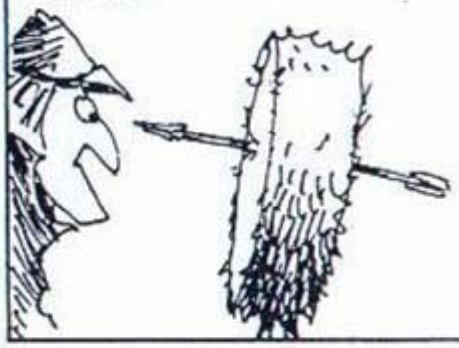
DE NUEVO, ESTA ECUACIÓN DESCRIBE LA PROBABILIDAD DE QUE LA AUTÉNTICA PROPORCIÓN DE LA POBLACIÓN FIJADA QUEDE DENTRO DEL INTERVALO ALEATORIO

$$\left(\hat{p} - 1{,}96\,SE(\hat{p}),\ \hat{p} + 1{,}96\,SE(\hat{p})\right).$$

SI LLEVÁSEMOS A CABO REPETIDAS MUESTRAS, ESTOS INTERVALOS INCLUIRÍAN p EN UN 95% DE LAS OCASIONES.

YA HEMOS HECHO EL CÁLCULO DE PROBABILIDADES Y HA LLEGADO LA HORA DEL...

Segundo paso: EL TRABAJO DETECTIVESCO. EN UNA ENCUESTA REAL, HOLMES SÓLO LLEVA A CABO UNA MUESTRA ALEATORIA SIMPLE DE 1.000 VOTOS, DESCUBRE QUE $\hat{p} = 0{,}550$, Y QUIERE INFERIR EL VALOR DE p.



ASÍ QUE UTILIZA EL PRIMER PASO PARA CALCULAR

$$SE(\hat{p}) = \sqrt{\frac{(0.55)(0.45)}{1000}} = 0.0157$$

Y LLEGA A LA CONCLUSIÓN DE QUE PODEMOS ESTAR UN 95% SEGUROS DE QUE p SE ENCUENTRA EN EL INTERVALO

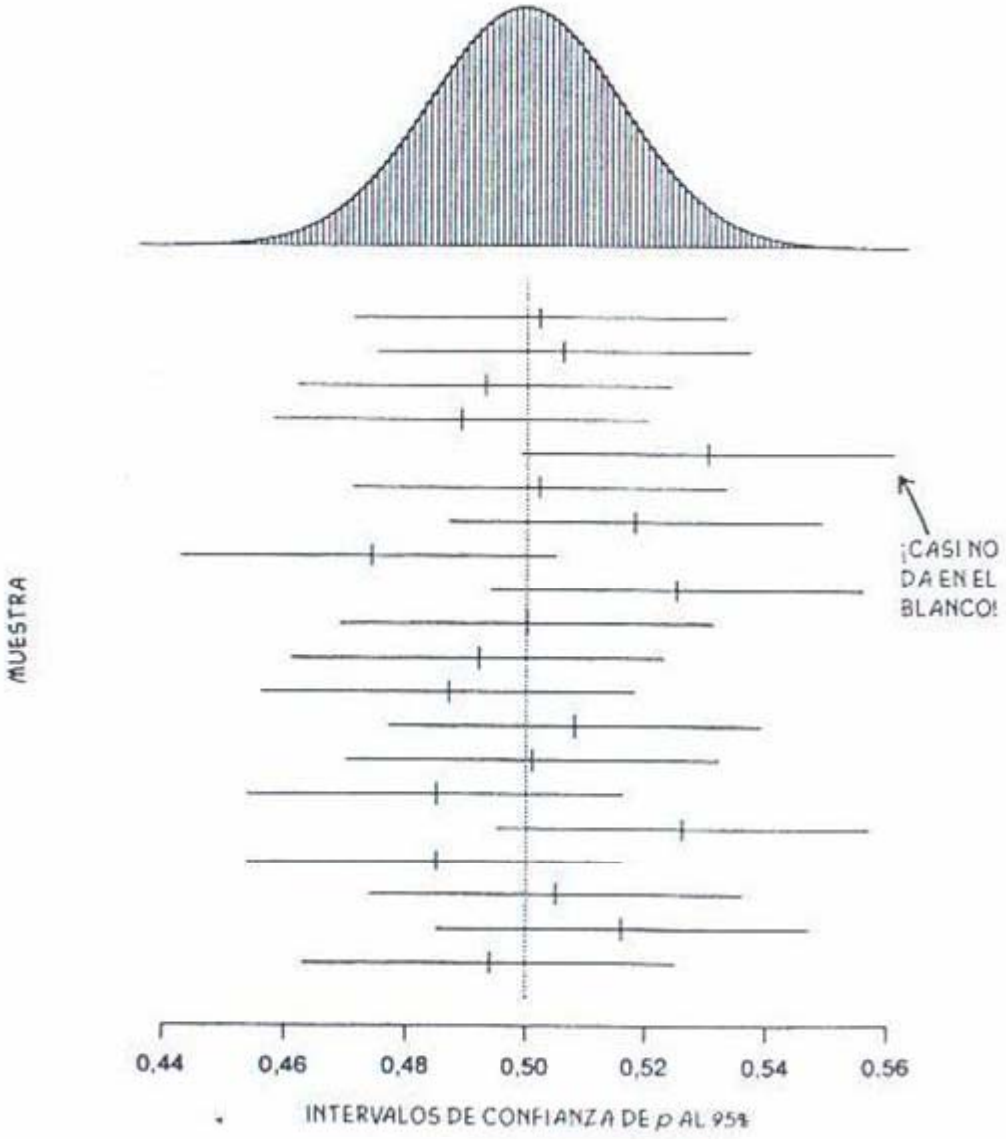
$$\hat{p} \pm 1{,}96\,SE(\hat{p})$$

$$= 0{,}550 \pm (1{,}96)(0{,}0157)$$

$$= 0{,}550 \pm 0{,}031$$

ESTO ES LO QUE QUIEREN DECIR LAS ENCUESTAS CUANDO SE REFIEREN A SU «MARGEN DE ERROR». EN ESTE CASO, HOLMES VIO QUE $0{,}519 \le p \le 0{,}581$, EN OTRAS PALABRAS, QUE $p = 55\%$ CON UN 3% DE MARGEN DE ERROR. (NORMALMENTE, LAS ENCUESTAS UTILIZAN UN 95% DE CONFIANZA.)

ESTA PÁGINA MUESTRA LOS RESULTADOS DE UNA SIMULACIÓN POR ORDENADOR DE VEINTE MUESTRAS DE TAMAÑO $n = 1.000$. SUPONEMOS QUE EL VALOR REAL DE $p = 0,5$. EN LA PARTE SUPERIOR PUEDES VER LA DISTRIBUCIÓN MUESTRAL DE $\hat{p}$ (NORMAL, CON MEDIA p Y $\sigma = \sqrt{\frac{p(1-p)}{n}}$). EN LA PARTE INFERIOR SE ENCUENTRAN LOS INTERVALOS DE CONFIANZA DE CADA MUESTRA, AL 95%. COMO MEDIA, UNO DE CADA VEINTE (O UN 5%) DE ESTOS INTERVALOS NO INCLUIRÁ EL PUNTO $p = 0,5$.



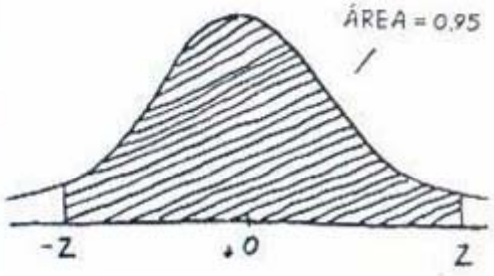
A PESAR DE QUE UN 95% DE CONFIANZA ESTÁ BASTANTE BIEN PARA LOS SONDEOS DE LA PRENSA, NO ES LO BASTANTE BUENO PARA EL SENADOR ASTUTO. ¡ÉL QUIERE UN 99%!

EL PRIMER MÉTODO EQUIVALE A AMPLIAR EL INTERVALO DE CONFIANZA. CUANTO MAYOR SEA EL MARGEN DE ERROR, MÁS SEGUROS PODEMOS ESTAR DE QUE EL VALOR REAL DE p SE ENCUENTRA EN EL INTERVALO.

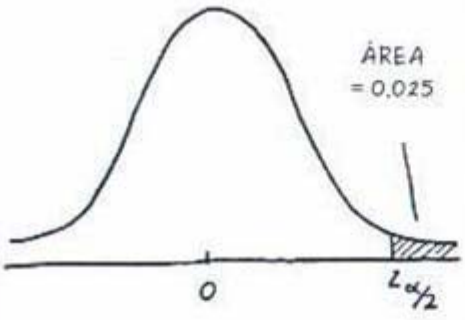
QUIZÁ HAYA LLEGADO LA HORA DE VER EXACTAMENTE CÓMO ENCONTRAR LOS EXTREMOS DE LOS INTERVALOS DE CONFIANZA...

AQUÍ, NORMALMENTE, AL NÚMERO IMPORTANTE LO LLAMAMOS α, Y MIDE LA DIFERENCIA ENTRE EL NIVEL DESEADO DE CONFIANZA Y CERTEZA. POR EJEMPLO, CUANDO EL NIVEL DE CONFIANZA ES 95%, O 0,95, α ES 0,05. ASÍ QUE HABLAMOS DEL NIVEL DE CONFIANZA $(1-\alpha)\cdot 100\%$.

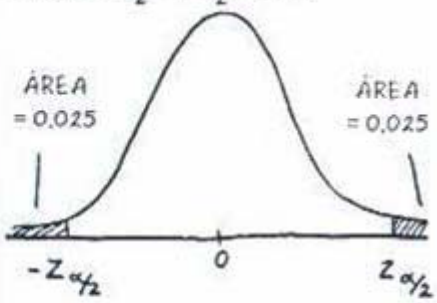
ENCONTRAR EL NIVEL DE CONFIANZA $(1-\alpha)\cdot 100\%$ IMPLICA OBSERVAR LA CURVA DE LA NORMAL TIPIFICADA Y BUSCAR LOS PUNTOS $\pm z$ ENTRE LOS QUE EL ÁREA ES $1-\alpha$.



ESTE PUNTO, LLAMADO $z_{\frac{\alpha}{2}}$, ES EL VALOR z MÁS ALLÁ DEL CUAL EL ÁREA ES $0{,}025 = \frac{\alpha}{2}$.



ESTO PASA PORQUE CORTAMOS LAS «COLAS» DE LOS DOS EXTREMOS DE LA CURVA, QUE TIENEN UN ÁREA TOTAL DE $\frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$.



PODEMOS CALCULAR $z_{\alpha/2}$ DIRECTAMENTE A PARTIR DE LA TABLA DE LA NORMAL TIPIFICADA (PÁGINA 84). ES EL PUNTO CON LA PROPIEDAD

$$Pr(z \geq z_{\alpha/2}) = \frac{\alpha}{2}$$

EN ESTE CASO,

$$Pr(z \geq z_{0,025}) = 0,025$$

z	-2,5	-2,4	-2,3	-2,2	-2,1
F(z)	0,006	0,008	0,011	0,014	0,018
z	-2,0	-1,9	-1,8	-1,7	-1,6
F(z)	0,023	0,029	0,036	0,045	0,055
z	-1,5				
F(z)	0,067	0,0			

AQUÍ TIENES UNA PEQUEÑA TABLA DE LOS VALORES CRÍTICOS DE VARIOS NIVELES DE CONFIANZA...

$1-\alpha$	0.80	0.90	0.95	0.99
α	0.20	0.10	0.05	0.01
$\alpha/2$	0.10	0.05	0.025	0.005
$z_{\frac{\alpha}{2}}$	1.28	1.64	1.96	2.58

PARA ESTABLECER UN INTERVALO DE CONFIANZA DEL 99%, UTILIZAMOS ESA TABLA Y ESCRIBIMOS

$$0.99 = Pr(\hat{p} - 2.58SE(\hat{p}) \leq p \leq \hat{p} + 2.58SE(\hat{p}))$$

Y LO ABREVIAMOS COMO

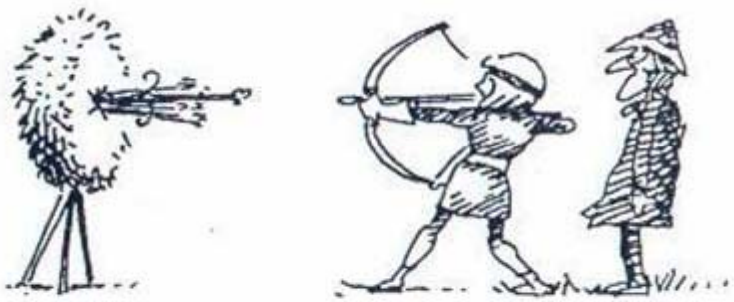
$$p = \hat{p} \pm 2.58\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$= 0.55 \pm 2.58\sqrt{\frac{(0.55)(0.45)}{1000}}$$

$$= 0,55 \pm 0,041$$

CON UNA CONFIANZA DEL 99%.

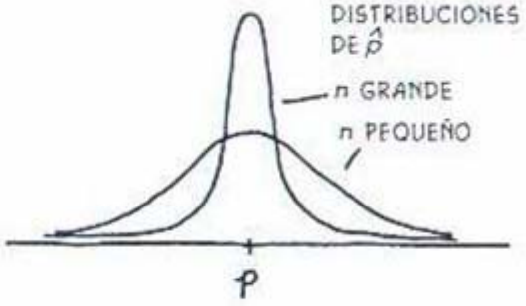
AUMENTAR EL INTERVALO ES UNA FORMA DE AUMENTAR LA CONFIANZA EN EL RESULTADO. COMO YA HEMOS DICHO ANTES, OTRA FORMA SERÍA DISPARAR LAS FLECHAS **CON MÁS PRECISIÓN**. SI SUPIÉRAMOS QUE LA ARQUERA CONSIGUE QUE UN 95% DE LAS FLECHAS DEN A 1 **CENTÍMETRO** DEL BLANCO DE LA DIANA, ¡NUESTRAS ESTIMACIONES PODRÍAN SER MUCHO MÁS PRECISAS!



¿CÓMO PODEMOS CONSEGUIRLO? ¡AUMENTANDO EL TAMAÑO DE LA MUESTRA! LA AMPLITUD DEL INTERVALO DE CONFIANZA DEPENDE DEL TAMAÑO MUESTRAL: EL INTERVALO TIENE LA FORMA $\hat{p} + E$, EN LA QUE E, EL ERROR, VIENE DADO POR

$$E = z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

ASÍ QUE CUANTO MAYOR SEA n, EL ERROR SERÁ MENOR. (ES DECIR, SI MULTIPLICAMOS n POR CUATRO, LA AMPLITUD DEL INTERVALO SE REDUCE A LA MITAD.)



ASTUTO LE PIDE A HOLMES QUE LE DÉ UN ERROR PEQUEÑO Y MUCHA CONFIANZA. PONGAMOS UN 99% DE CONFIANZA CON $E = \pm 0{,}01$. HOLMES CALCULA n.

$$n = \frac{z_{\frac{\alpha}{2}}^{\,2}\, p^*(1-p^*)}{E^2}$$

(p^* ES UNA APROXIMACIÓN A LA PROPORCIÓN REAL DE p; ¡RECUERDA QUE AÚN NO HEMOS REALIZADO EL MUESTREO!)

CON UNA SUPOSICIÓN CONSERVADORA DE $p^* = 0.5$, HOLMES DESCUBRE QUE

$$n = \frac{(2.58)^2(0.5)^2}{(0.01)^2}$$

$$= \frac{(6.65)(0.25)}{0.0001}$$

$$= 16,641$$

MIL VOTANTES DIERON UN 3% DE ERROR Y UNA CONFIANZA DE UN 95%. PARA CONSEGUIR UN ERROR DE UN 1% CON UNA CONFIANZA DEL 99%, ¡HOLMES TIENE QUE TOMAR UNA MUESTRA DE 16.641 VOTANTES!

ASÍ QUE LLEVAN A CABO LA ENCUESTA Y SE PRESENTAN A LAS ELECCIONES CON UN 99% DE CONFIANZA.

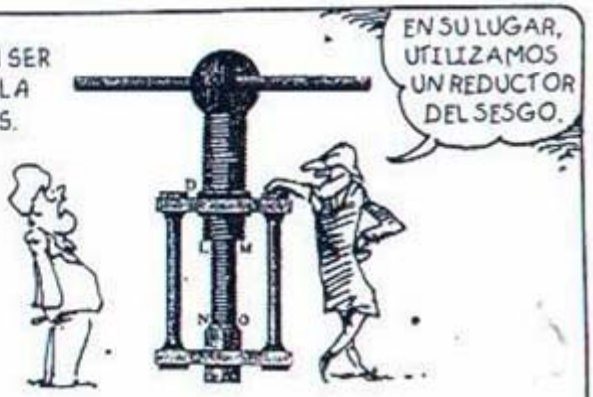
SIN EMBARGO... TODO ESTO DE LAS PROBABILIDADES SÓLO SIRVE ANTES DE UNAS ELECCIONES. ¡DESPUÉS, EL SENADOR ESTÁ ELEGIDO AL 100% O NO ELEGIDO AL 100%! Y A PESAR DE TODO, EL SENADOR ASTUTO PIERDE LAS ELECCIONES...

¡LO QUE HA PASADO ES QUE A LOS POLÍTICOS NO LOS ELIGEN LOS SONDEOS DE OPINIÓN!

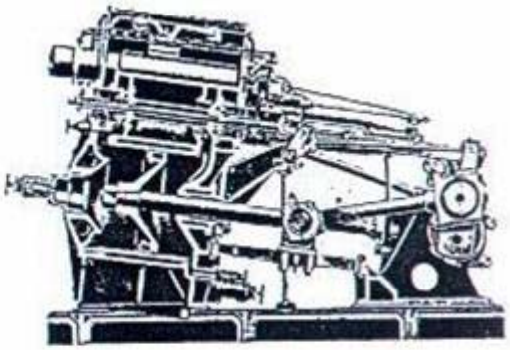
ALGUNOS DE LOS PROBLEMAS DE LOS SONDEOS FRENTE A LAS ELECCIONES SON:

NO EXISTE FORMA ALGUNA DE METERSE EN LA CABEZA DEL VOTANTE EN POTENCIA Y SABER SI VA A VOTAR, SI MIENTE O SI VA A CAMBIAR DE OPINIÓN ANTES DEL DÍA DE LAS ELECCIONES. INCLUSO CON MUESTRAS MAYORES NO SE PUEDE REDUCIR ESTE TIPO DE ERRORES.

Y A QUE ESTOS ERRORES PUEDEN SER MUY GRANDES, APENAS MERECE LA PENA PAGAR MUESTRAS MAYORES.



EN LAS ÚLTIMAS CINCO ELECCIONES PRESIDENCIALES DE LOS ESTADOS UNIDOS, EL SONDEO GALLUP HIZO UNA ENCUESTA CON MENOS DE 4.000 VOTANTES CADA VEZ. SIN EMBARGO, EN LAS CINCO ELECCIONES, LOS ERRORES DE LA ORGANIZACIÓN GALLUP EN SU PREDICCIÓN DE LOS RESULTADOS FUERON DE MENOS DE UN 2%.

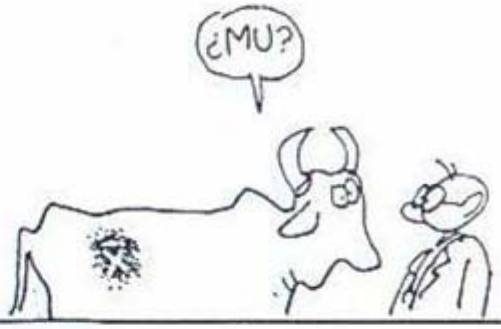


SU ÉXITO SE DEBE A LA UTILIZACIÓN DE ESTIMADORES QUE COMPENSAN LA FALTA DE RESPUESTA Y A LA DESESTIMACIÓN DE LOS VOTANTES QUE PROBABLEMENTE NO EJERCERÁN SU DERECHO EN LAS URNAS.

EN RESUMEN, PROPORCIÓN ESTIMADA = PROPORCIÓN REAL + SESGO + ERROR DE MUESTREO ALEATORIO.
INCLUSO LOS ENCUESTADORES TIENEN FONDOS LIMITADOS, ASÍ QUE DECIDEN SABIAMENTE GASTAR EL DINERO EN REDUCIR EL SESGO EN LUGAR DE INTENTAR AUMENTAR LA MUESTRA A MÁS DE 4.000 VOTANTES.

Intervalos de confianza de μ

HASTA AHORA, HEMOS VISTO INTERVALOS DE CONFIANZA DE UNA PROPORCIÓN p DE UNA POBLACIÓN. EXACTAMENTE EL MISMO RAZONAMIENTO FUNCIONA PARA LA MEDIA POBLACIONAL μ.



EN EL CAPÍTULO ANTERIOR (PÁGINA 105), VIMOS QUE LA DISTRIBUCIÓN DE LA MEDIA MUESTRAL $\bar{X}$ ES APROXIMADAMENTE NORMAL, CON EL CENTRO EN LA **AUTÉNTICA** MEDIA POBLACIONAL μ Y DESVIACIÓN TÍPICA $\sigma/\sqrt{n}$, EN LA QUE σ ES LA DESVIACIÓN TÍPICA DE LA POBLACIÓN. DE ESTE MODO, PARA UN VALOR n MUY GRANDE,

$$0{,}95 = Pr(-1.96 \leq Z \leq 1.96)$$

$$= Pr\left(-1.96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1.96\right)$$

DE NUEVO, YA QUE NO CONOCEMOS EL VALOR DE σ, LA SUSTITUIMOS POR s, LA DESVIACIÓN TÍPICA MUESTRAL:

$$0.95 \approx Pr\left(-1.96 \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq 1.96\right)$$

EL TÉRMINO $s/\sqrt{n}$ RECIBE EL NOMBRE DE **ERROR TÍPICO DE MUESTREO**, Y SE ESCRIBE $SE(\bar{X})$. LLEGAMOS A LA SIGUIENTE CONCLUSIÓN:

$$0.95 \approx Pr(\bar{X} - 1.96\,SE(\bar{X}) \leq \mu \leq \bar{X} + 1.96\,SE(\bar{X}))$$

DONDE

$$SE(\bar{X}) = \frac{s}{\sqrt{n}}$$

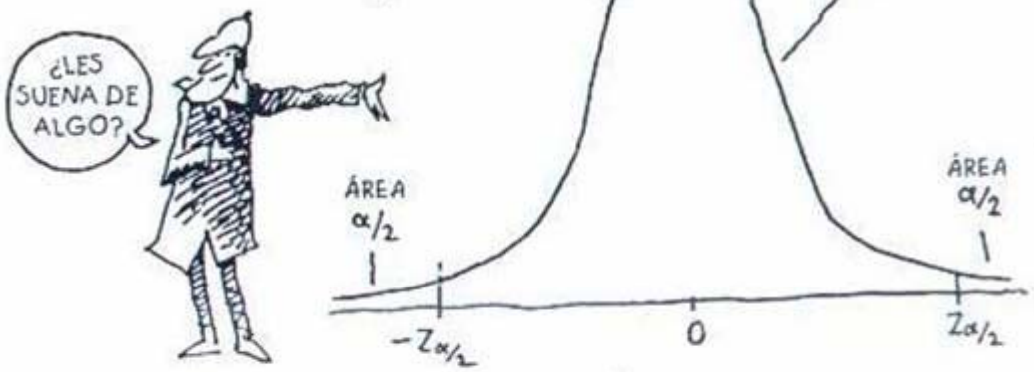
IGUAL QUE ANTES, HEMOS DESCUBIERTO QUE EL INTERVALO ALEATORIO

$$\bar{X} \pm 1.96\,SE(\bar{X})$$

INCLUYE LA MEDIA REAL, μ, CON UNA PROBABILIDAD DE 0,95... ASÍ QUE AHORA PODEMOS LLAMAR A SHERLOCK PARA QUE HAGA UNA INFERENCIA ESTADÍSTICA BASADA EN UNA SOLA MUESTRA DE TAMAÑO n Y MEDIA $\bar{x}$.

ÉL (Y NOSOTROS) ESTÁ UN 95% SEGURO DE QUE LA MEDIA μ SE ENCUENTRA EN EL INTERVALO $\bar{x} \pm 1{,}96SE(X)$.

IGUAL QUE ANTES, PARA OBTENER UN NIVEL ARBITRARIO DE SEGURIDAD $1-\alpha$, SUSTITUIMOS 1,96 POR $z_{\frac{\alpha}{2}}$.



VAMOS A REGRESAR A LOS DATOS DE LOS ESTUDIANTES DEL CAPÍTULO 2, SUPONIENDO QUE $n = 92$ ESTUDIANTES FUERA UNA MUESTRA ALEATORIA SIMPLE DE TODOS LOS ESTUDIANTES DEL ESTADO DE PENNSYLVANIA.

LA MEDIA MUESTRAL $\bar{x}$ ERA 145,2 LIBRAS Y LA DESVIACIÓN TÍPICA MUESTRAL s ERA 23,7. ASÍ QUE EL ERROR TÍPICO ES

$$SE(\bar{x}) = \frac{23,7}{\sqrt{92}} = 2,47$$

Y AHORA TENEMOS UNA CONFIANZA DEL 95% DE QUE EL PESO MEDIO DE TODOS LOS ESTUDIANTES DE ESE ESTADO QUEDA DENTRO DEL INTERVALO

$$\bar{x} \pm 1,96 SE(\bar{x})$$
$$= 145,2 \pm (1,96)(2,47)$$
$$= 145,2 \pm 4,8 \text{ LIBRAS}$$

EN RESUMEN: EN UNA MUESTRA ALEATORIA SIMPLE (MAS) DE TAMAÑO GRANDE, EL INTERVALO DE CONFIANZA $(1-\alpha)\cdot 100\%$ ES:

MEDIA POBLACIONAL μ	PROPORCIÓN POBLACIONAL, p
$\mu = \bar{x} \pm z_{\frac{\alpha}{2}} SE(\bar{x})$	$p = \hat{p} \pm z_{\frac{\alpha}{2}} SE(\hat{p})$
DONDE $SE(\bar{x}) = s/\sqrt{n}$	DONDE $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

EL TAMAÑO DE LOS DOS INTERVALOS ESTÁ CONTROLADO POR EL NIVEL DE CONFIANZA $(1-\alpha)\cdot 100\%$ Y EL TAMAÑO DE LA MUESTRA, n.

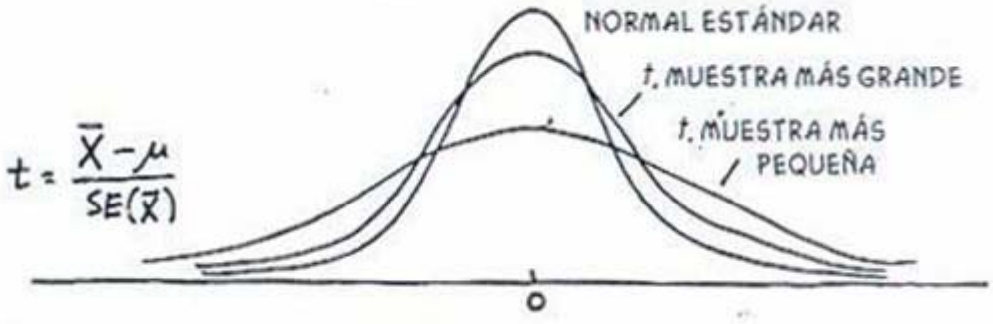
La *t* de Student (¡otra vez!)

COMO YA VIMOS EN EL CAPÍTULO 6, EL ESTADÍSTICO

$$\frac{\bar{X}-\mu}{SE(\bar{X})}$$

SÓLO TIENE UNA DISTRIBUCIÓN APROXIMADAMENTE NORMAL CUANDO SE CALCULA UTILIZANDO UNA MUESTRA MUY GRANDE. PARA MUESTRAS MÁS PEQUEÑAS ($n = 5, 10, 25...$), ÉSE YA NO ES EL CASO Y TENEMOS QUE UTILIZAR LA *t* DE STUDENT.

VAMOS A OBSERVAR MÁS DE CERCA LA *t*. YA MENCIONAMOS QUE LA DISTRIBUCIÓN *t* ES MÁS DISPERSA QUE LA NORMAL, Y QUE LA CANTIDAD DE DISPERSIÓN DEPENDE DEL TAMAÑO DE LA MUESTRA.



LO QUE HIZO SU DESCUBRIDOR, GOSSET, FUE CUANTIFICAR ESTA RELACIÓN. SI *n* ES EL TAMAÑO MUESTRAL, DECÍA, ENTONCES LLAMAREMOS A $n-1$ EL NÚMERO DE

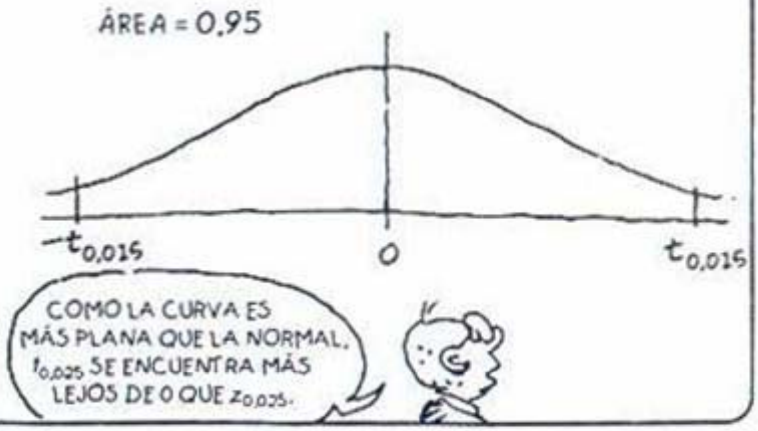
grados de libertad

DE LA MUESTRA.

GOSSET CALCULÓ TABLAS DE LA DISTRIBUCIÓN t PARA DIFERENTES TAMAÑOS MUESTRALES: ES DECIR, GRADOS DE LIBERTAD. Y NOSOTROS REPETIMOS, A MÁS GRADOS DE LIBERTAD, MÁS CERCA SE ENCUENTRA t DE LA NORMAL TIPIFICADA.

SI CONOCEMOS EL TAMAÑO MUESTRAL n, ESCOGEMOS LA DISTRIBUCIÓN t CON $n-1$ GRADOS DE LIBERTAD.

IGUAL QUE CON LA DISTRIBUCIÓN z (LA NORMAL TIPIFICADA), OBTENEMOS UN NIVEL DE CONFIANZA DEL 95% AL ENCONTRAR EL VALOR CRÍTICO $t_{0,025}$ MÁS ALLÁ DEL CUAL EL ÁREA DE LA CURVA ES 0,025.



PARA OBTENER UN INTERVALO DE CONFIANZA $(1-\alpha)\cdot 100\%$, ENCONTRAMOS UN VALOR CRÍTICO $t_{\frac{\alpha}{2}}$ TAL QUE $Pr(t \geq t_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$. AQUÍ TIENES UNA PEQUEÑA TABLA DE VALORES CRÍTICOS PARA LA DISTRIBUCIÓN t:

	$1-\alpha$	0,80	0,90	0,95	0,99
	α	0,20	0,10	0,05	0,01
	$\alpha/2$	0,10	0,05	0,025	0,005
GRADOS DE LIBERTAD	1	3,09	6,31	12,71	63,66
	10	1,37	1,81	2,23	4,14
	30	1,31	1,70	2,04	2,75
	100	1,29	1,66	1,98	2,63
	∞	1,28	1,65	1,96	2,58

CADA COLUMNA REPRESENTA UN NIVEL FIJO DE CONFIANZA, CON NÚMEROS CRECIENTES DE GRADOS DE LIBERTAD. CUANTOS MÁS GRADOS DE LIBERTAD, MÁS SE ACERCA EL VALOR CRÍTICO A $z_{\alpha/2}$, EL VALOR CRÍTICO DE LA DISTRIBUCIÓN NORMAL.

DERIVAMOS LA AMPLITUD DE NUESTRO INTERVALO DE CONFIANZA DIRECTAMENTE DE LA DEFINICIÓN DE t:

$$t = \frac{\bar{X} - \mu}{SE(\bar{X})}$$

ENTONCES, PARA OBTENER UN NIVEL DE CONFIANZA $(1-\alpha)\cdot 100\%$,

$$(1-\alpha) = Pr(\bar{x} - t_{\frac{\alpha}{2}} SE(\bar{X}) \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2}} SE(\bar{X}))$$

DE LO QUE INFERIMOS: DADA UNA SOLA MUESTRA DE TAMAÑO n Y MEDIA $\bar{x}$, PODEMOS ESTAR $(1-\alpha)\cdot 100\%$ SEGUROS DE QUE LA MEDIA POBLACIONAL μ QUEDA EN EL INTERVALO

$$\mu = \bar{x} \pm t_{\frac{\alpha}{2}} SE(\bar{x})$$

DONDE $SE(\bar{x}) = s/\sqrt{n}$ Y $t_{\frac{\alpha}{2}}$ ES EL VALOR CRÍTICO DE LA DISTRIBUCIÓN t CON $n-1$ GRADOS DE LIBERTAD.

NOTA: SI HABLAMOS CON EXACTITUD, LA DERIVACIÓN DE LA DISTRIBUCIÓN t DEPENDE DE LA PRESUNCIÓN DE QUE LA MUESTRA ERA DE UNA POBLACIÓN NORMAL. EN LA PRÁCTICA, LOS INTERVALOS DE CONFIANZA BASADOS EN t DAN RESULTADOS BASTANTE BUENOS, INCLUSO CUANDO LA DISTRIBUCIÓN POBLACIONAL TIENE UNA FORMA SÓLO APROXIMADAMENTE PARECIDA A UNA MONTAÑA.

Ejemplo: SUPONGAMOS QUE LA CAMALEÓN MOTORS TIENE QUE HACER PRUEBAS DE CHOQUE CON SUS COCHES PARA DETERMINAR EL COSTE MEDIO DE REPARACIÓN TRAS UNA COLISIÓN FRONTAL A UNOS 20 KILÓMETROS POR HORA. ¡RESULTA MUY CARO! ASÍ QUE DECIDEN PROBAR CON SÓLO CINCO CAMALEONES.

LOS DATOS DE LOS DESPERFECTOS EN DÓLARES SON 150, 400, 720, 500 Y 930.

LA MEDIA MUESTRAL:

$\bar{x} = 540$ DÓLARES

LA DESVIACIÓN ESTÁNDAR:

$s = 299$ DÓLARES

PUEDES COMPROBAR s CON UNA CALCULADORA. ES

$$\sqrt{\frac{1}{4}\left((150-540)^2+(400-540)^2+(720-540)^2+(500-540)^2+(930-540)^2\right)}$$

ASÍ QUE, ¿DÓNDE PODEMOS SITUAR LA MEDIA CON UNA CONFIANZA DEL 95%? ENCONTRAMOS NUESTRO VALOR CRÍTICO $t_{0.025}$ CON 4 GRADOS DE LIBERTAD:

	$1-\alpha$	0.80	0.90	0.95	0.99
	α	0.20	0.10	0.05	0.01
	$\alpha/2$	0.10	0.05	0.025	0.005
GRADOS DE LIBERTAD	1	3.09	6.31	12.71	63.66
	2	1.89	2.92	4.30	9.92
	3	1.64	2.35	3.18	5.84
	4	1.53	2.13	2.78	4.60
	5	1.48	2.01	2.57	4.03

Y ALLÁ VAMOS:

$$\mu = \bar{x} \pm 2.78\, s/\sqrt{n}$$
$$= 540 \pm 2.78\,(299/\sqrt{5})$$
$$= 540 \pm 372$$

ASÍ QUE LO MEJOR QUE PODEMOS DECIR CON UN 95% DE CONFIANZA ES QUE EL COSTE MEDIO POR REPARACIÓN DE DAÑOS ESTARÁ ENTRE 168 Y 912 DÓLARES.

LA COMPAÑÍA PUEDE DARSE POR SATISFECHA, O REALIZAR MÁS PRUEBAS...

PARA CALCULAR ESTE INTERVALO DE CONFIANZA UTILIZANDO LA *t* DE STUDENT, HEMOS REALIZADO UNA **PRESUNCIÓN NO COMPROBADA**: HEMOS DADO POR HECHO QUE LOS COSTES DE REPARACIÓN DE LOS COCHES TIENEN UNA **DISTRIBUCIÓN APROXIMADAMENTE NORMAL**. ES DECIR, QUE SI HICIÉSEMOS CHOCAR 1.000 CAMALEONES, EL HISTOGRAMA DE LOS COSTES DE REPARACIÓN SERÍA SIMÉTRICO Y CON FORMA DE MONTAÑA. **NO PODEMOS SABERLO** CON TAN SÓLO CINCO DATOS... PERO QUIZÁ AÑOS DE EXPERIENCIA CON ANTIGUOS MODELOS HAN PRODUCIDO HISTOGRAMAS DE LOS COSTES DE REPARACIÓN DE LA PARTE FRONTAL DEL COCHE CON DISTRIBUCIÓN NORMAL: UNA INFORMACIÓN QUE APOYARÍA NUESTRA DECISIÓN DE UTILIZAR LA *t* DE STUDENT.

PARA RESUMIR (!), YA TENEMOS TRES SIMPLES RECETAS PARA ENCONTRAR LOS INTERVALOS DE CONFIANZA. PARA LA PROPORCIÓN, O LA MEDIA DE MUESTRAS DE GRAN TAMAÑO, BUSCAMOS $z_{\frac{\alpha}{2}}$ EN UNA TABLA NORMAL. PARA LA MEDIA DE UNA MUESTRA PEQUEÑA (DIGAMOS $n \leq 30$), BUSCAMOS $t_{\frac{\alpha}{2}}$ EN LA TABLA t.

EN TODOS LOS CASOS, LA AMPLITUD DEL INTERVALO ES EL VALOR CRÍTICO POR EL ERROR ESTÁNDAR:

$$z_{\frac{\alpha}{2}} SE(\hat{p}) \qquad z_{\frac{\alpha}{2}} SE(\bar{X}) \qquad t_{\frac{\alpha}{2}} SE(\bar{X})$$

Y TODOS Y CADA UNO DE ESOS ERRORES ESTÁNDAR SON PROPORCIONALES A AQUEL NÚMERO MÁGICO:

$$\frac{1}{\sqrt{n}}$$

♦ Capítulo 8 ♦

CONTRASTE DE HIPÓTESIS

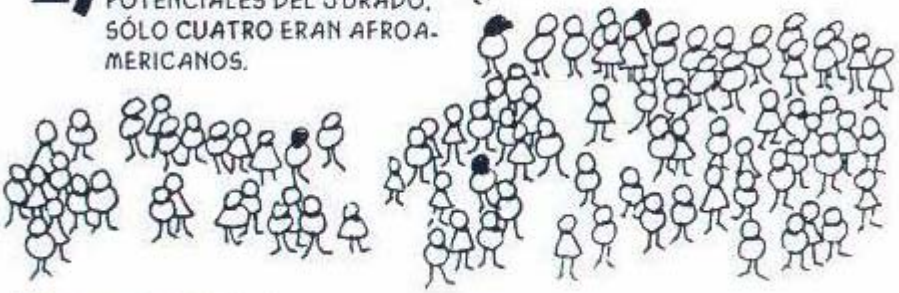
AHORA NOS ADENTRAREMOS EN NUEVOS TERRENOS... LA POLÍTICA, LA ECONOMÍA, Y LAS CIENCIAS EXACTAS Y LAS NO TAN EXACTAS ABUSAN MUY A MENUDO DE ESTAS PRUEBAS DE SIGNIFICACIÓN. Y PARA DESCUBRIR EL POR QUÉ, DEBEMOS PREGUNTARNOS: **«¿ESTAS OBSERVACIONES PUEDEN HABERSE DADO POR CASUALIDAD?»**

EMPEZAMOS CON UN EJEMPLO DE LA LEY: UN CONJUNTO DE DIFERENTES CASOS, QUE SE DIERON EN EL SUR DE LOS ESTADOS UNIDOS ENTRE 1960 Y 1980, EN LOS QUE TESTIGOS EXPERTOS DENUNCIABAN LA EXISTENCIA DE **DISCRIMINACIÓN RACIAL** EN LA SELECCIÓN DEL JURADO.

EN TEORÍA, LAS LISTAS DE JURADOS SE ELABORAN DE FORMA ALEATORIA A PARTIR DE UNA LISTA DE CIUDADANOS SUSCEPTIBLES DE SER ELEGIDOS. SIN EMBARGO, DURANTE LAS DÉCADAS DE LOS 50 Y 60, EN LOS ESTADOS DEL SUR, HABÍA MUY POCOS AFROAMERICANOS EN LAS LISTAS DE JURADOS. ASÍ QUE ALGUNOS ABOGADOS DE LA DEFENSA PUSIERON EN TELA DE JUICIO LOS VEREDICTOS. EN LA APELACIÓN, UN TESTIGO EXPERTO EN ESTADÍSTICA PRESENTÓ ESTA PRUEBA:

1) EL **50%** DE LOS CIUDADANOS SUSCEPTIBLES DE SER ELEGIDOS ERAN AFROAMERICANOS.

2) EN UNA LISTA DE 80 MIEMBROS POTENCIALES DEL JURADO, SÓLO **CUATRO** ERAN AFROAMERICANOS.



¿PUEDE SER ESTO FRUTO DE LA **PURA CASUALIDAD**?

PARA COMPLETAR EL ARGUMENTO, SUPONGAMOS QUE LA ELECCIÓN DEL JURADO POTENCIAL FUE **ALEATORIA**. ENTONCES, EL NÚMERO DE AFROAMERICANOS DE LA LISTA DE 80 PERSONAS SERÍA LA VARIABLE ALEATORIA **BINOMIAL** X CON $n = 80$ PRUEBAS Y $p = 0,5$.

ENTONCES, LA POSIBILIDAD DE FORMAR UN JURADO CON HASTA 4 MIEMBROS AFROAMERICANOS ES $Pr(x \leq 4)$ O SEA, ALREDEDOR DE 0,00000000000000000014 (!).

COMO LA PROBABILIDAD ES TAN PEQUEÑA, LA LISTA EN CUESTIÓN, CON SÓLO CUATRO AFROAMERICANOS, RESULTA UNA **PRUEBA** DE PESO CONTRA LA **HIPÓTESIS** DE LA SELECCIÓN ALEATORIA.

PARA HABLAR EN TÉRMINOS MÁS FAMILIARES, EL ESTADÍSTICO SEÑALA QUE ESTA PROBABILIDAD ES MENOR QUE LA DE CONSEGUIR TRES **ESCALERAS REALES** SEGUIDAS EN EL POKER.

POR ESO EL JUEZ **RECHAZA** LA **HIPÓTESIS** DE LA SELECCIÓN ALEATORIA.

REPASEMOS, DE NUEVO, TODO EL PROCESO PARA ENTENDER LOS CUATRO PASOS DEL CONTRASTE DE HIPÓTESIS ESTADÍSTICO.

Paso n.º1. FORMULAR TODAS LAS HIPÓTESIS

H_0, LA HIPÓTESIS NULA. LAS OBSERVACIONES SON EL RESULTADO DE LA PURA CASUALIDAD.

H_a, LA HIPÓTESIS ALTERNATIVA. HAY UN EFECTO REAL. LAS OBSERVACIONES SON EL RESULTADO DE ESTE EFECTO REAL, ADEMÁS DE LA VARIACIÓN CASUAL.

EN EL CASO DEL JURADO, H_0 INDICA QUE EL JURADO FUE ELEGIDO ALEATORIAMENTE ENTRE LA TOTALIDAD DE LA POBLACIÓN. LOS AFROAMERICANOS TIENEN UNA PROBABILIDAD $p = 0{,}5$ DE SER ELEGIDOS.

H_a INDICA QUE LOS AFROAMERICANOS TIENEN MENOS POSIBILIDADES DE SER ELEGIDOS PARA LA LISTA DE JURADOS QUE LA PROPORCIÓN DE LOS MISMOS EN LA POBLACIÓN: $p < 0{,}50$.

Paso n.º2. ANÁLISIS ESTADÍSTICO

ENCONTRAR UN ESTADÍSTICO QUE CONFIRME LA EVIDENCIA CONTRA LA HIPÓTESIS NULA.

EN EL CASO DEL JURADO LA PRUEBA ESTADÍSTICA ES LA VARIABLE ALEATORIA BINOMIAL X CON $p = 0{,}5$ Y $n = 80$.

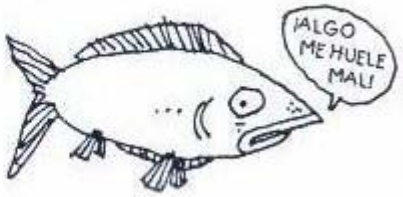
Paso n.°3. VALOR *p*: UNA AFIRMACIÓN DE PROBABILIDAD QUE RESPONDE A LA PREGUNTA: SI LA HIPÓTESIS NULA FUERA VERDADERA, ENTONCES ¿CUÁL SERÍA LA PROBABILIDAD DE OBSERVAR UN VALOR DE LA PRUEBA ESTADÍSTICA TAN EXAGERADO COMO EL QUE HEMOS VISTO?

Paso n.°4. COMPARAR EL VALOR *p* CON UN RIESGO α FIJO.

α ACTÚA COMO UN PUNTO DE CORTE POR DEBAJO DEL CUAL ACEPTAMOS QUE UN EFECTO ES ESTADÍSTICAMENTE SIGNIFICATIVO. O SEA, SI

VALOR P $\leq \alpha$

ENTONCES, DESCARTAMOS LA HIPÓTESIS NULA H_0 Y DECIDIMOS QUE PASA ALGO RARO.



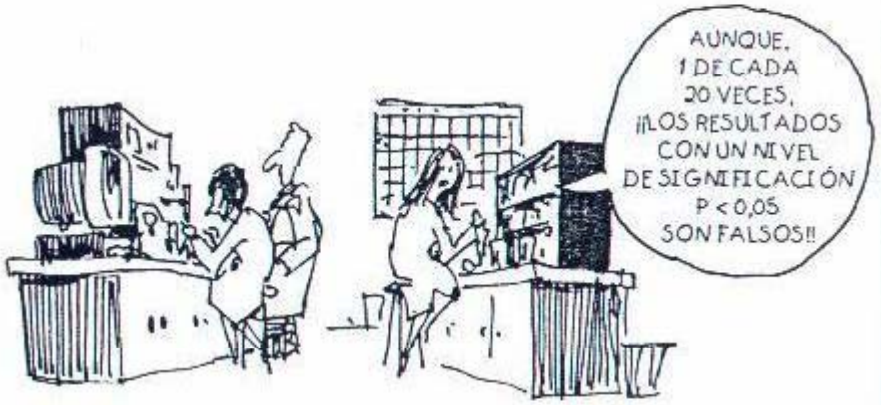
EN EL EJEMPLO, EL VALOR *p* ERA

$$Pr(x \leq 4 \mid p = 0{,}50 \text{ Y } n = 80) = 1{,}4 \times 10^{-16}$$

CALCULAMOS ESTE VALOR *p* DE FORMA MODERNA, USANDO UN PAQUETE DE SOFTWARE ESTADÍSTICO.

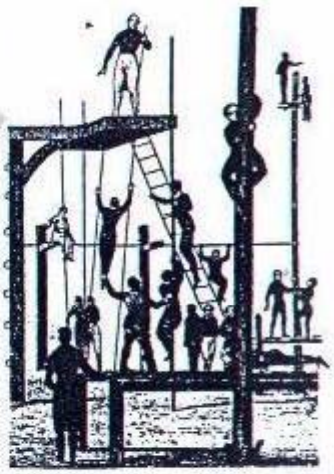
EN EL CASO DEL JURADO, EL ESTADÍSTICO ENCONTRÓ $p = 3{,}6 \times 10^{-16}$, EL MISMO NÚMERO DE OPORTUNIDADES DE QUE TE SALGAN TRES ESCALERAS REALES SEGUIDAS.

EN LOS ESTUDIOS CIENTÍFICOS, SE USA CON FRECUENCIA UN VALOR α FIJO DE 0,05 O 0,01. PODEMOS DECIR QUE ESTOS VALORES FIJOS SON RELIQUIAS DE LA ERA PREINFORMÁTICA, CUANDO NOS REFERÍAMOS A TABLAS QUE SÓLO SE PUBLICABAN PARA DETERMINADOS VALORES CRÍTICOS. AÚN HOY, EN ALGUNAS PUBLICACIONES CIENTÍFICAS SÓLO APARECEN LOS RESULTADOS SI EL VALOR $p \leq 0,05$.



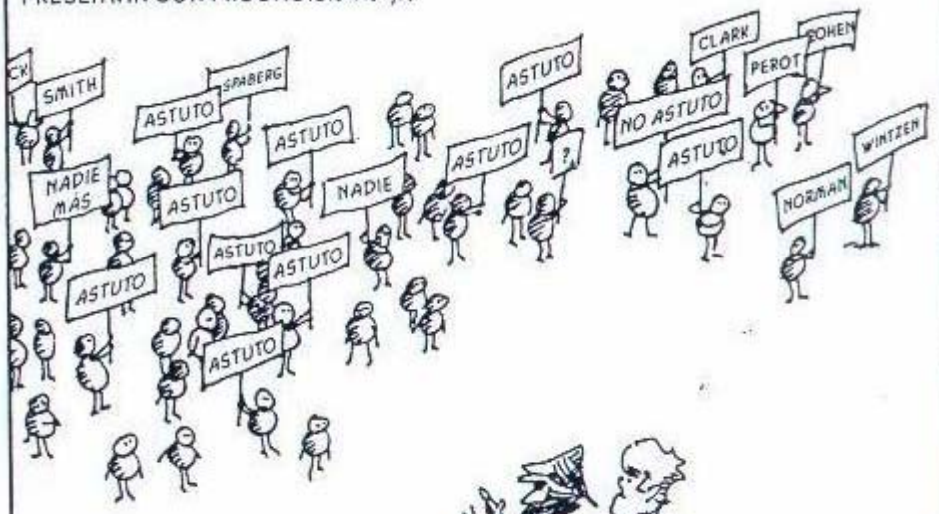
EN LOS PROCEDIMIENTOS LEGALES, EL ESTÁNDAR ES MÁS FLEXIBLE

MUESTRA GRANDE
PRUEBA DE SIGNIFICACIÓN PARA PROPORCIONES



EL EJEMPLO DEL JURADO ERA UN CASO ESPECÍFICO DE UN PROBLEMA GENERAL. LA HIPÓTESIS NULA ERA $p = p_0$. DONDE p_0 ERA UNA PROBABILIDAD (EN ESTE CASO, 0,5). VEAMOS AHORA PROBLEMAS COMO ÉSTE DESDE UN PUNTO DE VISTA GENERAL. CONTRASTEMOS LA HIPÓTESIS $p = p_0$.

COMO SIEMPRE, IMAGINAMOS TENER UNA POBLACIÓN NUMEROSA... OBSERVAMOS UNA MUESTRA GRANDE... Y VEMOS QUE ALGUNAS CARACTERÍSTICAS SE PRESENTAN CON PROBABILIDAD $\hat{p}$.



BASÁNDONOS EN ESTA OBSERVACIÓN, QUEREMOS SABER SI LA PROBABILIDAD REAL POBLACIONAL ES (POR EJEMPLO) MAYOR QUE LA DE OTRO VALOR p_0.
POR EJEMPLO, AL SENADOR ASTUTO, QUE TIENE UNA $\hat{p}$ DE 0,55 LE GUSTARÍA SABER QUE $p > 0,5$, O SEA, MAYORÍA ABSOLUTA.

Paso n.º1.
LA HIPÓTESIS NULA ES

$$H_0 : p = p_0$$

LA HIPÓTESIS ALTERNATIVA DEPENDE DE LA DIRECCIÓN TOMADA POR EL EFECTO QUE ESTAMOS BUSCANDO. EN EL CASO DEL SENADOR ASTUTO,

$$H_a : p > p_0$$

PERO, EN OTROS CASOS LA HIPÓTESIS ALTERNATIVA PODRÍA SER:

$$H_a : p < p_0$$

O

$$H_a : p \neq p_0$$

EN EL EJEMPLO DE LA SELECCIÓN DEL JURADO, LA HIPÓTESIS ALTERNATIVA ERA

$$H_a : p < 0.5$$

Y OTRAS VECES, NOS INTERESA SABER QUE p ES DIFERENTE A ALGÚN VALOR p_0. POR EJEMPLO, EN EL CONTRASTE DEL LANZAMIENTO DE UNA MONEDA, LA HIPÓTESIS ALTERNATIVA ES

$$H_a : p \neq 0.5$$

PERO NO TENEMOS UNA OPINIÓN A PRIORI DE SI SALDRÁN MÁS CARAS O CRUCES.

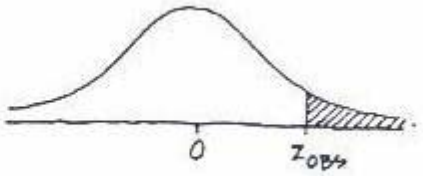
Paso n.º2. LA PRUEBA ESTADÍSTICA ES

$$z_{OBS} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}}$$

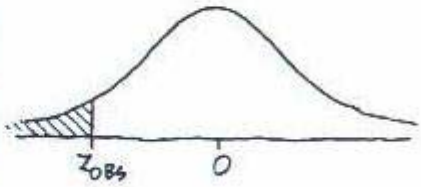
LO QUE MIDE LA GRAN DESVIACIÓN DE p CON RESPECTO A p_0. SI H_0 ES CIERTA, z_{OBS} TIENE UNA DISTRIBUCIÓN NORMAL TIPIFICADA.

Paso n.º3. EL VALOR p DEPENDE DE CUÁL SEA LA HIPÓTESIS ALTERNATIVA RELEVANTE:

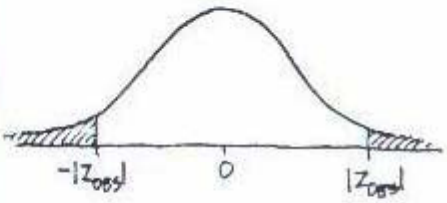
a) H_a «A LA DERECHA»: $p > p_0$
UTILIZA UN VALOR p $Pr(z > z_{OBS})$



b) H_a «A LA IZQUIERDA»: $p < p_0$
UTILIZA UN VALOR p $Pr(z < z_{OBS})$



c) H_a «BILATERAL»: p p_0 UTILIZA UN VALOR p $Pr(|z| > |z_{OBS}|)$



EN EL EJEMPLO DEL SENADOR ASTUTO:

1) LAS HIPÓTESIS SON:

$$H_0 : p = 0{,}5$$

$$H_a : p > 0{,}5$$

2) LA PRUEBA ESTADÍSTICA ES:

$$z_{obs} = \frac{0{,}55 - 0{,}50}{\sqrt{(0{,}5)(0{,}5)}/\sqrt{1.000}} = 3{,}16$$

3) EL VALOR p ES:

$$Pr(z > z_{obs}) = Pr(z \geq 3{,}16) = 0{,}0008$$

(DE LA TABLA NORMAL)

4) ASTUTO, QUE ES BASTANTE CONSERVADOR, TOMA UN NIVEL DE SIGNIFICACIÓN α DE 0,01 Y OBSERVA QUE

$$Pr(z > z_{obs}) = 0{,}0008 < \alpha$$

ASÍ, EL SENADOR RECHAZA LA HIPÓTESIS NULA. ÉL (Y SUS PARTIDARIOS) PUEDEN ESTAR SEGUROS DE SU VENTAJA ELECTORAL.

MUESTRA GRANDE
PRUEBA PARA LA MEDIA POBLACIONAL

PRIMERO, ESCOGEN SUS HIPÓTESIS.

$H_0: \mu = 16$ ONZAS

$H_a: \mu < 16$ ONZAS

RECHAZAR LA HIPÓTESIS NULA SUPONE RECHAZAR A GRANOLA

DESPUÉS, ELIGEN UN ANÁLISIS ESTADÍSTICO. AHORA, YA DEBERÍA SER UN ACTO REFLEJO SABER QUE LA DISPERSIÓN MUESTRAL DESDE LA MEDIA ES

$$\frac{\bar{X}-\mu_0}{SE(\bar{X})} = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$$

DÓNDE s ES LA DESVIACIÓN TÍPICA MUESTRAL. EN LA HIPÓTESIS NULA, ESTA SE APROXIMA A LA NORMAL ESTÁNDAR CUANDO LA MUESTRA ES GRANDE, POR EL TEOREMA CENTRAL DEL LÍMITE.

VOLVIENDO UN MOMENTO AL PASO N.º 3, ESTABLECEN UN LÍMITE PARA EL RIESGO α. COMO NO ACABARON LA CARRERA DE CIENCIAS, LOS DE ALIMENTOS DE VERDAD CREEN QUE $\alpha = 0{,}05$ SUENA BIEN.

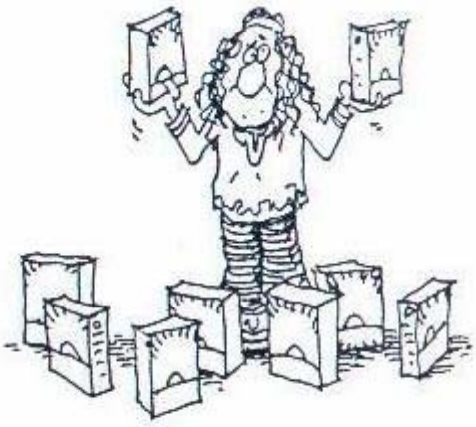
JUSTO EN ESE MOMENTO, LLEGA UN CAMIÓN CARGADO CON 10.000 CAJAS DE GRANOLA.

COGEN UNA PEQUEÑA MUESTRA ALEATORIA SIMPLE DE 49 CAJAS. LAS PESAN POR SEPARADO Y DETERMINAN EL RESUMEN ESTADÍSTICO:

$\bar{x} = 15{,}90$ ONZAS

$s = 0{,}35$ ONZAS

UN POCO LIGERO PERO, ¿SIGNIFICATIVO?

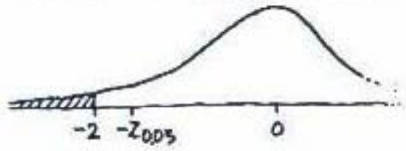


INSERTAN LOS VALORES EN LA PRUEBA ESTADÍSTICA PARA DESCUBRIR QUE:

$$z_{OBS} = \frac{15.9-16}{0.35/\sqrt{49}} = -2$$

AHORA CALCULAN EL VALOR p:

$$Pr(z < -2 \mid H_0) = 0{,}0227$$



AL SER ÉSTE MENOR QUE EL RIESGO $\alpha = 0{,}05$, ALIMENTOS DE VERDAD **RECHAZA** LA HIPÓTESIS NULA Y EL ENVÍO.

MUESTRA PEQUEÑA
PRUEBA PARA LA MEDIA POBLACIONAL

VOLVEMOS A CAMALEÓN MOTORS, Y A SU PRUEBA DE UN ACCIDENTE A UNOS 20 KILÓMETROS POR HORA. LA COMPAÑÍA DE SEGUROS LA HONRADA CUBRIRÁ AL ASEGURADO SÓLO SI EL COSTE MEDIO DE LA REPARACIÓN DE SU COCHE TRAS UN ACCIDENTE A 20 KILÓMETROS POR HORA ES INFERIOR A 1.000 DÓLARES. LA COMPAÑÍA UTILIZA EL ESTÁNDAR $\alpha = 0{,}05$. ASÍ QUE...

H_0: $\mu \geq 1.000$ DÓLARES. EL COSTE MEDIO ES DEMASIADO ALTO.

H_a: $\mu < 1.000$ DÓLARES. EL COSTE MEDIO ESTÁ BIEN.

EL ANÁLISIS ESTADÍSTICO SE BASA EN LA DISTRIBUCIÓN t

$$t = \frac{\bar{X} - \mu_0}{SE(\bar{X})}$$

EN LA QUE μ_0 ES LA MEDIA HIPOTÉTICA DE 1.000 DÓLARES.

Y QUEREMOS QUE NUESTRO VALOR t OBSERVADO PERMANEZCA A LA IZQUIERDA DE $-t_{0,05}$ (YA QUE LA $\bar{x}$ DE VALOR REDUCIDO ES PREFERIBLE. $\bar{x} - \mu_0$ DEBERÍA SER NEGATIVA PARA APOYAR A H_a).

GRADOS DE LIBERTAD	α = 0.05	0.025	0.005
1	6,31	12,71	63,66
2	2,92	4,30	9,92
3	2,35	3,18	5,84
4	2,13	2,78	4,60
5	2,01	2,57	4,03

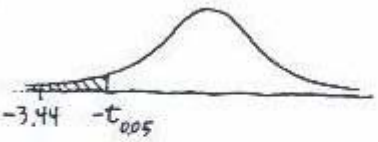
POR LA TABLA DE VALORES CRÍTICOS t, VEMOS QUE $t_{0.05} = 2{,}13$. ASÍ QUE DECIDIMOS RECHAZAR LA H_0 SI

$$t_{OBS} \leq -t_{0.05} = -2{,}13$$

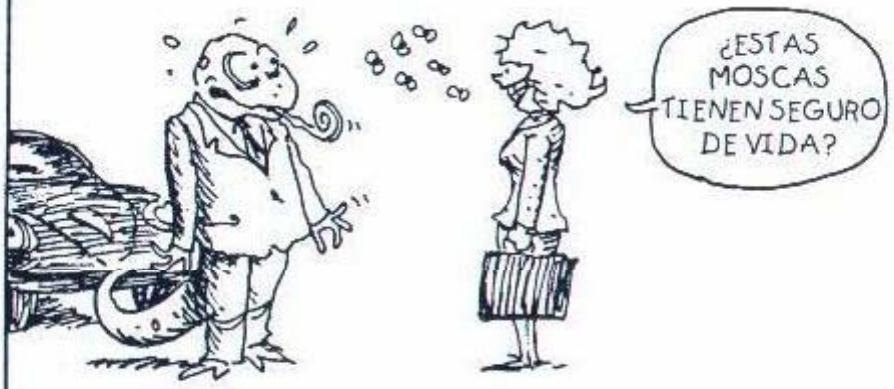
DEL CAPÍTULO 7 TENEMOS $\bar{x} = 540$ DÓLARES Y $s = 299$ DÓLARES PARA UNA MUESTRA PEQUEÑA DE CINCO COCHES, ASÍ QUE OBTENEMOS

$$t_{OBS} = \frac{540 - 1.000}{299/\sqrt{5}}$$

$$= -3{,}44 < -t_{0.05}$$



EL COCHE PASA LA PRUEBA... H_0 ES RECHAZADA... Y LA PÓLIZA DE SEGUROS ES TRAMITADA.

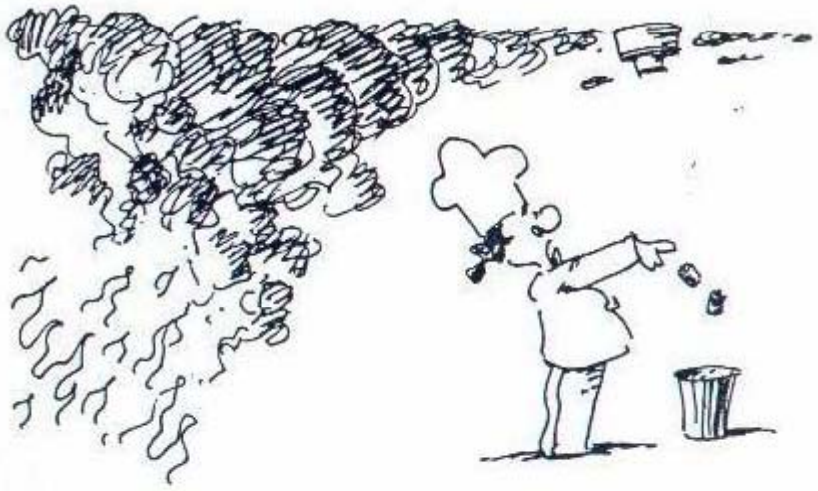


ESTE ES UN EJEMPLO DE **ACEPTACIÓN DE MUESTREO**. LA HIPÓTESIS NULA CONSISTE EN QUE LOS COSTES DE REPARACIÓN SON INADMISIBLES, Y LA CASA AUTOMOVILÍSTICA ASUME LA RESPONSABILIDAD HASTA QUE SE PRESENTEN PRUEBAS DE INOCENCIA SUFICIENTES, ES DECIR, QUE EL ACCIDENTE ENTRE DENTRO DEL PRESUPUESTO.

TEORÍA DE LA DECISIÓN

PODEMOS PENSAR EN EL CONTRASTE DE HIPÓTESIS Y EN LA PRUEBA DE SIGNIFICACIÓN EN TÉRMINOS DE **DETECTORES DE HUMO DOMÉSTICOS**. SI TIENES UNO DE ESTOS APARATOS EN CASA, TE HABRÁS DADO CUENTA DE QUE SUELEN DISPARARSE CADA VEZ QUE SE CHAMUSCAN LAS TOSTADAS.

ESTO ES LO QUE SE LLAMA UN **ERROR DE TIPO I**: UNA ALARMA SIN FUEGO. POR EL CONTRARIO, UN **ERROR DE TIPO II** ES UN FUEGO SIN ALARMA. TODOS LOS COCINEROS SABEN CÓMO EVITAR UN ERROR DE TIPO I: QUITANDO LAS PILAS. POR DESGRACIA ESTO AUMENTA LA INCIDENCIA DE ERRORES DE TIPO II.



DE IGUAL MODO, LA DISMINUCIÓN DE POSIBILIDADES DE ERRORES DE TIPO II, POR EJEMPLO, HACIENDO QUE LA ALARMA SEA HIPERSENSIBLE, PUEDE AUMENTAR EL NÚMERO DE FALSAS ALARMAS.

PODEMOS RESUMIR TODO ESTO EN UNA TABLA DE DECISIONES DE 2 × 2:

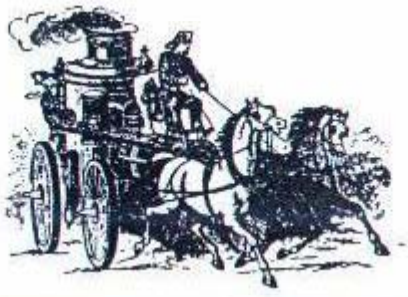
	SIN FUEGO	CON FUEGO
SIN ALARMA	NO HAY ERROR	TIPO II
CON ALARMA	TIPO I	NO HAY ERROR

AHORA PENSEMOS EN LA HIPÓTESIS NULA COMO CONDICIÓN DE «SIN FUEGO», MIENTRAS QUE LA HIPÓTESIS ALTERNATIVA ES QUE HAY UN INCENDIO. LA ALARMA CORRESPONDE AL RECHAZO DE LA HIPÓTESIS NULA.

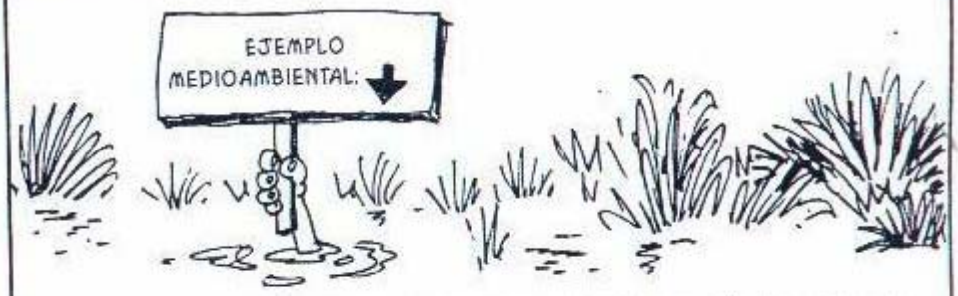
	ESTADO REAL	
	H_0	H_a
ACEPTAR H_0	NO HAY ERROR	TIPO II
RECHAZAR H_a	TIPO I	NO HAY ERROR

TODAS LAS PRUEBAS DE SIGNIFICACIÓN QUE HICIMOS ANTERIORMENTE EN ESTE CAPÍTULO SUBRAYABAN LA PROBABILIDAD DE ENFRENTARSE A UN ERROR DE TIPO I, ES DECIR, LA PROBABILIDAD DE QUE SE DIERAN NUESTRAS OBSERVACIONES SI H_0 FUERA CIERTA. PEDIMOS QUE:

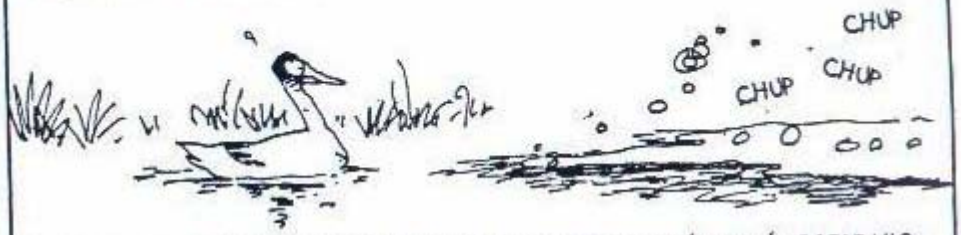
$$\Pr(\text{DE RECHAZO DE } H_0 \mid H_0) = \Pr(\text{ERROR TIPO I} \mid H_0) = \alpha$$



PERO, OTRAS VECES, LO QUE EN REALIDAD QUEREMOS CONOCER ES LA POSIBILIDAD DE QUE SE PRODUZCA UN **ERROR DE TIPO II**. ES DECIR, ¿QUÉ NIVEL DE SENSIBILIDAD TIENE NUESTRO «SISTEMA DE ALARMA» CUANDO LA HIPÓTESIS ALTERNATIVA ES **VERDADERA**?



ANTES, LAS FÁBRICAS QUE VERTÍAN SUS DESECHOS EN LAS VÍAS FLUVIALES DEBÍAN DEMOSTRAR QUE EL VERTIDO NO TENÍA EFECTOS NOCIVOS PARA LA VIDA ANIMAL. ESTO ES LA H_0. EL QUE CONTAMINABA PODÍA SEGUIR HACIÉNDOLO HASTA QUE LA HIPÓTESIS NULA FUERA RECHAZADA POR ALCANZAR EL NIVEL DE SIGNIFICACIÓN 0,05.



ASÍ QUE, CUANDO EL RESPONSABLE DE LA CONTAMINACIÓN CREÍA ESTAR VIOLANDO LOS LÍMITES ESTABLECIDOS POR LA AGENCIA PARA LA PROTECCIÓN DEL MEDIO AMBIENTE, LLEVABA A CABO UN PLAN DE SEGUIMIENTO DE CONTAMINACIÓN NADA EFECTIVO.

EL CULPABLE DE LA CONTAMINACIÓN SE SIENTE MUY SATISFECHO, PORQUE, COMO EN NUESTRA ALARMA DE HUMO SIN PILAS, SU ANÁLISIS TIENE POCAS POSIBILIDADES O NINGUNA DE HACER SALTAR LA ALARMA.

FORMALICEMOS ESTA IDEA. PARA DESCRIBIR LA PROBABILIDAD DE ERROR DE TIPO II, AÑADIMOS UNA NUEVA LETRA GRIEGA: BETA O β.

$\beta = Pr(\text{DE ACEPTACIÓN } H_0 | H_a)$

$= Pr(\text{ERROR TIPO II} | H_a)$

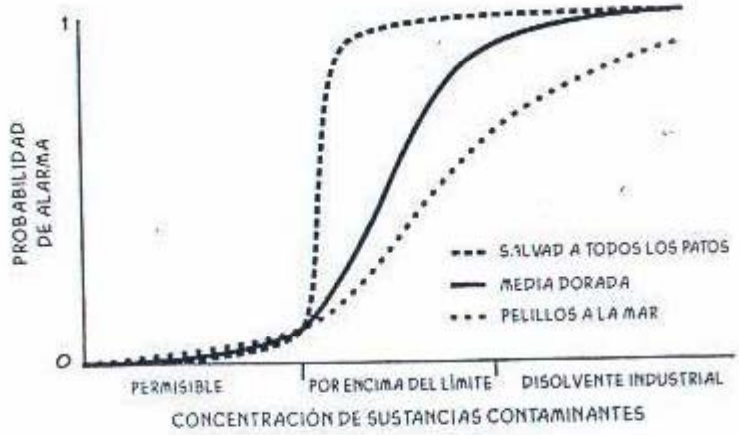
LA POTENCIA DE UNA PRUEBA SE DEFINE COMO $1 - \beta$.

$Pr(\text{DE RECHAZO DE } H_0 | H_a)$.

TE GUSTARÁ SABER QUE LOS LEGISLADORES MEDIOAMBIENTALES CADA VEZ EXIGEN MÁS PROGRAMAS DE SEGUIMIENTO PARA DEMOSTRAR QUE TIENEN UNA PROBABILIDAD MUY ALTA DE DETECTAR GRAVES CASOS DE CONTAMINACIÓN. EL ANÁLISIS DE POTENCIA REVELA A MENUDO DEFECTOS OCULTOS EN LOS PROGRAMAS DE SEGUIMIENTO.

UNA FORMA DE VISUALIZAR EL EFECTO DE LOS ANÁLISIS DE POTENCIA ES DIBUJAR LA GRÁFICA DE LA PROBABILIDAD DEL RECHAZO DE H_0 Y EL ESTADO REAL DEL SISTEMA DE ALARMA. EN EL CASO DE LA ALARMA PARA HUMOS, LA PROBABILIDAD ASCIENDE HASTA 1 A MEDIDA QUE EL HUMO SE HACE MÁS DENSO.

PARA EL EJEMPLO DE LA CALIDAD DEL AGUA DE LA AGENCIA PARA LA PROTECCIÓN DEL MEDIO AMBIENTE, EL EJE HORIZONTAL REPRESENTA LA CONCENTRACIÓN REAL DE CONTAMINANTE EN EL AGUA.



AQUÍ, ESTÁN REPRESENTADAS LAS CURVAS DE EFECTIVIDAD DE LOS TRES PROGRAMAS DE SEGUIMIENTO. LA DE **SALVAD A TODOS LOS PATOS** (CON UN COSTE DE 5 MILLONES DE DÓLARES), LA **MEDIA DORADA** (CON UN COSTE DE 500.000 DÓLARES) Y **PELILLOS A LA MAR** (TAMBIÉN, CON UN COSTE DE 500.000 DÓLARES). CUANTO MAYOR SEA LA POTENCIA DE LA PRUEBA MAYOR SERÁ LA PRONUNCIACIÓN DE LA CURVA.

ENTONCES ¿POR QUÉ TIENES ESA **SENSACIÓN DE VACÍO** EN EL ESTÓMAGO? PORQUE, PARA PONER EN PRÁCTICA ESTAS IDEAS, TENEMOS QUE APRENDER A APLICARLAS EN DIFERENTES SITUACIONES DE LAS QUE NI SIQUIERA HEMOS HABLADO HASTA AHORA. ES LO QUE VAMOS A HACER A CONTINUACIÓN, MEDIANTE LA **COMPARACIÓN** DE **DOS POBLACIONES**.

♦ Capítulo 9 ♦

COMPARACIÓN DE DOS POBLACIONES

DONDE APRENDEREMOS NUEVAS RECETAS
USANDO VIEJOS INGREDIENTES...

PERO, LO QUE CONVIERTE A LA ESTADÍSTICA EN ALGO CASI TAN DESAFIANTE COMO LA COCINA, ES LA VARIEDAD. AL IGUAL QUE UN COCINERO EXPERTO, EL ESTADÍSTICO PUEDE DEGUSTAR O «PROBAR» LOS INGREDIENTES EN UN PROBLEMA, PARA DESCUBRIR CUÁL ES LA FORMA MÁS EFECTIVA DE COMBINARLOS EN UNA RECETA ESTADÍSTICA.

(LA RAZÓN POR LA QUE TANTO LOS LIBROS DE COCINA COMO LOS DE ESTADÍSTICA SON TAN VOLUMINOSOS ES PORQUE AMBOS APORTAN SOLUCIONES EN UNA GRAN VARIEDAD DE SITUACIONES.)

EN ESTE CAPÍTULO UTILIZAREMOS NUESTROS MÉTODOS DEL PLATO COMBINADO CON ALGUNAS RECETAS NUEVAS QUE NOS AYUDARÁN A CONTESTAR LAS SIGUIENTES PREGUNTAS:

¿TOMAR ASPIRINAS CON REGULARIDAD PUEDE REDUCIR EL RIESGO DE INCIDENCIA DE INFARTO?

¿PUEDE UN PESTICIDA DETERMINADO AUMENTAR EL CRECIMIENTO DE MAÍZ POR HECTÁREA?

¿SON DIFERENTES LOS SALARIOS DE HOMBRES Y MUJERES QUE DESEMPEÑAN UN MISMO TRABAJO?

EL INGREDIENTE EN COMÚN DE TODAS ESTAS PREGUNTAS ES ESTE: PUEDEN SER CONTESTADAS MEDIANTE LA COMPARACIÓN DE DOS MUESTRAS ALEATORIAS INDEPENDIENTES, UNA DE CADA POBLACIÓN.

CON PESTICIDA

SIN PESTICIDA

Y AL FINAL DEL CAPÍTULO, CONSIDERAREMOS LAS DIFERENTES FORMAS DE COMPARAR DOS MEDIAS, LO CUAL NO IMPLICA ÚNICAMENTE TOMAR DOS MUESTRAS ALEATORIAS...

Comparando TASAS DE ÉXITO
(o de fracaso) en dos poblaciones.

EMPECEMOS CON UN EXPERIMENTO, QUE FORMÓ PARTE DE UN ESTUDIO DE LA UNIVERSIDAD DE HARVARD, CON EL QUE SE PRETENDÍA DECIDIR QUÉ GRADO DE EFECTIVIDAD TENÍA LA ASPIRINA EN LA REDUCCIÓN DE RIESGO DE INFARTO. Y COMO OCURRE EN MUCHAS PRUEBAS MÉDICAS, LAS PROBABILIDADES DE QUE ALGÚN INDIVIDUO SUFRA LA ENFERMEDAD, EN ESTE CASO UN INFARTO, EN EL TRANSCURSO DE UN AÑO, SON MUY PEQUEÑAS. PERO QUEREMOS RESPUESTAS RÁPIDAS. ¿QUÉ PODEMOS HACER?

LA SIMPLE, AUNQUE CARA, SOLUCIÓN ES EXAMINAR A UN GRAN NÚMERO DE INDIVIDUOS DURANTE UN PERÍODO REDUCIDO DE TIEMPO. EN ESTE ESTUDIO, SE FORMARON DOS GRUPOS A PARTIR DE 22.071 INDIVIDUOS (TODOS ELLOS MÉDICOS VOLUNTARIOS).

AL GRUPO 1 SE LE ADMINISTRA UN PLACEBO, UNA PASTILLA IDÉNTICA A LA ASPIRINA PERO QUE NO CONTIENE ASPIRINA.

AL GRUPO 2 SE LE ADMINISTRA UNA ASPIRINA DIARIA.

DURANTE UN PERÍODO DE APROXIMADAMENTE CINCO AÑOS*, LOS RESULTADOS REFLEJARON LAS SIGUIENTES RESPUESTAS: INFARTO O AUSENCIA DE INFARTO. EL RESULTADO: (EN LA TABLA QUE PRESENTAMOS A CONTINUACIÓN HEMOS COMBINADO INFARTOS MORTALES Y NO MORTALES.)

	INFARTO	NO INFARTO	n	INCIDENCIA DE INFARTO
PLACEBO	239	10.795	11.034	$\hat{p}_1 = \frac{239}{11.034} = 0,0217$
ASPIRINA	139	10.898	11.037	$\hat{p}_2 = \frac{139}{11.037} = 0,0126$

LA DIFERENCIA QUE SE OBSERVA EN EL NIVEL DE ÉXITO ES $\hat{p}_1 - \hat{p}_2 = 0,0091$. PARECE UNA CANTIDAD PEQUEÑA HASTA QUE NOS FIJAMOS EN EL RIESGO RELATIVO,

$$\frac{\hat{p}_1}{\hat{p}_2} = \frac{0,0217}{0,0126} = 1,72.$$

LOS INDIVIDUOS DEL GRUPO PLACEBO ERAN 1,72 VECES MÁS SUSCEPTIBLES DE SUFRIR UN INFARTO QUE LOS INDIVIDUOS DEL GRUPO DE LA ASPIRINA.

* EL EXPERIMENTO SE DETUVO PRONTO, POR SU RESULTADO POSITIVO. NO HUBIERA SIDO PRÁCTICO NI INTELIGENTE OCULTAR LOS RESULTADOS AL GRUPO DEL PLACEBO.

El modelo: LAS OBSERVACIONES HECHAS A PARTIR DE LOS GRUPOS DEL PLACEBO Y DE LA ASPIRINA SON MUESTRAS INDEPENDIENTES EXTRAÍDAS DE DOS POBLACIONES BINOMIALES. PARA LA CONSISTENCIA NOS REFERIMOS AL INFARTO COMO UN ÉXITO (!)

POBLACIÓN 1
DEL PLACEBO.
POSIBILIDAD DE ÉXITO = p_1

POBLACIÓN 2
DE LA ASPIRINA.
POSIBILIDAD DE ÉXITO = p_2

EL OBJETIVO ES LA ESTIMACIÓN DE LA DIFERENCIA REAL: $p_1 - p_2$.

PARA CADA POBLACIÓN (EN REALIDAD GRANDES MUESTRAS TOMADAS DE UNA POBLACIÓN EN GENERAL) TENEMOS LAS YA CONOCIDAS VARIABLES ALEATORIAS:

X_1 NÚMERO DE ÉXITOS DE LA POBLACIÓN 1

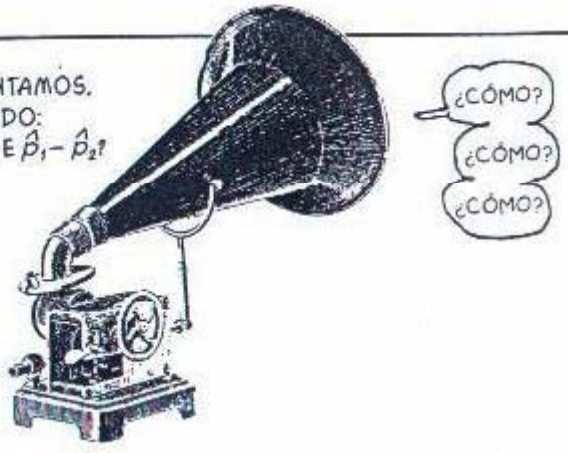
X_2 NÚMERO DE ÉXITOS DE LA POBLACIÓN 2

$\hat{P}_1 = \frac{X_1}{n_1}$ PROPORCIÓN DE ÉXITOS DE LA POBLACIÓN 1

$\hat{P}_2 = \frac{X_2}{n_2}$ PROPORCIÓN DE ÉXITOS DE LA POBLACIÓN 2

Y UN ESTIMADOR DE LA DIFERENCIA: $\hat{P}_1 - \hat{P}_2$

Y AHORA NOS PREGUNTAMOS, COMO UN DISCO RAYADO: ¿CÓMO SE DISTRIBUYE $\hat{p}_1 - \hat{p}_2$?



Distribución muestral de $\hat{P}_1 - \hat{P}_2$

PARA MUESTRAS GRANDES, $\hat{p}_1 - \hat{p}_2$ SE DISTRIBUYE CASI CON NORMALIDAD, MUCHO MÁS QUE EN EL CASO DE UNA SOLA MUESTRA. PODEMOS REALIZAR LA TÍPICA TRANSFORMACIÓN z PARA OBTENER LA VARIABLE ALEATORIA NORMAL ESTÁNDAR (APROXIMADAMENTE).

$$z = \frac{\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)}{\sigma(\hat{P}_1 - \hat{P}_2)}$$

PERO, ¿CÓMO CALCULAMOS LA DESVIACIÓN ESTÁNDAR EN EL DENOMINADOR?

COMO LAS DOS MUESTRAS SON INDEPENDIENTES, TAMBIÉN LO SON LAS VARIABLES ALEATORIAS $\hat{p}_1$ Y $\hat{p}_2$, Y LAS DOS VARIANZAS SE SUMAN.

$$\sigma^2(\hat{P}_1 - \hat{P}_2) = \sigma^2(\hat{P}_1) + \sigma^2(\hat{P}_2)$$

ENTONCES

$$\sigma(\hat{P}_1 - \hat{P}_2) = \sqrt{\sigma^2(\hat{P}_1) + \sigma^2(\hat{P}_2)}$$

Y AHORA, UNA VEZ QUE CONOCEMOS LA DISTRIBUCIÓN DE LA PRUEBA ESTADÍSTICA, PODEMOS PASAR A REALIZAR EL CÁLCULO DE LOS INTERVALOS DE CONFIANZA Y EL CONTRASTE DE LA HIPÓTESIS, QUE AFIRMA QUE LA INGESTIÓN DE ASPIRINA REDUCE EL RIESGO DE INFARTO.

Intervalos de confianza para $p_1 - p_2$

COMO SIEMPRE, LOS INTERVALOS DE CONFIANZA PARA NUESTRA ESTIMACIÓN SON ASÍ:

$$p_1 - p_2 = (\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} SE(\hat{p}_1 - \hat{p}_2)$$

- DIFERENCIA REAL ENTRE LAS PROPORCIONES POBLACIONALES
- DIFERENCIA OBSERVADA
- VALOR CRÍTICO
- ERROR TÍPICO

SE SUMAN LAS VARIANZAS $\hat{P}_1$ Y $\hat{P}_2$. AHORA EL ERROR TÍPICO ES:

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

EN EL ESTUDIO SOBRE LA ASPIRINA, EL ERROR TÍPICO ES:

$$\sqrt{\frac{(0{,}0217)(0{,}9783)}{11.034} + \frac{(0{,}0126)(0{,}9874)}{11.037}}$$

$$= 0{,}00175$$

PARA CONSEGUIR UN INTERVALO DE CONFIANZA DEL 95% EN EL ESTUDIO DE LA ASPIRINA, SÓLO TENEMOS QUE AÑADIR LOS VALORES OBSERVADOS:

$$p_1 - p_2 = 0.0091 \pm (1.96)(0.00175)$$
$$= 0{,}0091 \pm 0{,}0034$$
$$= 0{,}0057;\ 0{,}0125$$

ESTAMOS SEGUROS, COMO MÍNIMO EN UN 95%, DE QUE LA DIFERENCIA EN LA INCIDENCIA DE INFARTO SE ENCUENTRA ENTRE 0,0057 Y 0,0125. SE TRATA, DEFINITIVAMENTE, DE UNA CIFRA POSITIVA. AHORA ESTAMOS SEGUROS, COMO MÍNIMO EN UN 95%, DE QUE LA ASPIRINA DISMINUYE LA INCIDENCIA DE INFARTO.

Contraste de hipótesis

LA PREGUNTA RELATIVA AL CONTRASTE DE HIPÓTESIS ES:

H_0. LA HIPÓTESIS NULA, ES QUE LA ASPIRINA NO TIENE EFECTO: $p_1 = p_2$.

H_a. LA HIPÓTESIS ALTERNATIVA, ES QUE LA ASPIRINA REDUCE LA INCIDENCIA DE INFARTO: $p_1 > p_2$.

AHORA NECESITAMOS UN ESTADÍSTICO CON UNA DISTRIBUCIÓN NORMAL CUANDO H_0 ES CIERTA...

OBSERVA QUE BAJO H_0 LAS DOS PROPORCIONES SON IGUALES, $p_1 = p_2 = p$... AHORA PODEMOS JUNTAR TODOS LOS DATOS PARA CONSEGUIR LA PROPORCIÓN DE INFARTO EN AMBAS MUESTRAS A LA VEZ:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

CUANDO LA HIPÓTESIS NULA ES VERDADERA, EL ERROR ESTÁNDAR DEPENDE ÚNICAMENTE DE ESTA ESTIMACIÓN CONJUNTA:

$$SE_0(\hat{P}_1 - \hat{P}_2) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Y PODEMOS FORMULAR UNA PRUEBA ESTADÍSTICA

$$Z = \frac{\hat{P}_1 - \hat{P}_2}{SE_0(\hat{P}_1 - \hat{P}_2)}$$

(NORMALMENTE, EL NUMERADOR SERÍA $\hat{P}_1 - \hat{P}_2 - (p_1 - p_2)$. PERO H_0 ASUME QUE $p_1 - p_2 = 0$.)

EN EL ESTUDIO SOBRE LA ASPIRINA NOS ENCONTRAMOS CON:

$$\hat{p} = \frac{378}{22.071}$$

$$SE_0(\hat{P}_1 - \hat{P}_2) = 0{,}00175$$

ENTONCES

$$Z_{OBS} = \frac{0{,}0091}{0{,}00175} = 5{,}20$$

z_{OBS} ESTÁ A MÁS DE CINCO DESVIACIONES TÍPICAS DE CERO, UN RESULTADO ALTAMENTE SIGNIFICATIVO. HALLAREMOS EL VALOR p CON AYUDA DE UNA TABLA O UN ORDENADOR PERSONAL.

$$\text{VALOR P} = PR(Z \geq Z_{OBS}) = PR(Z \geq 5.2) = 0{,}0000001$$

SI LA HIPÓTESIS NULA FUERA VERDADERA, LA PROBABILIDAD DE OBSERVAR UN EFECTO ASÍ DE GRANDE SERÍA DE UNA ENTRE DIEZ MILLONES. ¡ES UNA PRUEBA DE MUCHO PESO CONTRA H_0!

La receta básica:

PARA PROBAR LA HIPÓTESIS NULA

$$H_0: p_1 = p_2$$

CALCULAMOS EL ESTADÍSTICO

$$z_{OBS} = \frac{\hat{p}_1 - \hat{p}_2}{SE_0(\hat{P})}$$

(EN LA QUE SE_0 SE CALCULA USANDO LA PROBABILIDAD CONJUNTA EN LA COMBINACIÓN DE LOS DOS GRUPOS).

EL VALOR p RELEVANTE DEPENDE DE LA HIPÓTESIS ALTERNATIVA

A) H_a BILATERAL: $p_1 = p_2$

VALOR P= Pr ($|Z| Z_{OBS}|$)

B) H_a A LA DERECHA: $p_1 > p_2$

VALOR P= Pr ($Z > Z_{OBS}$)

C) H_a A LA IZQUIERDA: $p_1 < p_2$

VALOR P= Pr ($Z < Z_{OBS}$)

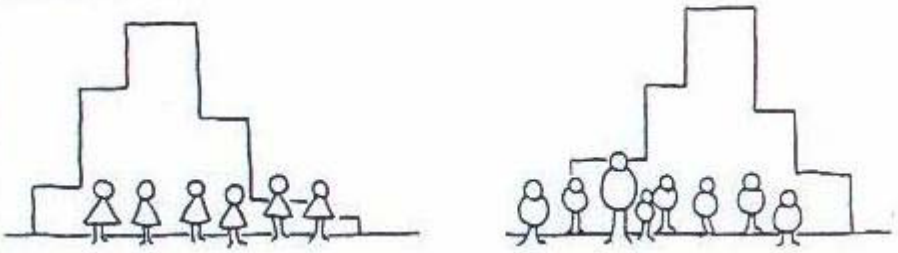
EL ANÁLISIS DEL ESTUDIO SOBRE LA ASPIRINA DEPENDE DE CIERTAS CARACTERÍSTICAS DEL EXPERIMENTO, DISEÑADAS PARA ASEGURAR LA ALEATORIEDAD Y ELIMINAR LA IMPARCIALIDAD:

LOS PUNTOS 1 Y 2 CONSTITUYEN PARTES ESENCIALES DE LA MAYORÍA DE LOS DISEÑOS DE PRUEBAS MÉDICAS CON SERES HUMANOS, PERO EL PUNTO 3 NO ES SIEMPRE NECESARIO. SE PUEDEN ENCONTRAR BUENAS PRUEBAS ESTADÍSTICAS CON MUESTRAS PEQUEÑAS EN PAQUETES DE SOFTWARE. ESTOS PROCEDIMIENTOS NO **PARAMÉTRICOS** DEPENDEN DE UNOS CÁLCULOS DE PROBABILIDAD SIMPLES PERO LARGOS, PARECIDOS A LOS CÁLCULOS DEL JUEGO QUE YA VIMOS EN EL CAPÍTULO 4...

Comparación de las **MEDIAS** de dos poblaciones

SUPONGAMOS QUE QUEREMOS COMPARAR EL SALARIO MEDIO DE LOS TRABAJADORES Y LAS TRABAJADORAS QUE DESEMPEÑAN EL MISMO TRABAJO EN UNA EMPRESA.

LA POBLACIÓN UNO ESTÁ FORMADA POR MUJERES, LA POBLACIÓN DOS, POR HOMBRES.



LA POBLACIÓN UNO TIENE UN SALARIO MEDIO μ_1 Y UNA DESVIACIÓN TÍPICA σ_1.

LA POBLACIÓN DOS TIENE UN SALARIO MEDIO μ_2 Y UNA DESVIACIÓN TÍPICA σ_2.

DOS MUESTRAS ALEATORIAS DE TAMAÑO n_1 DEL GRUPO 1 Y UNA n_2 DEL GRUPO 2 NOS DA UNAS MEDIAS MUESTRALES $\bar{x}_1$ Y $\bar{x}_2$ Y UNAS DESVIACIONES TÍPICAS σ_1 Y σ_2 RESPECTIVAMENTE. EL ESTIMADOR DE μ_1 Y μ_2 ES

$$\bar{X}_1 - \bar{X}_2$$

¿UN ESTIMADOR $\bar{X}_1 - \bar{X}_2$ ES BUENO O NO? PARA LAS MUESTRAS GRANDES ES APROXIMADAMENTE NORMAL (POR EL TEOREMA CENTRAL DEL LÍMITE) Y EL ERROR ESTÁNDAR ES

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

(LAS VARIANZAS SE SUMAN, PORQUE LAS MUESTRAS SON INDEPENDIENTES.) AHORA PODEMOS PASAR DIRECTAMENTE A LOS

intervalos de confianza: PARA MUESTRAS GRANDES, EL INTERVALO DE CONFIANZA $(1-\alpha)\,100\%$ PARA LA DIFERENCIA ENTRE MEDIAS ES

$$\mu_1 - \mu_2 = \bar{x}_1 - \bar{x}_2 \pm z_{\frac{\alpha}{2}}\, SE(\bar{X}_1 - \bar{X}_2)$$

*"RIGHT IN THE FORMULA" POR "CORRECTO" Y "EN LA DERECHA". [N.T.]

Contraste de hipótesis: ESTABLECEMOS LA HIPÓTESIS NULA DE QUE LAS MEDIAS DE LAS DOS POBLACIONES SON IGUALES:

$$H_0: \ \mu_1 = \mu_2$$

LA PRUEBA ESTADÍSTICA ES:

$$z_{OBS} = \frac{\bar{X}_1 - \bar{X}_2}{SE(\bar{X}_1 - \bar{X}_2)}$$

Y LOS VALORES P SON COMO SIEMPRE.

¿Y cómo se comparan las medias de

MUESTRAS PEQUEÑAS?

¿TE ACUERDAS DE LA CAMALEÓN MOTORS? LA COMPETENCIA, AUTO IGUANA, AFIRMA QUE SU ACCESORIO DE POLIESTIRENO COLOCADO EN LA PARTE DELANTERA DE LA CARROCERÍA, PROPORCIONA UNA MAYOR PROTECCIÓN EN CASO DE CHOQUE FRONTAL. PARA DEMOSTRARLO, HAN ESTRELLADO SIETE IGUANAS.

ESTOS SON SUS RESULTADOS COMPARADOS CON LOS DE CAMALEÓN:

CAMALEÓN	
1	*$150
2	$400
3	$720
4	$500
5	$930
n_1	5
$\bar{x}_1$	$540
s_1	$299

*(COSTE EN DÓLARES)

IGUANA	
1	$50
2	$200
3	$150
4	$400
5	$750
6	$400
7	$150
n_2	7
$\bar{x}_2$	$300
s_2	$238

LA DISTRIBUCIÓN t PUEDE APLICARSE SI DOS POBLACIONES TIENEN FORMA DE MONTAÑA Y TIENEN LA MISMA DESVIACIÓN TÍPICA $\sigma = \sigma_1 = \sigma_2$. EL ÚNICO PROBLEMILLA ES QUE TENEMOS QUE JUNTAR LAS VARIANZAS MUESTRALES S_1^2 Y S_2^2 PARA FORMAR UNA ESTIMACIÓN ÚNICA DE σ:

$$s^2_{\text{CONJUNTA}} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

EL ERROR ESTÁNDAR ES EL MISMO QUE EN LAS MUESTRAS GRANDES, SUSTITUYENDO S_{CONJUNTA} A S_1 Y S_2.

$$SE(\bar{X}_1-\bar{X}_2) = \sqrt{\frac{s^2_{\text{CONJUNTA}}}{n_1} + \frac{s^2_{\text{CONJUNTA}}}{n_2^2}}$$

$$= s_{\text{CONJUNTA}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

EL INTERVALO DE CONFIANZA $(1-\alpha)\cdot 100\%$ ES:

$$\mu_1-\mu_2 = (\bar{x}_1-\bar{x}_2) \pm t_{\frac{\alpha}{2}}\, SE(\bar{X}_1-\bar{X}_2)$$

DONDE $t_{\frac{\alpha}{2}}$ ES UN VALOR CRÍTICO DE t CON n_1+n_2-2 GRADOS DE LIBERTAD.

LOS REPTILES FABRICANTES DE COCHES CONVIENEN EN QUE SUS RESPECTIVAS DESVIACIONES TÍPICAS ESTÁN MUY PRÓXIMAS Y REPARAN EN QUE LOS HISTOGRAMAS TIENEN FORMA DE MONTAÑA. Y CALCULAN:

$$s_{\text{CONJUNTA}} = \sqrt{\frac{4\cdot 299^2 + 6\cdot 328^2}{10}} = 264$$

$$SE(\bar{X}_1-\bar{X}_2) = 264\sqrt{\tfrac{1}{5}+\tfrac{1}{7}} = 154$$

EL INTERVALO DE CONFIANZA DEL 95% ES:

$$\mu_1-\mu_2 = 540-300 \pm t_{0.025}(154)$$
$$= 240 \pm (2{,}23)(154)$$
$$= 240 \pm 340$$

PUESTO QUE ÉSTE INCLUYE EL VALOR CERO AUTO IGUANA NO HA EXPERIMENTADO UNA MEJORA SIGNIFICATIVA EN LOS GASTOS DE REPARACIÓN.

A CONTINUACIÓN, UN EJEMPLO ILUSTRATIVO DE LAS PIFIAS QUE PUEDEN COMETERSE POR LEER EL LIBRO DE RECETAS CON LOS PIES: EL PROPIETARIO DE UNA GRAN FLOTA DE TAXIS QUIERE COMPARAR LA CANTIDAD DE GASOLINA CONSUMIDA CON **GASOLINA A** Y **GASOLINA B**.

EMPIEZA CON 100 TAXIS, Y ASIGNA ALEATORIAMENTE 50 A CADA TIPO DE GASOLINA, Y, TRAS UNOS DÍAS DE CONDUCCIÓN, AFIRMA

	TAMAÑO MUESTRAL	MEDIA DE MILLAS RECORRIDAS	DESVIACIÓN TÍPICA
A	50	25	5,00
B	50	26	4,00

LA DIFERENCIA ENTRE MUESTRAS ES

$\bar{x}_1 - \bar{x}_2 = 25 - 26 = -1$

¿DE VERDAD LA GASOLINA B ES MEJOR QUE LA A?

DEBIDO AL ELEVADO VALOR DE LAS DESVIACIONES TÍPICAS, EL ERROR ESTÁNDAR ES BASTANTE SUSTANCIAL:

$$SE(\bar{X}_1-\bar{X}_2) = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

$$= \sqrt{\frac{25}{50}+\frac{16}{50}}$$

$$= 0.905$$

EN EL NIVEL DE CONFIANZA DEL 95%, TENEMOS

$$\mu_1-\mu_2 = \bar{x}_1-\bar{x}_2 \pm z_{0.025}(0.905)$$

$$= -1 \pm (1.96)(0.905)$$

$$= -1 \pm 1{,}774$$

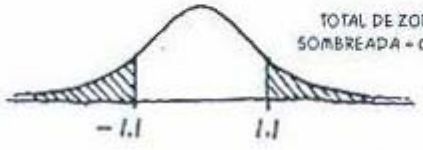
ESTO INCLUYE EL VALOR CERO, QUE CORRESPONDE A $\mu_1 = \mu_2$

EL VALOR P PARA LA HIPÓTESIS ALTERNATIVA, H_a, ES $\mu_1 \neq \mu_2$

$$Pr(|z| \geq |z_{obs}|) = Pr(|z| \geq \frac{1}{0.905})$$

$$= Pr(|z| \geq 1.1) = 2(0.1357)$$

$$= 0{,}2714$$



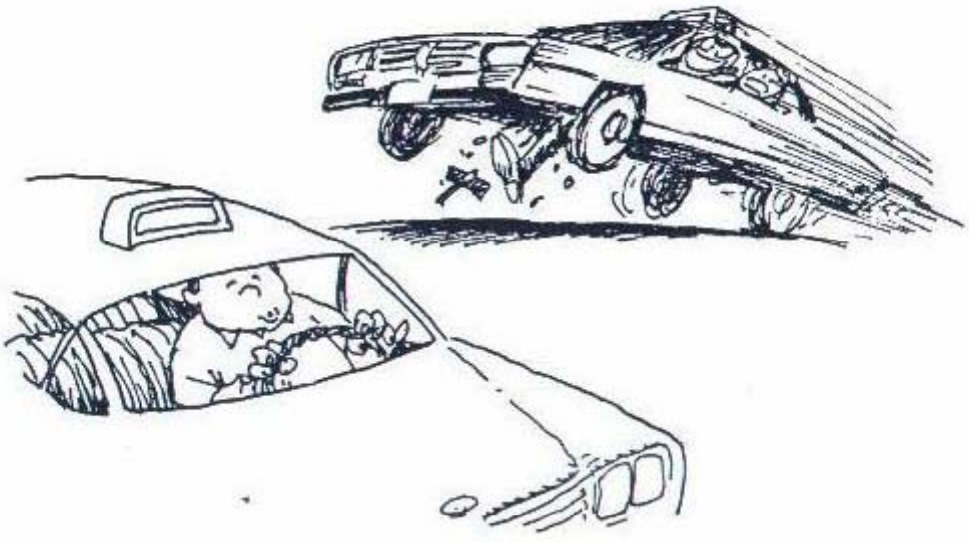
ESTA CIFRA EXCEDE EL VALOR DE SIGNIFICACIÓN $\alpha = 0{,}05$, ASÍ QUE LLEGAMOS A LA CONCLUSIÓN DE QUE LAS PRUEBAS A FAVOR DE UNA U OTRA GASOLINA SON POCO FIABLES.

COMPARACIONES APAREADAS

Una forma mejor de comparar tipos de gasolina

EL PROPIETARIO DE LOS TAXIS SIGUIÓ PASO A PASO EL LIBRO DE RECETAS. SUS MUESTRAS ERAN ALEATORIAS, Y LOS TAMAÑOS MUESTRALES ERAN LO SUFICIENTEMENTE GRANDES. SIMPLEMENTE, SE EQUIVOCÓ AL NO PENSAR CUANDO ERA NECESARIO.

AUNQUE LA GASOLINA B PARECE UN POCO MEJOR QUE LA GASOLINA A, EL INTERVALO DE CONFIANZA ES AMPLIO POR LAS GRANDES DESVIACIONES TÍPICAS. EL NÚMERO DE MILLAS RECORRIDAS VARÍA AMPLIAMENTE DE UN TAXI A OTRO. ¿POR QUÉ ESTA GRAN VARIABILIDAD? ¡PORQUE LOS TAXIS, Y LOS TAXISTAS, TIENEN PERSONALIDADES DIFERENTES!



UNA FORMA MUCHO MEJOR DE REALIZAR ESTE ESTUDIO ES PONER GASOLINA A Y GASOLINA B EN EL **MISMO** TAXI EN DÍAS DIFERENTES.

TODAVÍA TENEMOS QUE APLICAR LA ALEATORIEDAD PARA ELEGIR, LANZANDO UNA MONEDA, CUÁNDO PONEMOS GASOLINA A, EL MARTES O EL MIÉRCOLES. TAMBIÉN SE PUEDE REDUCIR LA APLICACIÓN DEL EXPERIMENTO A 10 TAXIS, Y ASÍ LE AHORRAMOS AL PROPIETARIO MUCHO TIEMPO Y DINERO.

TAXI	GASOLINA A	GASOLINA B	DIFERENCIA
1	27,01	26,95	0,06
2	20,00	20,44	−0,44
3	23,41	25,05	−1,64
4	25,22	26,32	−1,10
5	30,11	29,56	0,55
6	25,55	26,60	−1,05
7	22,23	22,93	−0,70
8	19,78	20,23	−0,45
9	33,45	33,95	−0,50
10	25,22	26,01	−0,79
MEDIA	25,20	25,80	−0,60
DESVIACIÓN TÍPICA	4,27	4,10	0,61

OBSERVA QUE LAS MEDIAS Y LAS DESVIACIONES TÍPICAS DE LA GASOLINA A Y LA GASOLINA B SON MÁS O MENOS LAS MISMAS. ESTO ERA DE ESPERAR, YA QUE POSEEN LA MISMA FUENTE DE VARIABILIDAD, COMO OCURRÍA EN EL EXPERIMENTO NO APAREADO. PERO EN ESTA OCASIÓN, LA **COLUMNA DE LA DIFERENCIA** TIENE UNA DESVIACIÓN TÍPICA **MUY PEQUEÑA**. LA COLUMNA DE LA DIFERENCIA, AL COMPARAR LA RESPUESTA DE UN SOLO COCHE CON AMBOS TIPOS DE GASOLINA, ELIMINA LA VARIABILIDAD ENTRE LOS TAXIS.

LAS DIFERENCIAS d_i PROPORCIONAN UNA MEDIDA **ÚNICA DE LA DIFERENCIA PARA CADA TAXI**, Y ASÍ, PODEMOS UTILIZARLAS PARA REALIZAR UNA PRUEBA t (DE MUESTRA PEQUEÑA)

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

EL INTERVALO DE CONFIANZA DEL 95% ALREDEDOR DE $\bar{d}$ ES

$$\mu_d = \bar{d} \pm t_{0,025}\left(s_d/\sqrt{n}\right)$$

MEDIA MUESTRAL — VALOR CRÍTICO — ERROR TÍPICO

$$= -0,6 \pm (2,26)\left(\frac{0,61}{\sqrt{10}}\right)$$

$$= \mathbf{-0,60 \pm 0,44}$$

AHORA TENEMOS $-1,04 \leq \mu_d \leq -0,16$ CON UNA SEGURIDAD DEL 95%. ES UNA BUENA PRUEBA DE QUE LA GASOLINA B ES REALMENTE MEJOR.

EL VALOR P DEL CONTRASTE DE HIPÓTESIS PUEDE SER CALCULADO CON LA AYUDA DE UN PAQUETE DE SOFTWARE:

$H_a: \mu_d \neq 0$

$$\text{VALOR P} = Pr(|t| \geq |t_{OBS}|)$$

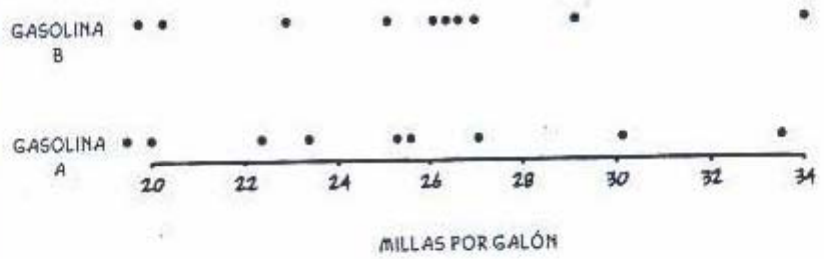
$$= Pr\left(|t| \geq \frac{0,6}{0,19}\right)$$

$$= Pr(|t| \geq 3,15)$$

$$= 0,012 < 0,05$$

DE NUEVO, LA GASOLINA B SUPERA LA PRUEBA.

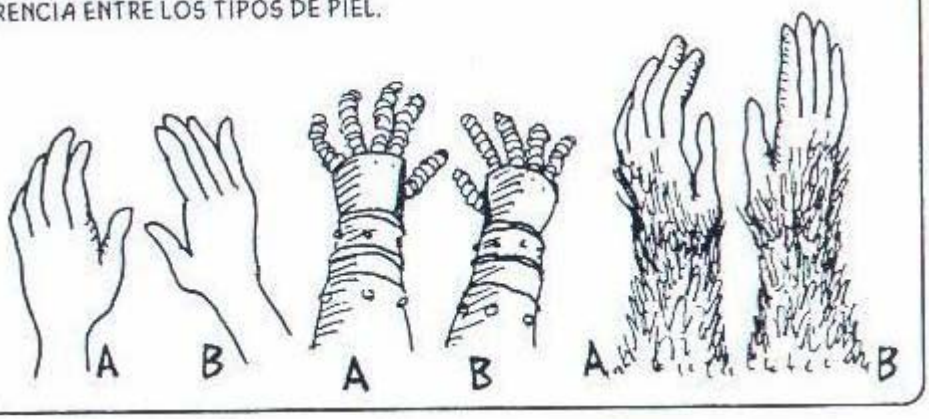
AQUÍ TENEMOS UNOS DIAGRAMAS DE PUNTOS SOBRE LOS DATOS DEL CONSUMO POR MILLAS RECORRIDAS: EL PRIMERO REPRESENTA LAS MILLAS NO APAREADAS:



Y AHORA, LOS MISMOS DATOS APAREADOS, SEGÚN TAXIS:

EL PREDOMINIO DE LÍNEAS HACIA LA DERECHA DEMUESTRA QUE LA GASOLINA B APORTA MEJORES RESULTADOS.

UN EXPERIMENTO DE COMPARACIÓN APAREADA ES UNA DE LA FORMAS MÁS EFICACES DE REDUCIR LA VARIABILIDAD NATURAL AL COMPARAR FORMAS DE TRATAMIENTO. POR EJEMPLO, SI COMPARAMOS CREMAS PARA MANOS, LAS DOS MARCAS SE ASIGNAN DE FORMA ALEATORIA A LA MANO DERECHA O IZQUIERDA DE CADA INDIVIDUO. ESTO ELIMINA LA VARIABILIDAD DEBIDA A LA DIFERENCIA ENTRE LOS TIPOS DE PIEL.



O, SI COMPARAMOS DOS MARCAS DE CEREALES, CADA CATADOR PUNTÚA AMBAS MARCAS (POR ORDEN ALEATORIO). LA COMPARACIÓN APAREADA ELIMINA EL SESGO NATURAL DE QUE AL CATADOR LE GUSTEN O NO LOS CEREALES EN GENERAL.

EN ESTE CAPÍTULO, HEMOS APLICADO LAS NOCIONES BÁSICAS SOBRE LOS INTERVALOS DE CONFIANZA Y EL CONTRASTE DE HIPÓTESIS PARA LA COMPARACIÓN DE DOS POBLACIONES. EXISTEN INNUMERABLES POSIBILIDADES. PODRÍAMOS HABER CONTINUADO DESCRIBIENDO COMPARACIONES DE:

- DESVIACIONES TÍPICAS DE DOS POBLACIONES CUANDO EL TAMAÑO MUESTRAL ES PEQUEÑO.
- LAS MEDIAS DE MÁS DE DOS POBLACIONES CUANDO EL TAMAÑO MUESTRAL ES GRANDE.
- LAS MEDIAS DE MÁS DE DOS POBLACIONES CUANDO EL TAMAÑO MUESTRAL ES PEQUEÑO.

¡ETC.!

EN LA PRÁCTICA, LOS ESTADÍSTICOS PROFESIONALES DETERMINAN LA NATURALEZA GENERAL DEL PROBLEMA Y, DESPUÉS, CONSULTAN EL LIBRO ADECUADO.

LA ÚNICA NOVEDAD DE ESTE CAPÍTULO HA SIDO LA IDEA DE **EXPERIMENTO DE COMPARACIÓN POR PAREJAS.** EN EL CAPÍTULO SIGUIENTE, VEREMOS OTROS TIPOS DE DISEÑOS EXPERIMENTALES.


¿QUIERE COMPRAR UN CAMALEÓN USADO?

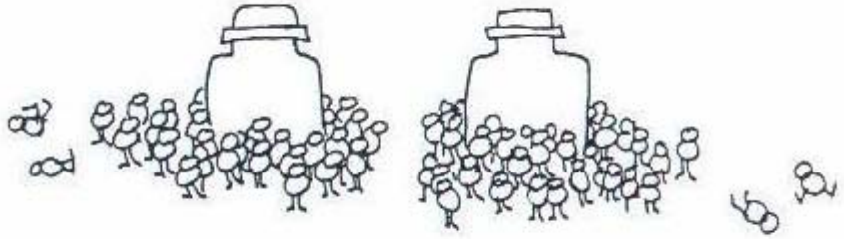
♦ Capítulo 10 ♦

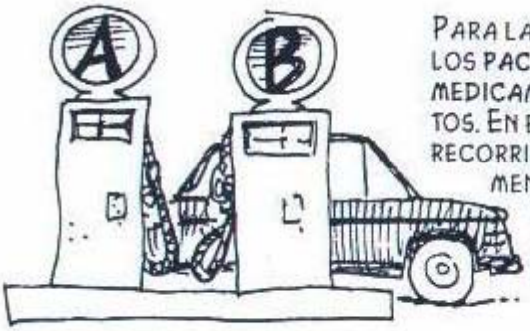
DISEÑO EXPERIMENTAL

A MENUDO, EL DISEÑO DE UN EXPERIMENTO ORIGINA SU ÉXITO O SU FRACASO. EN EL EJEMPLO DE COMPARACIÓN APAREADA, NUESTRO ESTADÍSTICO PASÓ DE ACUMULAR Y ANALIZAR DATOS DE FORMA PASIVA A PARTICIPAR ACTIVAMENTE EN EL DISEÑO EXPERIMENTAL.

EN ESTE CAPÍTULO, PRESENTAMOS LAS IDEAS FUNDAMENTALES DEL DISEÑO EXPERIMENTAL, Y DEJAREMOS LOS DETALLADOS ANÁLISIS NUMÉRICOS PARA EL ÚTIL PAQUETE DE SOFTWARE DE TU ORDENADOR.

LOS ELEMENTOS DE UN DISEÑO SON LAS **UNIDADES EXPERIMENTALES** Y LOS **TRATAMIENTOS** ASIGNADOS A LAS UNIDADES. EL OBJETIVO DE TODO DISEÑO ES LA COMPARACIÓN DE LOS TRATAMIENTOS.



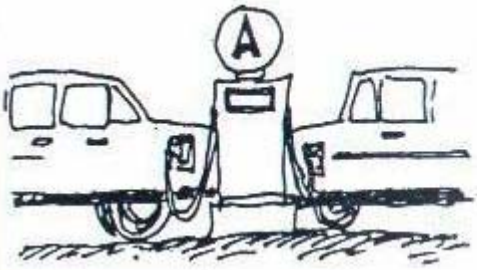


PARA LAS PRUEBAS MÉDICAS, LOS **PACIENTES** SON UNIDADES Y LOS **MEDICAMENTOS** SON LOS TRATAMIENTOS. EN EL EJEMPLO DE LAS MILLAS RECORRIDAS, LAS UNIDADES EXPERIMENTALES SON LOS TAXIS Y LOS TRATAMIENTOS QUE SE COMPARAN SON LOS TIPOS DE GASOLINA A Y B.

EN LOS EXPERIMENTOS AGRÍCOLAS, LAS UNIDADES EXPERIMENTALES SON, A MENUDO, LAS PARCELAS DE UN TERRENO, Y LOS TRATAMIENTOS PUEDEN SER LAS DIFERENTES APLICACIONES DE TIPOS DE GRANO, PESTICIDAS, FERTILIZANTES, ETC.

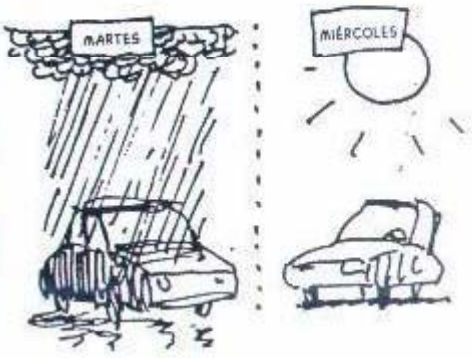
EN LA ACTUALIDAD, LAS IDEAS PARA EL DISEÑO EXPERIMENTAL SE APLICAN GENERALMENTE EN EL PROCESO DE **OPTIMIZACIÓN INDUSTRIAL, EN LA MEDICINA Y EN LAS CIENCIAS SOCIALES**. EL DISEÑO EXPERIMENTAL UTILIZA **TRES PRINCIPIOS FUNDAMENTALES**, QUE ESTÁN CLARAMENTE ILUSTRADOS EN NUESTRO EJEMPLO DE LOS TAXIS.

Repetición: SE ASIGNAN LOS MISMOS TRATAMIENTOS A LAS DIFERENTES UNIDADES EXPERIMENTALES. SIN LA REPETICIÓN, RESULTA IMPOSIBLE ESTABLECER LA VARIABILIDAD NATURAL Y EL ERROR DE LA MEDIDA.



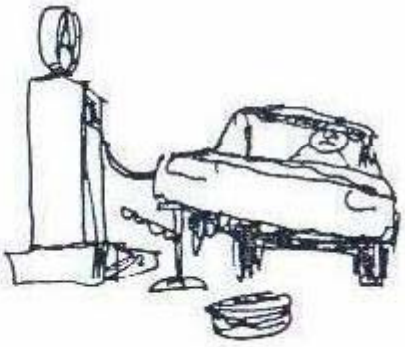
Control local: HACE REFERENCIA A CUALQUIER MÉTODO QUE REPRESENTE Y REDUZCA LA VARIABILIDAD NATURAL. UNA DE SUS FORMAS ES LA AGRUPACIÓN DE LAS UNIDADES EXPERIMENTALES EN **BLOQUES**. EN EL EJEMPLO DE LOS TAXIS, SE USARON LOS DOS TIPOS DE GASOLINA EN TODOS LOS TAXIS, Y DECIMOS QUE CADA TAXI ES UN BLOQUE.

Aleatorización: ¡ES EL PASO PRIMORDIAL DE TODAS LAS ESTADÍSTICAS! LOS TRATAMIENTOS DEBEN SER ASIGNADOS DE FORMA ALEATORIA A LAS UNIDADES EXPERIMENTALES. A CADA TAXI LE ASIGNAMOS GASOLINA A EL MARTES O EL MIÉRCOLES LANZANDO UNA MONEDA. DE NO HABERLO HECHO ASÍ, LOS RESULTADOS PODRÍAN HABERSE VISTO ARRUINADOS POR LAS DIFERENCIAS ENTRE MARTES Y MIÉRCOLES.



AHORA SUPONGAMOS QUE QUEREMOS INVESTIGAR EL EFECTO DE DOS MARCAS DE NEUMÁTICOS Y TAMBIÉN DE DOS TIPOS DE GASOLINA. TENEMOS CUATRO TRATAMIENTOS POSIBLES, QUE PODEMOS APLICAR EN UN DISEÑO FACTORIAL DOS POR DOS. LOS DOS FACTORES SON LAS MARCAS DE LOS TIPOS DE GASOLINA Y LOS NEUMÁTICOS.

	GASOLINA A	GASOLINA B
NEUMÁTICO A	*a*	*b*
NEUMÁTICO B	*c*	*d*



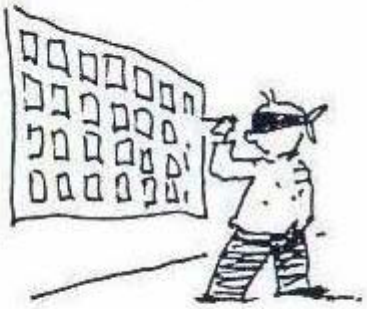
PODEMOS ASIGNAR LOS CUATRO TRATAMIENTOS A CUATRO DÍAS DIFERENTES EN CADA TAXI. LOS CUATRO TRATAMIENTOS (*A, B, C Y D*) SE REPITEN EN CADA BLOQUE (TAXI). ESTO SE DENOMINA DISEÑO COMPLETO DE BLOQUES **ALEATORIZADOS**.

HASTA AHORA, HEMOS ASUMIDO QUE TODOS LOS DÍAS DE LA SEMANA SON IGUALES, PERO TAMBIÉN PODEMOS CONTROLAR ESTE ASPECTO DE LA SIGUIENTE FORMA: USANDO SÓLO CUATRO TAXIS Y ASIGNÁNDOLES EL TRATAMIENTO COMO EN LA TABLA DE LA DERECHA:

	DÍA 1	2	3	4
TAXI 1	*a*	*b*	*c*	*d*
2	*b*	*c*	*d*	*a*
3	*c*	*d*	*a*	*b*
4	*d*	*a*	*b*	*c*

UNA TABLA CUATRO POR CUATRO, CON CUATRO ELEMENTOS DIFERENTES, EN LA QUE APARECEN TODOS LOS ELEMENTOS EN LAS CUATRO COLUMNAS Y EN LAS CUATRO FILAS, RECIBE EL NOMBRE DE **cuadrado latino.**

EN ESTE EXPERIMENTO, LOS CUATRO DÍAS Y LOS CUATRO TAXIS RECIBEN LOS CUATRO TRATAMIENTOS UNA VEZ.



EL PASO DE LA ALEATORIZACIÓN ELIGE AL AZAR UN ÚNICO CUADRADO LATINO ENTRE TODOS LOS POSIBLES.

SI LAS CUATRO UNIDADES NO SON SUFICIENTES, PODEMOS AUMENTAR EL NÚMERO DE UNIDADES EXPERIMENTALES REPITIENDO EL DISEÑO EXPERIMENTAL. SI EMPEZAMOS CON OCHO TAXIS, PODEMOS DIVIDIRLOS EN DOS GRUPOS DE CUATRO Y REPETIR EL DISEÑO EN CADA GRUPO.

HEMOS PROMETIDO NO ENTRAR EN DETALLES DEL ANÁLISIS DE DATOS, PERO ÉSTA ES, *GROSSO MODO*, LA FORMA DE TRATAR UN TIPO COMPLEJO DE DISEÑO COMO ESTE.

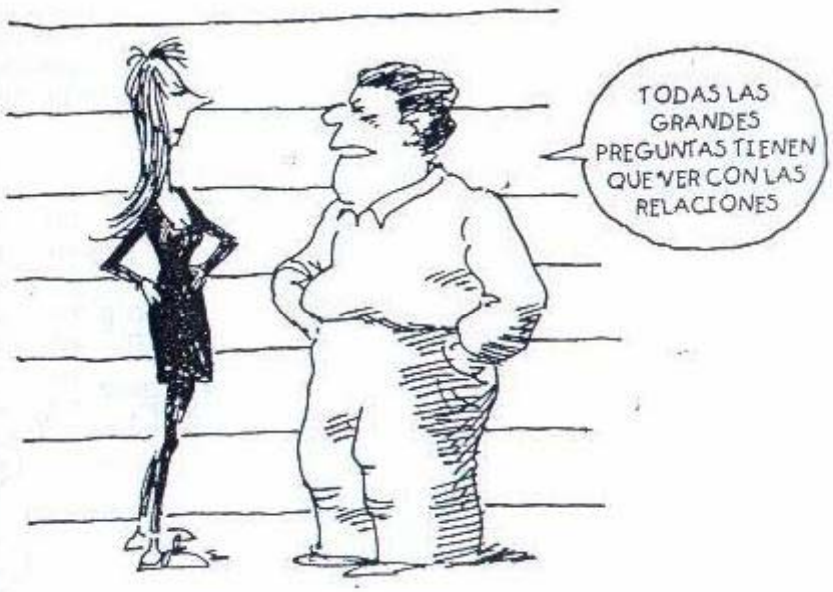
LOS DISEÑOS EXPERIMENTALES SE ANALIZAN DISTRIBUYENDO LA VARIABILIDAD TOTAL ENTRE LAS DIFERENTES FUENTES. EN EL EJEMPLO DE LOS TAXIS, LAS FUENTES DE VARIABILIDAD SON: EL TAXI, LA MARCA DEL NEUMÁTICO, EL TIPO DE GASOLINA, EL DÍA Y EL ERROR ALEATORIO. EL ANÁLISIS DE LA VARIANZA, **ANOVA*** PARA ABREVIAR, DIVIDE LA VARIACIÓN TOTAL EN PARTES Y LOCALIZA LAS PORCIONES DE CADA FUENTE.

EN EL SIGUIENTE CAPÍTULO, EXPLICAREMOS DETALLADAMENTE UN TIPO DE MODELO PARA ANALIZAR DISEÑOS COMPLEJOS: **EL MODELO DE REGRESIÓN LINEAL**. EN LA REGRESIÓN LINEAL PODRÁS VER EL **ANOVA** DE CERCA Y EXPRESADO EN FORMA NUMÉRICA...

* AUNQUE EN CASTELLANO SERÍA MÁS CORRECTO ANVA O ANDEVA, SEGUIREMOS LAS SIGLAS INGLESAS ANOVA. [N.T.]

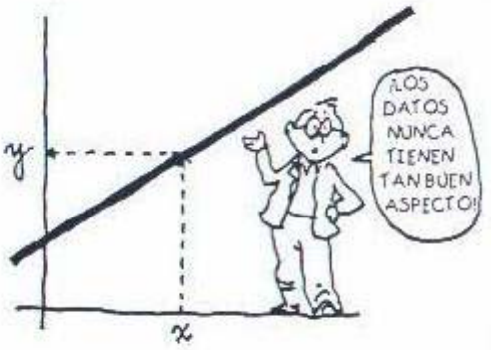
♦ Capítulo 11 ♦
REGRESIÓN

HASTA AHORA, HEMOS ESTUDIADO UNA SOLA **VARIABLE** CADA VEZ, TANTO SI SE TRATABA DE UNA POBLACIÓN DE PERSONAS A LAS QUE SE LES ADMINISTRABA UNA PÍLDORA, O DE UNA DE PEPINILLOS, COMO DE COCHES ACCIDENTADOS. EN ESTE CAPÍTULO, APRENDEREMOS A RELACIONAR **DOS VARIABLES**: DADOS LOS PESOS DE LOS 92 ESTUDIANTES DEL CAPÍTULO 2, NOS PREGUNTAREMOS QUÉ RELACIÓN TIENE EL PESO CON LA **ESTATURA** DE LOS ESTUDIANTES.

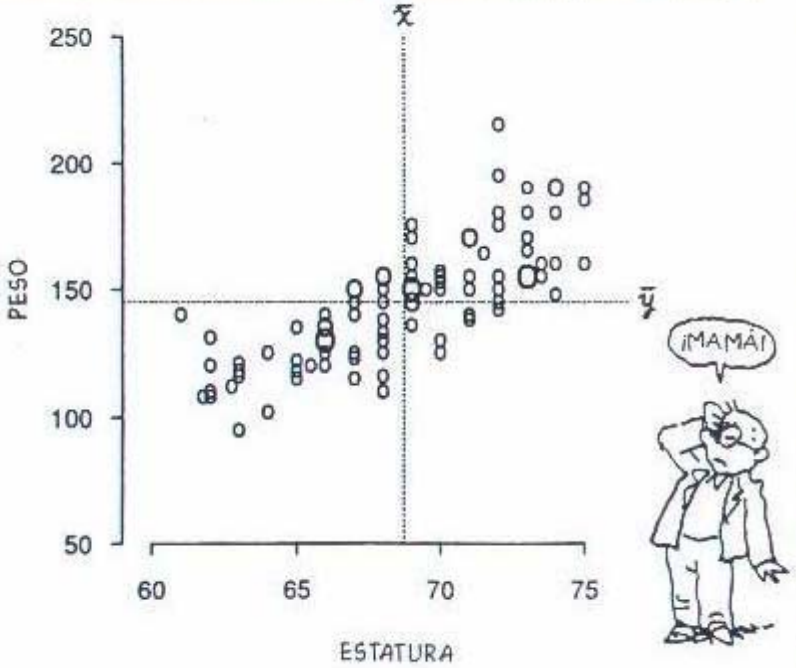


ÉSTE ES UN EJEMPLO DE UNA AMPLIA SERIE DE PREGUNTAS IMPORTANTES: ¿**PUEDE** PREDECIRSE LA **ESPERANZA** DE VIDA MIDIENDO LA **TENSIÓN ARTERIAL**? ¿LAS NOTAS DE LA **SELECTIVIDAD** PREDICEN EL **COMPORTAMIENTO ACADÉMICO** EN LA UNIVERSIDAD? ¿LEER LIBROS DE ESTADÍSTICA TE CONVIERTE EN MEJOR PERSONA?

SEGURAMENTE, EN CLASE DE MATEMÁTICAS HAS APRENDIDO A VER LAS RELACIONES REPRESENTADAS EN GRÁFICOS. DADA LA x PUEDES PREDECIR LA y. PERO, EN ESTADÍSTICA, ¡LAS COSAS NUNCA SON TAN SENCILLAS! SABEMOS (O CREEMOS SABER) QUE LA ESTATURA INFLUYE EN EL PESO, PERO NO SE TRATA DE LA ÚNICA INFLUENCIA. EXISTEN OTROS FACTORES COMO EL SEXO, LA EDAD, LA COMPLEXIÓN FÍSICA Y LA **VARIABLE ALEATORIA**.



EN ESTE CAPÍTULO ETIQUETAREMOS LOS DATOS RELATIVOS AL PESO CON LA y, Y LOS RELATIVOS A LA ESTATURA CON LA x. ASÍ (x_i, y_i) ES LA ESTATURA Y EL PESO DEL ESTUDIANTE i. REPRESENTAMOS LOS PUNTOS (x_i, y_i) EN UN DIAGRAMA BIDIMENSIONAL QUE RECIBE EL NOMBRE DE **GRÁFICO DE DISPERSIÓN DE PUNTOS**.



(ALGUNOS PUNTOS SON MÁS GRANDES PORQUE REPRESENTAN A DOS O TRES ESTUDIANTES DEL MISMO PESO Y ESTATURA.)

¿PODEMOS PREDECIR EL PESO y DE UN ESTUDIANTE A PARTIR DE SU ESTATURA x?

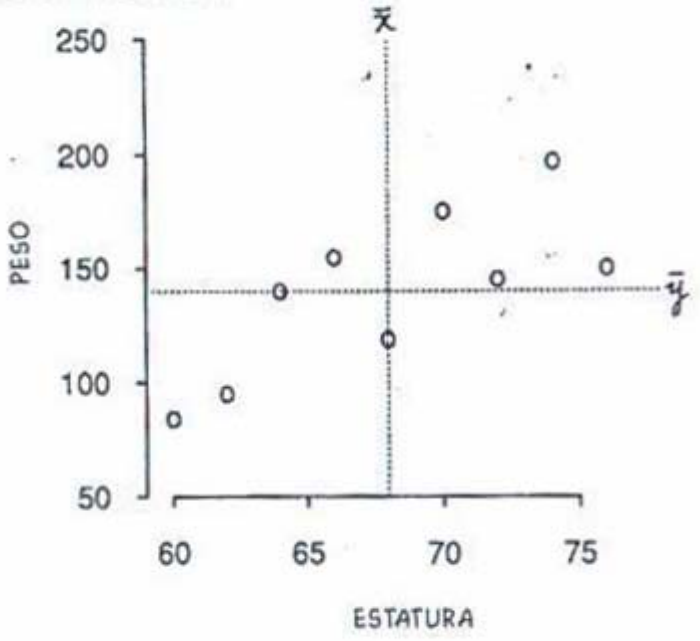
El análisis de regresión

AJUSTA UNA LÍNEA RECTA EN ESTE DESORDENADO GRÁFICO DE PUNTOS.
x RECIBE EL NOMBRE DE VARIABLE INDEPENDIENTE O REGRESORA O PREDICTORA, E y ES LA VARIABLE DEPENDIENTE O RESPUESTA. LA RECTA DE REGRESIÓN O DE PREDICCIÓN TIENE LA FORMA:

$$y = a + bx$$

PARA ILUSTRAR EL EJEMPLO DE AJUSTE DE LA RECTA, UTILIZAREMOS UN CONJUNTO MÁS REDUCIDO DE DATOS FICTICIOS CON SÓLO NUEVE PAREJAS DE PESOS Y ESTATURAS DE ESTUDIANTES.

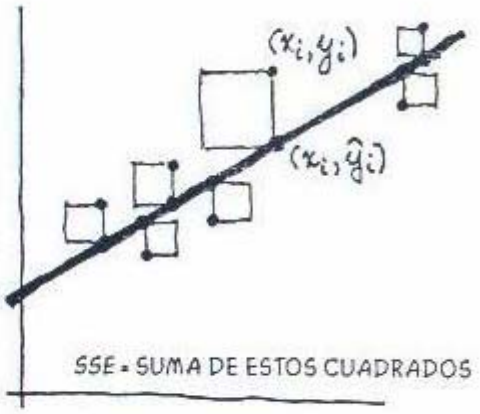
ESTATURA	PESO
60	84
62	95
64	140
66	155
68	119
70	175
72	145
74	197
76	150



ENTONCES, ¿CÓMO PODEMOS CONSEGUIR LA MEJOR RECTA DE AJUSTE?

LA IDEA CONSISTE EN **MINIMIZAR** LA DISTANCIA TOTAL DE LOS VALORES *Y* A LA RECTA. IGUAL QUE CUANDO DEFINÍAMOS LA VARIANZA, BUSCAMOS LAS **DISTANCIAS AL CUADRADO** DE *y* CON LA RECTA Y LAS SUMAMOS PARA OBTENER LA **SUMA DE LOS ERRORES CUADRÁTICOS** (*SSE*):

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



ES UNA MEDIDA AGREGADA DE CUÁNTO PUEDEN DIFERIR LAS «PREDICCIONES y_i», LLAMADAS $\hat{y}_i$, CON RESPECTO A LOS VALORES REALES y_i.

La recta de **regresión** o recta de **mínimos cuadrados**

ES LA RECTA CON LA MÍNIMA *SSE*

NOTA HISTÓRICA: ¿POR QUÉ DENOMINAMOS ESTE PROCESO ANÁLISIS DE REGRESIÓN? A PRINCIPIOS DE SIGLO, EL ESTUDIOSO DE LA GENÉTICA **FRANCIS GALTON** DESCUBRIÓ UN FENÓMENO LLAMADO **REGRESIÓN A LA MEDIA**. BUSCANDO LEYES DE HERENCIA GENÉTICA, DESCUBRIÓ QUE LA ESTATURA DE LOS HIJOS SOLÍA SER UNA REGRESIÓN A LA ESTATURA MEDIA POBLACIONAL, EN COMPARACIÓN CON LA ESTATURA DE SUS PADRES. LOS PADRES ALTOS SOLÍAN TENER HIJOS ALGO MÁS BAJOS, Y VICEVERSA. GALTON DESARROLLÓ EL ANÁLISIS DE REGRESIÓN PARA ESTUDIAR ESTE FENÓMENO, AL QUE SE REFIRIÓ DE MANERA OPTIMISTA COMO «REGRESIÓN A LA MEDIOCRIDAD».

PARA NO ANDARNOS POR LAS RAMAS, PRESENTAMOS SIN MÁS EXPLICACIONES LA FÓRMULA DE LA REGRESIÓN LINEAL: ES LIADITA PERO CALCULABLE.

$$y = a + bx$$

DONDE

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Y

$$a = \bar{y} - b\bar{x}$$

AQUÍ $\bar{x}$ E $\bar{y}$ SON LAS MEDIAS DE $\{x_i\}$ Y $\{y_i\}$ RESPECTIVAMENTE.

COMO ESTAS EXPRESIONES VOLVERÁN A SALIR, LAS ABREVIAREMOS:

$$ss_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$ss_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

LA SUMA DE LOS CUADRADOS ALREDEDOR DE LA MEDIA MIDE LA DISPERSIÓN DE x_i Y DE y_i.

$$ss_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

EL PRODUCTO CRUZADO DETERMINA (CON ss_{xx}) EL COEFICIENTE b.

ESTE ES EL CÁLCULO TOTAL DE LOS VALORES FICTICIOS:

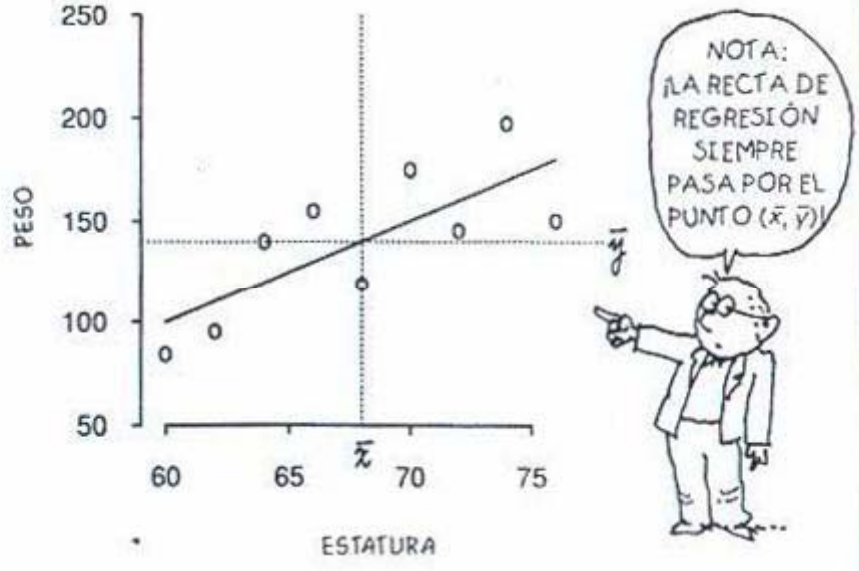
x_i	y_i	$(x_i-\bar{x})$	$(y_i-\bar{y})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
60	84	-8	-56	64	3136	448
62	95	-6	-45	36	2025	270
64	140	-4	0	16	0	0
66	155	-2	15	4	225	-30
68	119	0	-21	0	441	0
70	175	2	35	4	1225	70
72	145	4	5	16	25	20
74	197	6	57	36	3249	342
76	150	8	10	64	100	80
SUMA = 612	1.260			$SS_{xx} = 240$	$SS_{yy} = 10.426$	$SS_{xy} = 1.200$

$\bar{x} = 68$ $\bar{y} = 140$

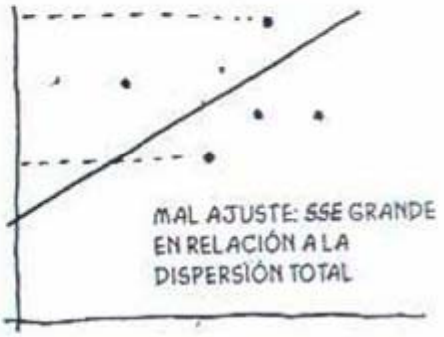
LO CUAL NOS DA VALORES PARA a Y b:

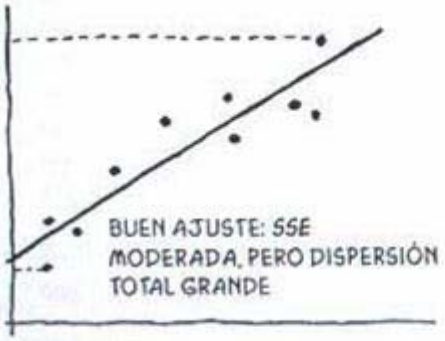
$$b = \frac{1.200}{240} = 5 \qquad a = \bar{y} - b\bar{x} = 140 - 5(68) = -200$$

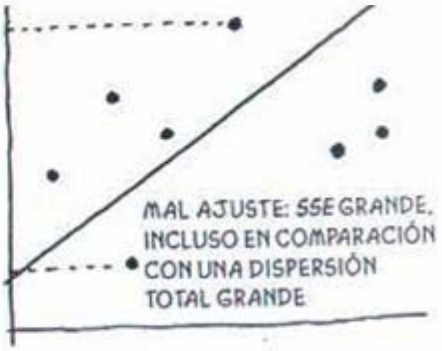
ENTONCES $y = -200 + 5x$



COMO IMAGINAS, LA RESPUESTA A ESTA PREGUNTA DEPENDE DE LA FORMA EN QUE SE ESPARCEN LOS PUNTOS DE LOS DATOS. ES DECIR, ES LA MAGNITUD DE LA SSE RELATIVA A LA DISPERSIÓN TOTAL DE LOS DATOS. ALGUNOS EJEMPLOS:







VAMOS A CUANTIFICAR ESTO DESGLOSANDO LA VARIABILIDAD DE y. SEGUIREMOS COMO GUÍA EL DIBUJO DE LA DERECHA. TENEMOS

$$\hat{y}_i = a + bx_i$$

ENTONCES, $\hat{y}_i$ SON LOS PESOS PREDICHOS POR LA RECTA DE REGRESIÓN.

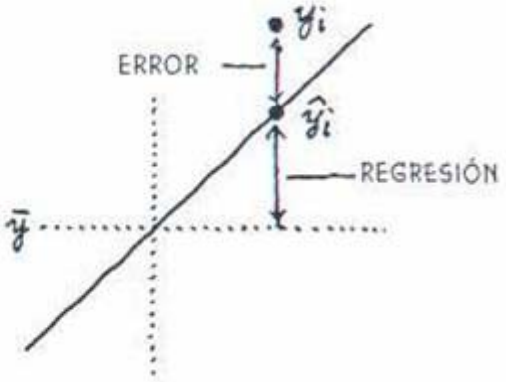


Tabla ANOVA

FUENTE DE VARIABILIDAD	SUMA DE CUADRADOS	VALOR DE LOS DATOS FICTICIOS
REGRESIÓN	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	6.000
ERROR	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	4.426
TOTAL	$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$	10.426

(POR CIERTO, AUNQUE NO ES EVIDENTE QUE $SS_{yy} = SSR + SSE$, ES VERDAD) BUENO, DE TODOS MODOS, ASÍ ES COMO SE CALCULAN LAS SUMAS DE LA REGRESIÓN Y LOS ERRORES DE LOS CUADRADOS PARA EL CONJUNTO DE LOS DATOS REALES, CON $y = -200 + 5x$

			REGRESIÓN		ERROR	
x_i	y_i	$\hat{y}_i$	$(\hat{y}_i - \bar{y})$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
60	84	100	-40	1.600	-16	256
62	95	110	-30	900	-15	225
64	140	120	-20	400	20	400
66	155	130	-10	100	25	625
68	119	140	0	0	-21	441
70	175	150	10	100	25	625
72	145	160	20	400	-15	225
74	197	170	30	900	27	729
76	150	180	40	1.600	-30	900
$\bar{x}=68$	$\bar{y}=140$			$SSR = 6.000$		$SSE = 4.426$

SSR MIDE LA VARIABILIDAD TOTAL DEBIDA A LA REGRESIÓN, O SEA, EXPLICADA POR LOS VALORES PREDICHOS DE *y*. YA NOS HEMOS ENCONTRADO CON *SSE*. OBSERVA QUE:

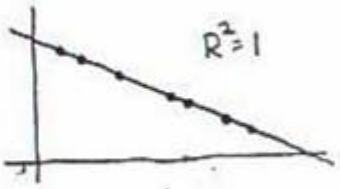
$$\frac{SSE}{SS_{yy}}$$

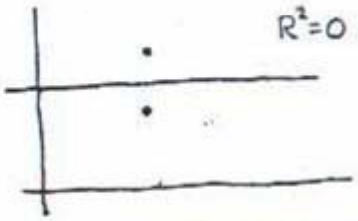
ES LA PROPORCIÓN DEL ERROR, RELATIVO A LA DISPERSIÓN TOTAL.

El coeficiente de determinación

ES LA PROPORCIÓN DE TODAS LAS SS_{yy} EXPLICABLES POR LA REGRESIÓN:

$$R^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$





(PORQUE $SSR = SS_{yy} - SSE$), R^2 ES SIEMPRE MENOR QUE 1. CUANTO MÁS SE APROXIMA A 1, MÁS PRECISO ES EL AJUSTE DE LA CURVA. $R^2 = 1$ CORRESPONDE AL AJUSTE PERFECTO.

EL CÁLCULO DE R^2 DEL CONJUNTO DE DATOS FICTICIOS ES

$$R^2 = \frac{6.000}{10.426} = 0{,}58$$

LA VARIACIÓN DEL 58% EN EL PESO SE EXPLICA POR LA ESTATURA. EL 42% RESTANTE ES EL «ERROR».

POR OTRA PARTE, TENEMOS EL

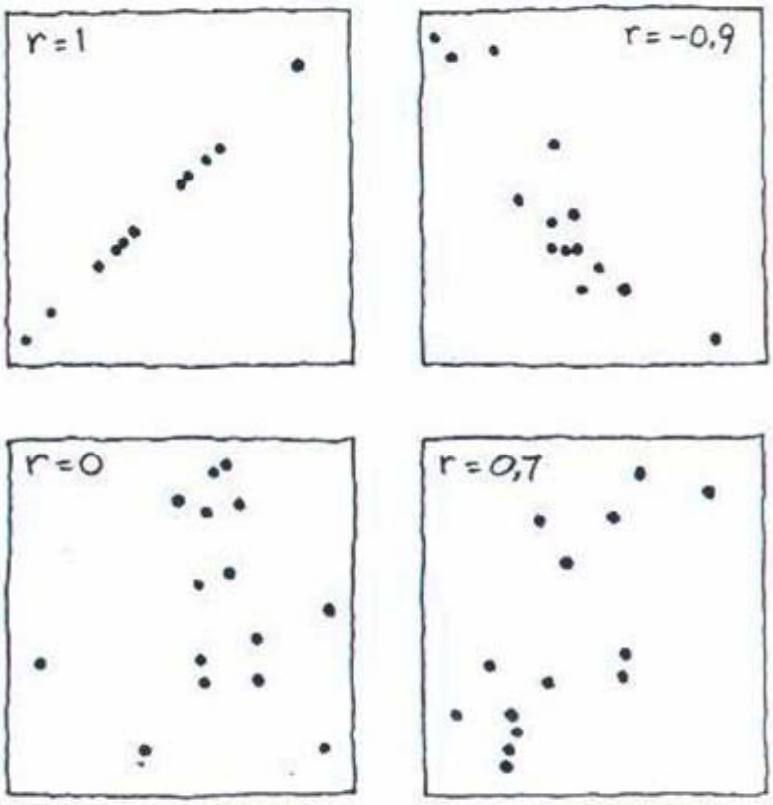
coeficiente de correlación

QUE ES LA RAÍZ CUADRADA DE R^2 CON EL SIGNO DE b.

$$r = (\text{SIGNO DE } b)\sqrt{R^2}$$

ENTONCES, r ES POSITIVA SI LA RECTA ES ASCENDENTE HACIA LA DERECHA, Y NEGATIVA SI LA RECTA TIENE FORMA DESCENDENTE HACIA LA DERECHA.

r MIDE LA PRECISIÓN DEL AJUSTE E INDICA SI AUMENTA LA x HACE SUBIR O HACE BAJAR LA y.



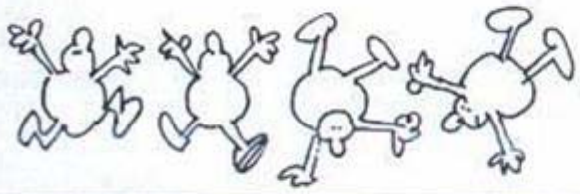
PERO SEAMOS SINCEROS: NADIE (BUENO, CASI NADIE) HACE YA ESTOS CÁLCULOS A MANO. CON EL ORDENADOR TODO ESTE TRABAJO PUEDE REALIZARSE ESCRIBIENDO UNA SOLA LÍNEA DE CÓDIGO...

EN EL SISTEMA DE SOFTWARE MINITAB, DISEÑADO EN EL ESTADO DE PENNSYLVANIA, EL ÚNICO COMANDO NECESARIO TIENE ESTE ASPECTO:

```
MTB > regress «PESO» on 1 independent variable «ESTATURA»
```

Y LOS RESULTADOS SON


```
The regression equation is

  PESO  - 200 + 5.00 ESTATURA

Predictor   Coef     Stdev   t-ratio     p
Constant   -200.0    110.7    -1.81   0.114
height      5.000    1.623     3.08   0.018

s = 25.15     R-sq = 57.5% R-sq(adj) = 51.5%


Analysis of Variance

SOURCE       DF      SS        MS        F       p
Regression    1    6000.0    6000.0    9.49   0.018
Error         7    4426.0     632.3
Total         8   10426.0
```



AHORA VAMOS A HACERLO CON LOS DATOS DE LOS 92 ESTUDIANTES:

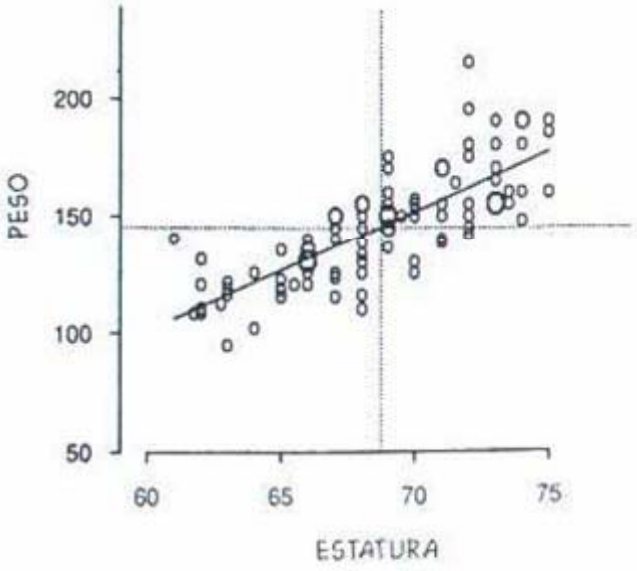
```
MTB > regress «PESO» on 1 independent variable «ESTATURA»
```

Y LOS RESULTADOS SON

```
The regression equation is
WEIGHT = - 205 + 5.09 HEIGHT

Predictor     Coef      Stdev    t-ratio      p
Constant   -204.74      29.16      -7.02  0.000
height       5.0918     0.4237     12.02  0.000

s = 14.79     R-sq = 61.6%  R-sq(adj) = 61.2%


Analysis of Variance

SOURCE       DF      SS        MS        F       p
Regression    1      31592     31592   144.38  0.000
Error        90      19692       219
Total        91      51284
```

ESTE ES EL DIAGRAMA DE DISPERSIÓN DE PUNTOS CON LA RECTA DE REGRESIÓN AJUSTADA. EL COEFICIENTE DE CORRELACIÓN PARA ESTE CONJUNTO DE DATOS ES:

$$r = +\sqrt{0,616} = 0,78$$

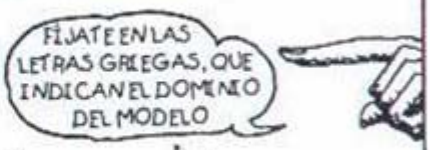


INFERENCIA ESTADÍSTICA

HASTA AHORA, HEMOS HECHO ANÁLISIS DE DATOS Y DESCRITO LA RELACIÓN LINEAL MÁS PRÓXIMA ENTRE LOS DATOS OBSERVADOS x E y. VAMOS A CAMBIAR NUESTRO PUNTO DE VISTA. RECORDEMOS A LOS 92 ESTUDIANTES COMO UNA MUESTRA POBLACIONAL DE TODOS LOS ESTUDIANTES. ¿QUÉ PODEMOS INFERIR?

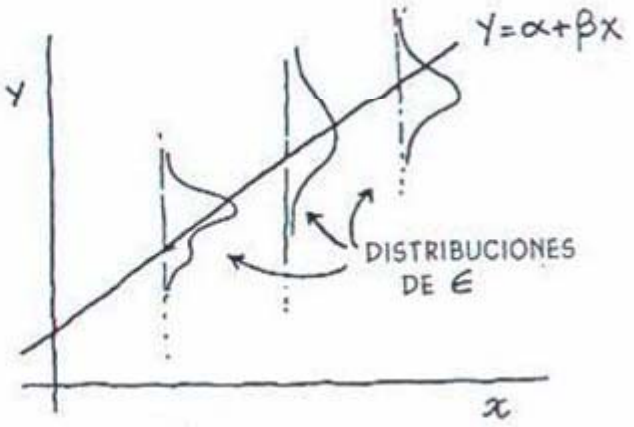
UN MODELO DE REGRESIÓN DEL TOTAL DE LA POBLACIÓN ES UNA RELACIÓN LINEAL

$$Y = \alpha + \beta x + \epsilon$$

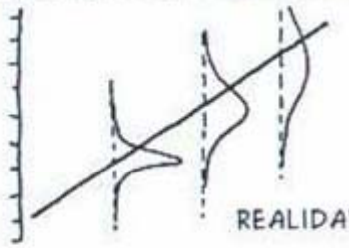


y ES LA VARIABLE ALEATORIA DEPENDIENTE; x ES LA VARIABLE INDEPENDIENTE (QUE PUEDE SER ALEATORIA O NO); α Y β SON LOS PARÁMETROS QUE QUEREMOS ESTIMAR; Y ϵ REPRESENTA LAS FLUCTUACIONES DEL ERROR ALEATORIO.

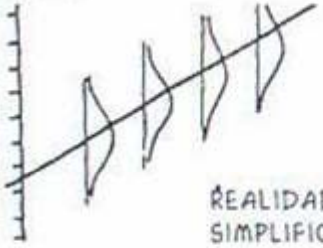
EN EL MODELO DE LA ESTATURA FRENTE AL PESO, x ES LA ESTATURA, α Y β SON LOS PARÁMETROS A ESTIMAR, Y PODEMOS CONSIDERAR ϵ COMO EL COMPONENTE ALEATORIO DE LOS PESOS y PARA CADA VALOR DE ESTATURA x.



DE HECHO, LA DISTRIBUCIÓN DE ϵ ES DIFERENTE PARA DISTINTOS VALORES DE x: LOS INDIVIDUOS QUE MIDEN 5 PIES (ALREDEDOR DE 1,52 METROS) VARÍAN MENOS EN EL PESO QUE LOS QUE MIDEN 6 PIES (ALREDEDOR DE 1,82 METROS). SIN EMBARGO, PODEMOS SIMPLIFICAR ESTA AFIRMACIÓN: SUPONGAMOS QUE PARA TODOS LOS VALORES DE x, LAS ϵ SON INDEPENDIENTES, NORMALES Y TIENEN LA **MISMA DESVIACIÓN TÍPICA** $\sigma = \sigma(\epsilon)$ Y MEDIA $\mu = 0$.



REALIDAD



REALIDAD SIMPLIFICADA

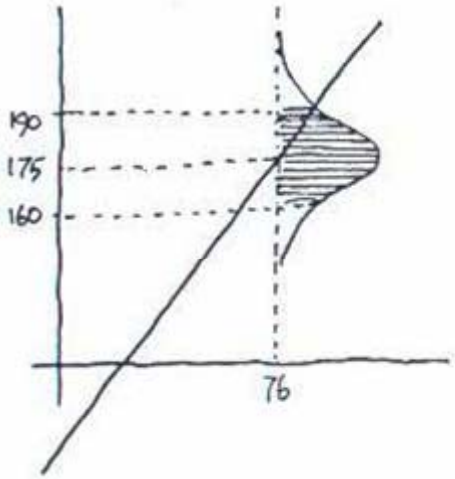
ASÍ QUE EL MODELO DE PESOS PUEDE SER

$$Y = -125 + 4x + \epsilon$$

ϵ ES NORMAL CON $\mu = 0$ Y $\sigma = 15$ LIBRAS (SUPONGAMOS). ENTONCES, DE ACUERDO CON ESTE MODELO, LOS ESTUDIANTES QUE TIENEN UNA ALTURA DE 6 PIES Y 4 PULGADAS (76 PULGADAS, O UNOS 193,4 CENTÍMETROS) TIENEN UNA DISTRIBUCIÓN DE

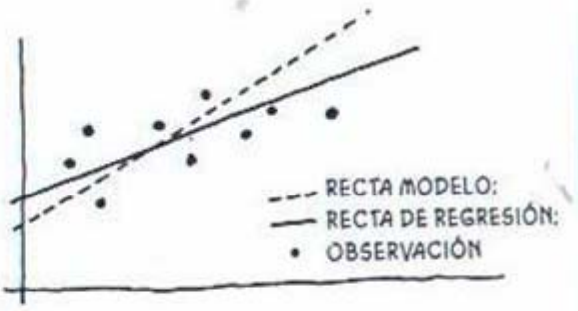
$$Y = -125 + 4(76) + \epsilon$$
$$= 175 + \epsilon$$

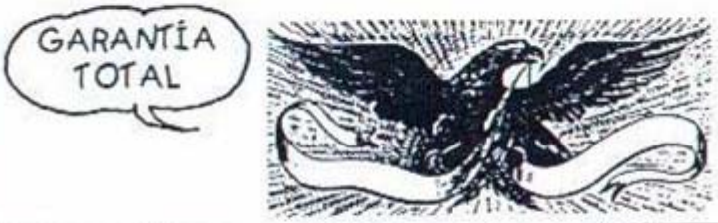
ASÍ QUE, PARA $x = 76$, Y ES NORMAL CON MEDIA 175 Y DESVIACIÓN TÍPICA DE 15 LIBRAS. *



AHORA, DADO EL MODELO $Y = \alpha + \beta x + \epsilon$, QUEREMOS HACER LO MISMO QUE HEMOS HECHO EN ESTOS ÚLTIMOS CAPÍTULOS: TOMAR UNA **MUESTRA** Y UTILIZARLA PARA **ESTIMAR** α Y β.

SE PUEDE DEMOSTRAR QUE LAS a Y b OBTENIDAS POR EL MÉTODO ANTERIOR DE MÍNIMOS CUADRADOS SON LOS **ESTIMADORES LINEALES NO SESGADOS DE MENOR VARIANZA** (SEA ESTO LO QUE SEA).





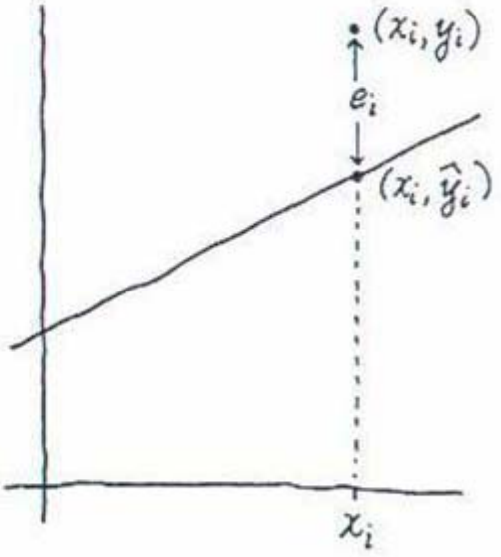
COMO SIEMPRE, **MUESTRAS** DIFERENTES PROPORCIONAN CONJUNTOS DE DATOS DIFERENTES, LO CUAL GENERA RECTAS DE REGRESIÓN DIFERENTES. ESTAS RECTAS SE **DISTRIBUYEN** ALREDEDOR DE $Y = \alpha + \beta x + \epsilon$. ENTONCES LA PREGUNTA ES: ¿CÓMO SE DISTRIBUYEN a Y b ALREDEDOR DE α Y β, RESPECTIVAMENTE, Y CÓMO CONSTRUIMOS LOS **INTERVALOS DE CONFIANZA** Y EL **CONTRASTE DE HIPÓTESIS**?

PARA CADA PUNTO (x_i, y_i) TENEMOS

$$y_i = a + bx_i + e_i$$

DONDE $e_i = y_i - \hat{y}_i$ ES LA DISTANCIA DE y_i HASTA LA RECTA DE REGRESIÓN. LOS e_i SON LOS **VALORES MUESTRALES DE** ϵ, Y NOS PROPORCIONAN UN ESTIMADOR s DE $\sigma(\epsilon)$:

$$s = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$



(¿POR QUÉ $n-2$ ES EL DENOMINADOR? POR QUÉ HEMOS UTILIZADO HASTA DOS GRADOS DE LIBERTAD PARA CALCULAR a Y b, DEJANDO $n-2$ PIEZAS INDEPENDIENTES DE INFORMACIÓN PARA ESTIMAR σ.)

AUNQUE NO RESULTE OBVIO, TAMBIÉN PODEMOS EXPRESAR s COMO:

$$s = \sqrt{\frac{SS_{yy} - bSS_{xy}}{n-2}}$$

UNA FÓRMULA QUE NOS PERMITE CALCULAR s DIRECTAMENTE A PARTIR DE LA ESTADÍSTICA MUESTRAL.

REPETIMOS, s ES UN ESTIMADOR DEL GRADO DE DISPERSIÓN QUE TENDRÁN LOS PUNTOS ALREDEDOR DE LA RECTA.

Intervalos de confianza

LOS INTERVALOS DE CONFIANZA DEL 95% PARA α Y β TIENEN ESTA YA CONOCIDA FORMA:

$$\beta = b \pm t_{0,025}\, SE(b)$$

$$\alpha = a \pm t_{0,025}\, SE(a)$$

DONDE USAMOS LA DISTRIBUCIÓN t CON $n-2$ GRADOS DE LIBERTAD (POR LA MISMA RAZÓN QUE ANTES)

SIN EMBARGO, LOS ERRORES ESTÁNDAR NO NOS SUENAN PARA NADA. SON (SIN LA DERIVACIÓN):

$$SE(b) = \frac{s}{\sqrt{SS_{xx}}}$$

$$SE(a) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}}$$

¿QUÉ HA PASADO CON NUESTRO MARAVILLOSO $\frac{1}{\sqrt{n}}$? HA SIDO SUSTITUIDO POR SS_{xx}. AL IGUAL QUE n, SS_{xx} AUMENTA A MEDIDA QUE AÑADIMOS MÁS PUNTOS. PERO TAMBIÉN REFLEJA LA DISPERSIÓN TOTAL DE LOS DATOS x. POR EJEMPLO, SI TODOS LOS ESTUDIANTES MUESTREADOS TUVIERAN LA MISMA ESTATURA, NO TENDRÍAMOS NINGUNA JUSTIFICACIÓN PARA REPRESENTAR UNA CONCLUSIÓN SOBRE LA DEPENDENCIA DEL PESO CON RESPECTO A LA ESTATURA. SI ASÍ FUERA, $SS_{xx} = 0$, Y OBTENDRÍAMOS $b = \infty$ Y UNOS INTERVALOS DE CONFIANZA CON AMPLITUD INFINITA.

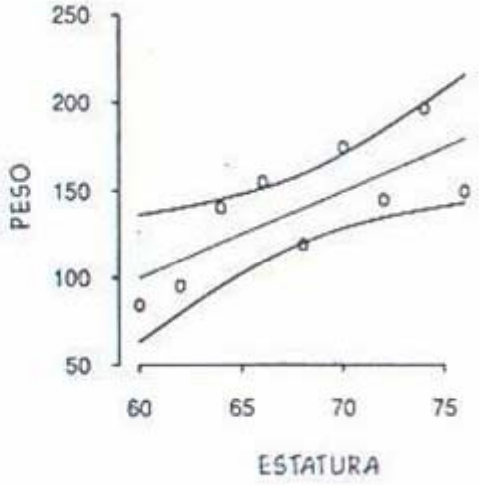
MÁS PREGUNTAS:

¿CON QUÉ PRECISIÓN PODEMOS INFERIR LA RESPUESTA MEDIA Y EN UN VALOR FIJO x_0? POR EJEMPLO, ¿CUÁL ES EL PESO MEDIO DE LOS ESTUDIANTES QUE MIDEN 76 PULGADAS? EL INTERVALO DE CONFIANZA DEL 95% PARA $Y = \alpha + \beta x_0$ ES:

$$\alpha+\beta x_0 = a + bx_0 \pm t_{0,025}\,SE(\hat{y})$$

DONDE

$$SE(\hat{y}) = s\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SS_{xx}}}$$



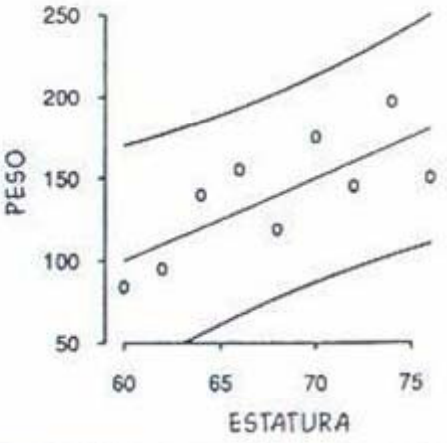
SUPONGAMOS QUE ENTRA UN NUEVO ESTUDIANTE QUE TIENE UNA ALTURA x_{nuevo}. ¿CON QUÉ PRECISIÓN PODEMOS INFERIR y_{nuevo} SIN PESARLE?

EL INTERVALO DE CONFIANZA DEL 95% DE y_{nuevo} PARA UN INDIVIDUO CON UNA x_{nueva} OBSERVADA ES

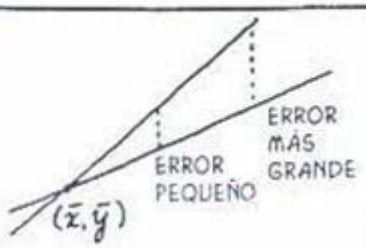
$$Y_{nuevo} = a + bx_{nuevo} \pm t_{0,025}\,SE(Y_{nuevo})$$

DONDE

$$SE(Y_{nuevo}) = s\sqrt{1 + \frac{1}{n} + \frac{(x_{nuevo}-\bar{x})^2}{SS_{xx}}}$$



AMBOS ERRORES ESTÁNDAR CONTIENEN UN TÉRMINO QUE CRECE A MEDIDA QUE EL VALOR x_0 O x_{nueva} SE ALEJA DEL VALOR MEDIO $\bar{x}$. ¿POR QUÉ EL ERROR SE ALEJA MÁS DE $\bar{x}$? PORQUE, SI DESPLAZAMOS LA RECTA DE REGRESIÓN, ¡SE QUEDA MUY ALEJADO DE LA MEDIA! (RECUERDA, LA RECTA SIEMPRE PASA POR $(\bar{x}, \bar{y})$.)



HAGAMOS LO MISMO CON LOS DATOS FICTICIOS: PARA EL PESO MEDIO CUANDO $x = 76$ PULGADAS, TENEMOS QUE $\beta = -200$ Y $\alpha = 5$. ENTONCES

$$y = -200 + 5(76) \pm (2{,}365)(25{,}15)$$

$$= 180 \pm (2{,}365)(25{,}15)\sqrt{0{,}3777}$$

$$= 180 \pm 36{,}34 \text{ LIBRAS} = [144, 216]$$

LA MEDIA ESTIMADA DE LOS ESTUDIANTES QUE MIDEN 6 PIES Y 4 PULGADAS ES DE 180 LIBRAS, Y TENEMOS UNA SEGURIDAD DEL 95% DE QUE ESTAMOS A MENOS DE 36 LIBRAS DE LA MEDIA REAL.

PARA UN NUEVO ESTUDIANTE QUE MIDA 76 PULGADAS, UTILIZAMOS NUESTRA MUESTRA FICTICIA DE NUEVE PUNTOS PARA INFERIR QUE

$$y_{\text{nuevo}} = -200 + 5(76) \pm (2{,}365)(25{,}15)\sqrt{1 + \frac{1}{9} + \frac{(76-68)^2}{290}}$$

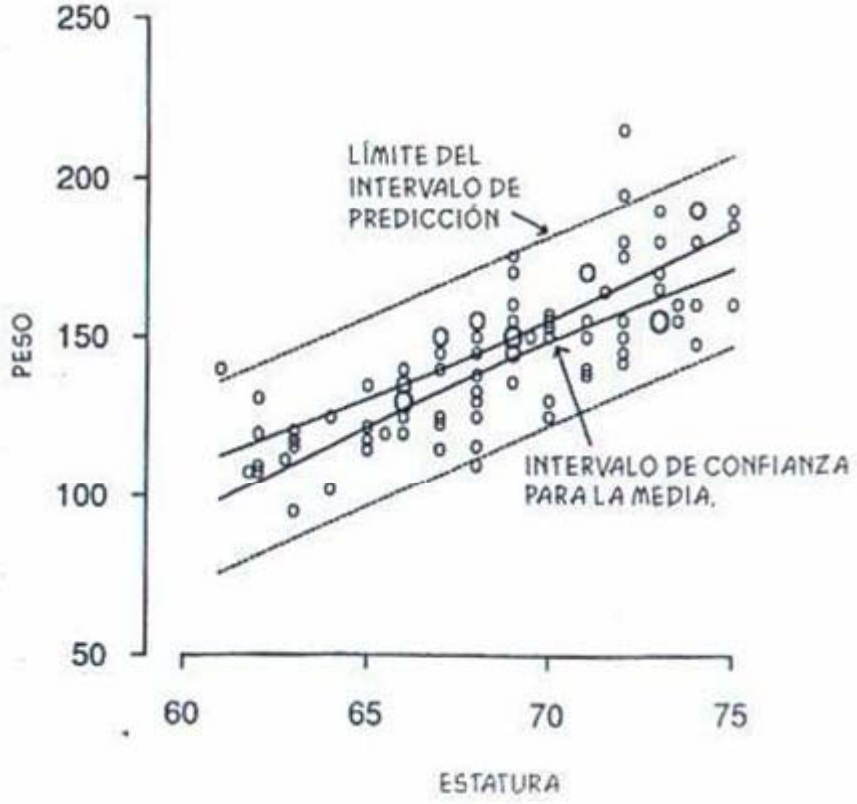
$$= 180 \pm (2{,}365)(29{,}51)$$

$$= 180 \pm 70 \text{ LIBRAS} = [110, 250]$$

LE DECIMOS AL ENTRENADOR DE FÚTBOL QUE ESTAMOS BASTANTE SEGUROS DE QUE EL NUEVO PESA ¡ENTRE 110 Y 250! (ENTRE 50 Y 115 KILOS)

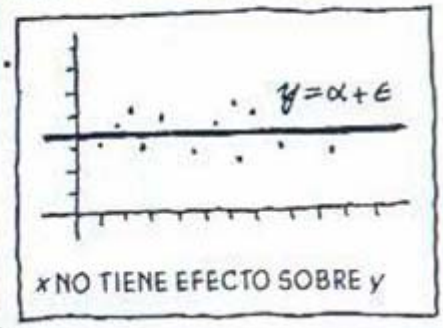
¡LOS INTERVALOS SON BASTANTE HORRIBLES! ¿CUÁL ES EL PROBLEMA? EN REALIDAD HAY DOS PROBLEMAS:

LOS ESTUDIANTES DE PENNSYLVANIA PRESENTAN MEJORES ESTIMACIONES.



Contraste de hipótesis

QUIEN SEA TOTALMENTE ESCÉPTICO PUEDE SUGERIR QUE NO EXISTE NINGUNA RELACIÓN ENTRE LA ESTATURA Y EL PESO. ESTO EQUIVALE A DECIR QUE $\beta = 0$.



TOMAMOS ESTO COMO HIPÓTESIS NULA.

$$H_0 : \beta = 0$$

EN ESTE CASO, EL ESTADÍSTICO

$$t = \frac{b}{SE(b)}$$

TIENE DISTRIBUCIÓN t CON $n-2$ GRADOS DE LIBERTAD. COMO SIEMPRE, LA PRUEBA DE SIGNIFICACIÓN DEPENDE DE LA HIPÓTESIS ALTERNATIVA.

$t > t_\alpha$ PARA $H_a : \beta > 0$

$t < t_\alpha$ PARA $H_a : \beta < 0$

$|t| > |t_{\alpha/2}|$ PARA $H_a : \beta \neq 0$

PARA LOS DATOS DE PESO FICTICIOS, TENEMOS LA FIRME SOSPECHA DE QUE LA HIPÓTESIS ALTERNATIVA DEBERÍA SER

$$H_a : \beta > 0$$

LO PROBAMOS

$$t_{obs} = \frac{5}{SE(b)} = \frac{5}{1{,}62}$$

$$= 3{,}08$$

PARA 7 GRADOS DE LIBERTAD, $t_{0{,}05} = 1{,}895$. DADO QUE $t_{OBS} > t_{0{,}05}$ RECHAZAMOS LA HIPÓTESIS NULA AL NIVEL DE SIGNIFICACIÓN Y CONCLUIMOS AFIRMANDO QUE EXISTE UNA RELACIÓN POSITIVA ENTRE LA ESTATURA Y EL PESO.

Regresión lineal múltiple

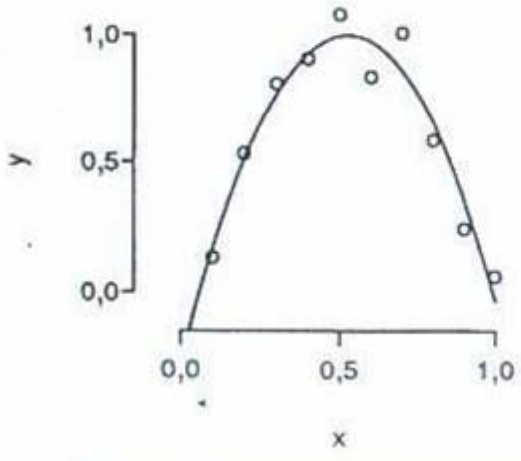
PODEMOS UTILIZAR LAS MISMAS IDEAS FUNDAMENTALES PARA ANALIZAR LA RELACIÓN ENTRE UNA VARIABLE DEPENDIENTE Y DISTINTAS VARIABLES INDEPENDIENTES:

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_n x_n + \epsilon$$

POR EJEMPLO, EL PESO ESTÁ DETERMINADO POR UNA SERIE DE FACTORES DIFERENTES A LA ESTATURA: LA EDAD, EL SEXO, LA DIETA, LA COMPLEXIÓN FÍSICA, ETC.

EL ÁLGEBRA MATRICIAL Y EL ORDENADOR SE COMPLEMENTAN PARA FACILITAR EL ANÁLISIS DE ESTOS PROBLEMAS.

Regresión no lineal



OBVIAMENTE, A VECES LOS DATOS DIBUJAN UNA CURVA NO LINEAL. LOS ESTADÍSTICOS TIENEN UN MONTÓN DE TRUCOS PARA UTILIZAR TÉCNICAS DE REGRESIÓN LINEAL PARA PROBLEMAS NO LINEALES. LO MÁS FÁCIL ES ESCRIBIR Y COMO UNA POLINOMIAL

$$Y = \alpha + \beta_1 x + \beta_2 x^2 + \epsilon$$

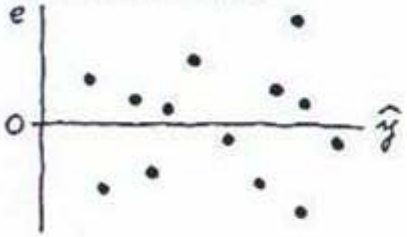
Y TRATAR A x Y A x^2 COMO VARIABLES INDEPENDIENTES EN UN MODELO LINEAL.

Diagnóstico de la regresión

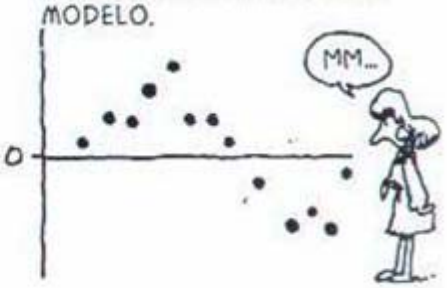
AJUSTAR UN MODELO COMPLEJO A LOS DATOS PUEDE OCULTAR MUCHAS DIFICULTADES. UTILIZAMOS LOS PROCEDIMIENTOS DE DIAGNÓSTICO DE LA REGRESIÓN PARA DESCUBRIR TODO TIPO DE SORPRESAS INDEFINIBLES Y DESAGRADABLES.

EL PROCEDIMIENTO MÁS SIMPLE CONSISTE EN REPRESENTAR EN UN DIAGRAMA DE PUNTOS LOS **RESIDUOS** e_i FRENTE A LA **PREDICCIÓN** $\hat{y}_i$. RECUERDA QUE ASUMIMOS QUE EL ERROR ε ES INDEPENDIENTE DE x.

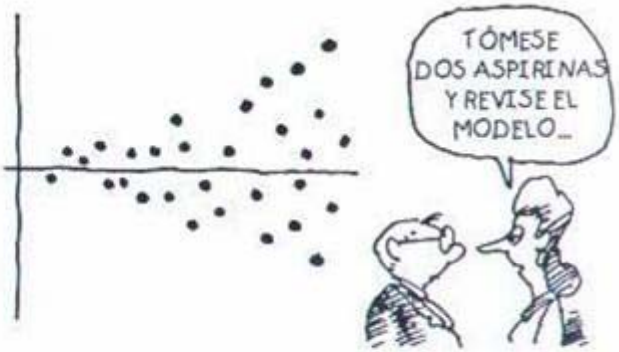
UNA DISPERSIÓN DE PUNTOS ALEATORIA INDICA QUE LAS PRESUNCIONES DEL MODELO SON PROBABLEMENTE CORRECTAS.



CUALQUIER FORMA QUE ADOPTE EL GRÁFICO INDICA PROBLEMAS CON LAS PREMISAS DEL MODELO.



UNA TÍPICA SORPRESA DESAGRADABLE (QUE SE DA EN LOS DATOS DE PESO/ESTATURA) ES QUE LOS ERRORES SON **HETEROCEDÁSTICOS**, ES DECIR, QUE LA DISPERSIÓN DE e AUMENTA A MEDIDA QUE AUMENTA y.

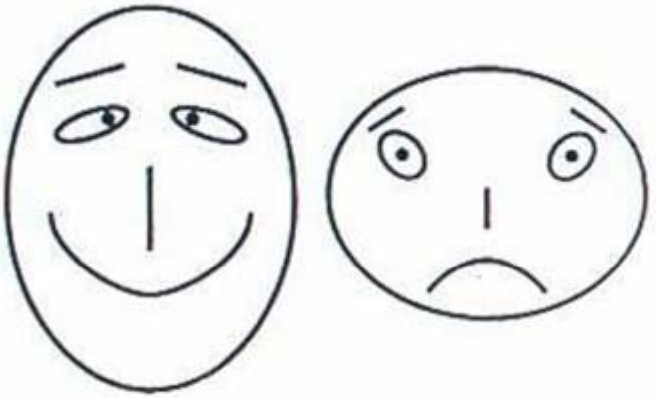


EN ESTE CAPÍTULO, HEMOS RESUMIDO LAS IDEAS FUNDAMENTALES Y LAS TÉCNICAS DEL ANÁLISIS DE REGRESIÓN, QUE ESTUDIA RELACIONES ENTRE VARIABLES. CON ESTO CONCLUIMOS NUESTRA DETALLADA DISCUSIÓN SOBRE LOS MÉTODOS BÁSICOS DE LA ESTADÍSTICA. EN NUESTRO ÚLTIMO CAPÍTULO HAREMOS UN REPASO RÁPIDO DE ALGUNOS PUNTOS QUE FALTAN.

♦ Capítulo 12 ♦
CONCLUSIÓN

REPRESENTACIÓN DE DATOS

HEMOS APRENDIDO A REPRESENTAR UNA VARIABLE EN UN DIAGRAMA DE PUNTOS Y DOS VARIABLES EN UN DIAGRAMA DE DISPERSIÓN DE PUNTOS. PERO, ¿CÓMO PODEMOS REPRESENTAR MÁS DE DOS VARIABLES EN UNA HOJA? ENTRE LAS DISTINTAS POSIBILIDADES, UNA GUÍA EN CÓMIC DEBE MENCIONAR LA SENCILLA IDEA DE HERMAN CHERNOFF: UTILIZANDO EL ROSTRO HUMANO, ASIGNA CADA RASGO A UNA VARIABLE Y DIBUJA LAS RESULTANTES CARAS DE CHERNOFF:



x = CEJAS
y = TAMAÑO DE LOS OJOS
z = LARGO DE LA NARIZ
t = LARGO DE LA BOCA
β = LARGO DE LA CARA
ETC.

Análisis estadístico de DATOS MULTIVARIANTES

UN SURTIDO DE MODELOS MULTIVARIABLES NOS AYUDA A ANALIZAR Y A REPRESENTAR LOS DATOS n-DIMENSIONALES. ALGUNAS DE LAS TÉCNICAS MULTIVARIABLES SON:

Análisis de conglomerados

INTENTA DIVIDIR LA POBLACIÓN EN SUBGRUPOS HOMOGÉNEOS. POR EJEMPLO, SI ANALIZAMOS LOS PATRONES ELECTORALES DEL CONGRESO, DESCUBRIMOS QUE LOS REPRESENTANTES DEL SUR Y DEL OESTE FORMAN DOS CONGLOMERADOS DIFERENTES.

Análisis discriminante

ES EL PROCESO INVERSO. POR EJEMPLO, LA COMISIÓN DE ADMISIONES DE UNA FACULTAD PUEDE QUERER DESCUBRIR DATOS QUE LE AYUDEN A INTUIR SI UN ASPIRANTE SERÁ BUEN ESTUDIANTE (HARÁ GRANDES APORTACIONES MONETARIAS AL FONDO DE LICENCIADOS) O MAL ESTUDIANTE (SALDRÁ PARA HACER ALGO BUENO EN ESTE MUNDO Y NO VOLVERÁ A SABERSE MÁS DE ÉL).

Análisis factorial

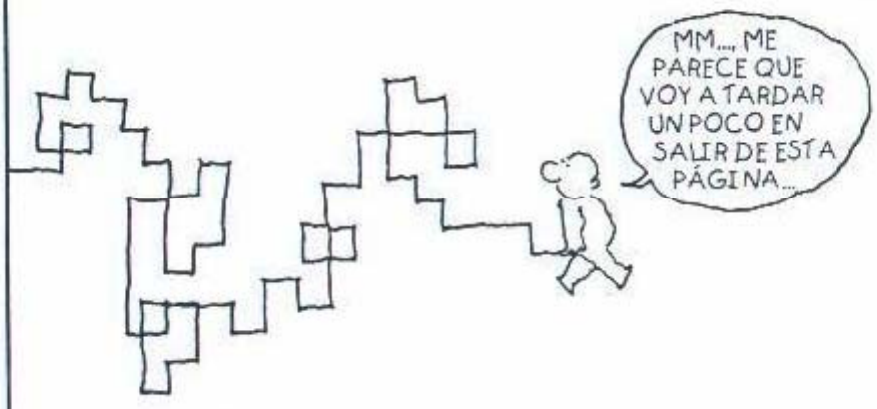
PRETENDE EXPLICAR LOS DATOS DE GRANDES DIMENSIONES CON UN NÚMERO REDUCIDO DE VARIABLES. UN PSICÓLOGO PUEDE ENTREGAR UN TEST CON 100 PREGUNTAS, Y AL MISMO TIEMPO ASUMIR EN SECRETO QUE LAS RESPUESTAS DEPENDEN ÚNICAMENTE DE UNOS POCOS FACTORES: LA EXTROVERSIÓN, EL AUTORITARISMO, EL ALTRUISMO, ETC. LOS RESULTADOS DEL TEST SE RESUMIRÁN ENTONCES UTILIZANDO TAN SÓLO UNAS CUANTAS PUNTUACIONES COMPUESTAS DE LOS FACTORES YA MENCIONADOS.

TODAVÍA HAY MÁS PARA LA

PROBABILIDAD:

Paseos aleatorios. EMPIEZAN CON EL LANZAMIENTO DE UNA MONEDA. SUPONGAMOS QUE AVANZAS UN PASO SI SALE CARA Y RETROCEDES SI SALE CRUZ. (SI USAS DOS MONEDAS LO PUEDES HACER EN DOS DIMENSIONES.) REPETIDOS LANZAMIENTOS PRODUCEN UN PROCESO ESTOCÁSTICO LLAMADO **PASEO ALEATORIO**. LOS MODELOS DE PASEO ALEATORIO SE EMPLEAN EN LAS **OPCIONES DE INVERSIÓN EN BOLSA** Y EN LA **ADMINISTRACIÓN DE LA CARTERA DE ACCIONES**.

Análisis de series temporales. ESTÁ RELACIONADO CON CONJUNTOS DE DATOS QUE, COMO EL PASEO ALEATORIO, SE **VAN ACUMULANDO CON EL TIEMPO**: LAS TEMPERATURAS LOCALES Y GLOBALES, EL PRECIO DEL PETRÓLEO, ETC. EN EL **ANÁLISIS DE SERIES TEMPORALES** LOS VALORES ALEATORIOS SE USAN PARA PREDECIR VALORES FUTUROS.



YA HEMOS VISTO CÓMO NOS AYUDA EL **ORDENADOR** CON LOS ANÁLISIS Y CON LA ARITMÉTICA. EXISTEN MÁS IDEAS ESTADÍSTICAS QUE DEBEN SU EXISTENCIA A LOS ORDENADORES:

Análisis de imagen

LA IMAGEN DE UN ORDENADOR ES DE 1.000 POR 1.000 PÍXELES, CON CADA PUNTO REPRESENTADO A PARTIR DE UNA GAMA DE 16,7 MILLONES DE COLORES EN CADA PÍXEL. EL ANÁLISIS ESTADÍSTICO DE IMAGEN INTENTA EXTRAER ALGÚN SIGNIFICADO DE UNA «INFORMACIÓN» COMO ÉSTA.

Remuestreo

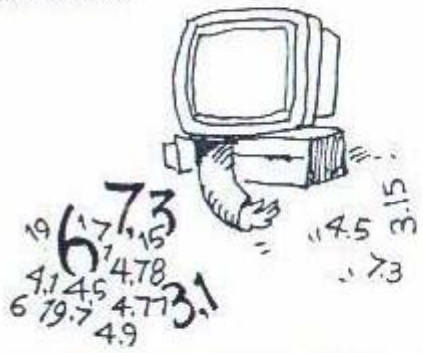
A VECES, LOS ERRORES ESTÁNDAR Y LOS LÍMITES DE CONFIANZA RESULTAN IMPOSIBLES DE ENCONTRAR. ENTONCES, SE INTRODUCE EL **REMUESTREO**, UNA TÉCNICA QUE TRATA LA MUESTRA **COMO SI FUERA LA POBLACIÓN**. ESTAS TÉCNICAS, TAN ATRACTIVAS PARA EL ESTADÍSTICO, RECIBEN NOMBRES COMO ESTIMACIÓN AUTOSUFICIENTE («BOOTSTRAPPING») O HERRAMENTAL («JACKKNIFE»).*

* EL BARÓN DE MUNCHAUSEN AFIRMABA HABER ALCANZADO LA LUNA TIRANDO DE LOS CORDONES DE SUS BOTAS. «BOOTSTRAPPING». [N.T.]

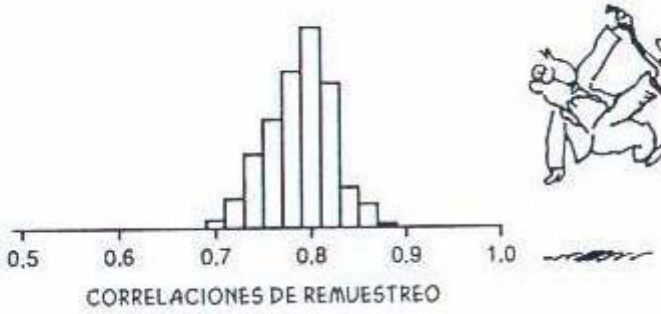
Remuestreo (continuación)

PARA REALIZAR EL REMUESTREO, EL ORDENADOR

- REMUESTREA LA MUESTRA
- CALCULA LA ESTIMACIÓN PARA LA REMUESTRA
- REPITE LOS DOS PRIMEROS PASOS Y ENCUENTRA LA DISPERSIÓN DE LAS ESTIMACIONES REMUESTREADAS.



¿RECUERDAS EL COEFICIENTE DE CORRELACIÓN *r* DE LOS 92 PESOS Y ESTATURAS APAREADOS DEL CAPÍTULO 11? ¿CUÁL ES EL **ERROR ESTÁNDAR DE *r*?** EL ORDENADOR REMUESTREA OTRAS 200 VECES LOS 92 PUNTOS, CALCULA *r* CADA VEZ, Y DIBUJA UN HISTOGRAMA DE LOS VALORES DE *r*.



CORRELACIONES DE REMUESTREO

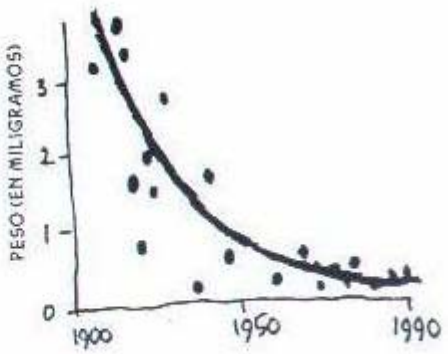
OBSERVA QUE LA DISPERSIÓN DE LAS ESTIMACIONES DEL REMUESTREO ES RELATIVAMENTE **PEQUEÑA**.

CALIDAD DE LOS DATOS

LOS ERRORES APARENTEMENTE PEQUEÑOS DEL MUESTREO, LA MEDICIÓN Y EL REGISTRO DE DATOS PUEDEN ACABAR CON CUALQUIER ANÁLISIS. **R. A. FISHER**, ESTUDIOSO DE LA GENÉTICA Y FUNDADOR DE LA ESTADÍSTICA MODERNA, NO SÓLO DISEÑABA Y ANALIZABA LA CRÍA DE ANIMALES, SINO QUE TAMBIÉN **LIMPIABA SUS JAULAS Y CUIDABA DE ELLOS**, PORQUE SABÍA QUE LA PÉRDIDA DE UN ANIMAL INFLUIRÍA EN SUS RESULTADOS.

LOS ESTADÍSTICOS MODERNOS, CON SUS ORDENADORES, SUS BASES DE DATOS, Y LAS AYUDAS DEL GOBIERNO, HAN PERDIDO ESTE ESPÍRITU DE DEDICACIÓN PERSONAL.

SI REPRESENTAMOS EN UN GRÁFICO LA CANTIDAD MEDIA DE EXCREMENTOS DE RATA DEPOSITADOS EN LAS UÑAS DE ESTADÍSTICOS A LO LARGO DE LA HISTORIA, LA CURVA TENDRÍA PROBABLEMENTE ESTE ASPECTO:



Innovación

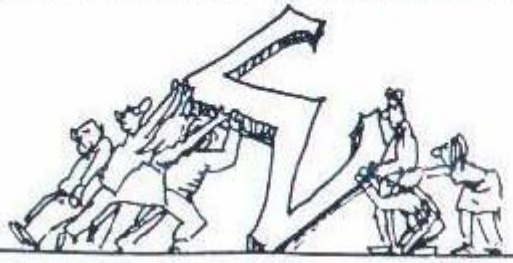
¡LAS MEJORES SOLUCIONES NO SIEMPRE ESTÁN EN LOS LIBROS! POR EJEMPLO, UNA COMPAÑÍA CONTRATADA PARA ESTIMAR LA COMPOSICIÓN DE UN **VERTEDERO DE BASURA** SE ENCONTRÓ CON ALGUNOS PROBLEMAS INTERESANTES QUE NO SE REFLEJAN EN LOS MANUALES...

Comunicación

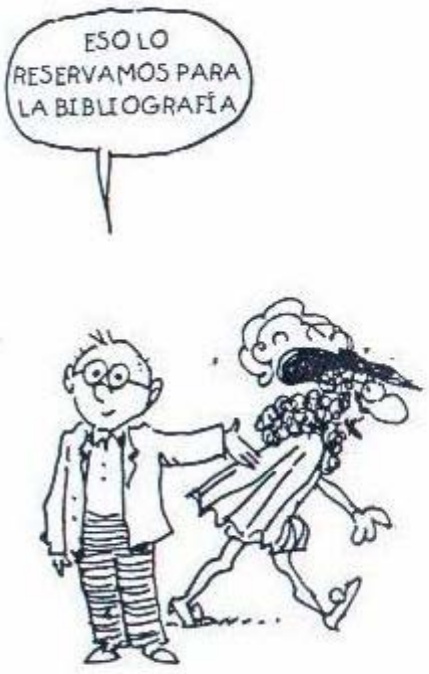
UN ANÁLISIS BRILLANTE RESULTA INÚTIL SI LOS RESULTADOS NO SON COMUNICADOS CON UN LENGUAJE SENCILLO Y CLARO, INCLUIDO EL GRADO DE INCERTIDUMBRE ESTADÍSTICA DE LAS CONCLUSIONES. POR EJEMPLO, EN LA ACTUALIDAD, LOS MEDIOS DE COMUNICACIÓN PUBLICAN MÁS A MENUDO LOS MÁRGENES DE ERROR DE LOS RESULTADOS DE LAS ENCUESTAS QUE REALIZAN.

Trabajo en equipo

EN LA COMPLEJA SOCIEDAD EN QUE VIVIMOS, LA SOLUCIÓN A MUCHOS PROBLEMAS REQUIERE **UN ESFUERZO DE EQUIPO**. INGENIEROS, ESTADÍSTICOS Y TRABAJADORES DE CADENAS DE MONTAJE COOPERAN PARA MEJORAR LA CALIDAD DE SUS PRODUCTOS. BIOESTADÍSTICOS, DOCTORES Y DEFENSORES DE LA LUCHA CONTRA EL SIDA, TRABAJAN EN COLABORACIÓN PARA DISEÑAR PRUEBAS MÉDICAS QUE PERMITAN EVALUAR CON MAYOR RAPIDEZ LAS TERAPIAS.



BUENO, ¡ESO ES TODO! AHORA DEBERÍAS SER CAPAZ DE HACER DE TODO CON LA ESTADÍSTICA, DE TODO, MENOS MENTIR, HACER TRAMPAS, ROBAR O DEDICARTE AL JUEGO.




¿TIENE UN BUEN SEGURO PROFESIONAL DE NEGLIGENCIA ESTADÍSTICA?

BIBLIOGRAFÍA

PARA EL ESTUDIANTE:

MOORE, DAVID S., *STATISTICS: CONCEPTS AND CONTROVERSIES*, 1991, NEW YORK, W.H. FREEMAN. HACE HINCAPIÉ EN LAS IDEAS MÁS QUE EN LOS PROCEDIMIENTOS.

FREEDMAN, DAVID; PISANI, ROBERT; PURVES, ROGER, *STATISTICS*, 1989, NEW YORK, W.H. FREEMAN, W. W. NORTON.

MOORE, DAVID S.; MCCABE, GEORGE P., *INTRODUCTION TO THE PRACTICE OF STATISTICS*, 1989, NEW YORK, W.H. FREEMAN.

SMITH, GARY, *STATISTICAL REASONING*, 1990, BOSTON, ALLYN AND BACON, INC. MÁS TÉCNICO, SUBRAYA LOS ASPECTOS ECONÓMICOS Y FINANCIEROS, PERO CONTIENE EJEMPLOS DE TODOS LOS CAMPOS.

ESTOS TEXTOS SON ACTUALES, CORRECTOS, LITERARIOS E INGENIOSOS. ADEMÁS DE LOS QUE CITAMOS, HAY CIENTOS DE LIBROS DE TEXTO EN EL MERCADO Y PODEMOS DECIR QUE LA MAYORÍA SON MÁS QUE ACEPTABLES.

PARA EL ESTUDIANTE APASIONADO:

PYRCZAK, FRED, *STATISTICS WITH A SENSE OF HUMOR*, 1989, LOS ANGELES, FRED PYRCZAK PUBLISHER. UN LIBRO ELEMENTAL DE EJERCICIOS Y UNA GUÍA PARA LA RESOLUCIÓN DE PROBLEMAS ESTADÍSTICOS.

CÓMO MENTIR, HACER TRAMPAS Y APOSTAR. VUESTROS ANGELICALES AUTORES TIENEN MUY POCA EXPERIENCIA EN ESTOS TEMAS. AQUÍ PRESENTAMOS UN PAR DE RECOMENDACIONES DE LOS ENTENDIDOS:

HUFF, DARRELL, *HOW TO LIE WITH STATISTICS*, WITH PICTURES BY IRVING GEIS, NEW YORK, 1954, W.W. NORTON. ES BARATO Y SIGUE EN EL MERCADO.

JAFFE, A. J.; SPIRER, HERBERT F., *MISUSED STATISTICS: STRAIGHT TALK FOR TWISTED NUMBERS*, 1987, NEW YORK, MARCEL DECKER. DE UNA SERIE MUY POPULAR DE LIBROS DE ESTADÍSTICA.

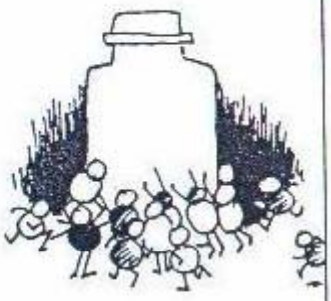
ORKIN, MIKE, *CAN YOU WIN?*, 1991, NEW YORK, W.H. FREEMAN. CONSEJOS DE UN EXPERTO EN PROBABILIDAD Y JUEGO.

MCGERVEY, JOHN D., *PROBABILITIES IN EVERY DAY LIFE*, 1989, N.Y. IVY BOOKS. JUEGOS DE CARTAS DESDE EL BLACKJACK AL SMOKING.

LEGISLACIÓN Y SOCIEDAD:

GASTWIRTH, JOSEPH L., *STATISTICAL REASONING IN LAW AND POLICY, VOL. 1 & 2*, 1988, SAN DIEGO, ACADEMIC PRESS. LOS INTRÍNGULIS LEGALES, CON CASOS DE SELECCIÓN DE JURADOS COMO EL QUE VIMOS EN EL CAPÍTULO 9.

STEERING COMMITTEE OF THE PHYSICIANS' HEALTHY STUDY RESEARCH GROUP, *FINAL REPORT ON THE ASPIRIN COMPONENT OF THE ONGOING PHYSICIANS' HEALTHY STUDY*, THE NEW ENGLAND JOURNAL OF MEDICINE, VOL. 321, PP.129-135.



EL COMENTARIO NO JURÍDICO SOBRE POKER HECHO DESDE EL ESTRADO, QUE APARECE EN EL CAPÍTULO 8, SE EXTRAJO DE UN CASO REAL. NOS LO HA CONFIRMADO PERSONALMENTE EL DR. JOHN DE CANI, DE LA UNIVERSIDAD DE PENNSYLVANIA.

REPRESENTACIÓN GRÁFICA DE DATOS:

TUFTE, EDWARD R., *THE VISUAL DISPLAY OF QUANTITATIVE INFORMATION*, 1983, CHESHIRE, CONNECTICUT, GRAPHICS PRESS.

TUFTE, EDWARD R., *ENVISIONING INFORMATION*, 1990, CHESHIRE, CONNECTICUT, GRAPHICS PRESS. HISTORIA, ARTE Y CIENCIA DE LOS GRÁFICOS. AMBOS LIBROS SON YA UNOS CLÁSICOS.

CLEVELAND, WILLIAM S., *THE ELEMENTS OF GRAPHING DATA*, 1985, PACIFIC GROVE CA., WADSWORTH ADVANCED BOOKS AND SOFTWARE. PRINCIPIOS PARA EL DISEÑO DE GRÁFICOS POR ORDENADOR.

HISTORIA:

DAVID, F. N., *GAMES, GODS AND GAMBLINGS*, 1962, NEW YORK, HAFNER, NEW YORK.

STIGLER, STEPHEN M., *THE HISTORY OF STATISTICS: THE MEASUREMENT OF UNCERTAINTY BEFORE 1900*, 1985, CAMBRIDGE, MA., BELKMAN PRESS OF HARVARD UNIVERSITY PRESS.

BOX, JOAN FISHER, *R.A. FISHER, THE LIFE OF A SCIENTIST*, 1978, NEW YORK, WILEY. BIOGRAFÍA, ESCRITA POR SU HIJA, DEL PERSONAJE MÁS INFLUYENTE Y CONTROVERTIDO DE LA ESTADÍSTICA DEL SIGLO XX.

KRUSKAL, WILLIAM, «*THE SIGNIFICANCE OF FISHER: A REVIEW OF R. A. FISHER: THE LIFE OF A SCIENTIST*», 1980, JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, VOL. 75, 1030. ANALIZA LA BIOGRAFÍA DE FISHER CON OBJETIVIDAD Y CONTIENE UNA FORMIDABLE BIBLIOGRAFÍA.

SOFTWARE ESTADÍSTICO:

EN ESTE LIBRO HEMOS USADO EL SISTEMA DE SOFTWARE ESTADÍSTICO **MINITAB** (MINITAB INC., STATE COLLEGE, PA). LOS DATOS DE PESOS Y ESTATURAS DE LOS ESTUDIANTES DE PENNSYLVANIA SON DEL **PULSE DATA** INSTALADO EN ESTE SISTEMA. LOS GRÁFICOS POR ORDENADOR HAN SIDO GENERADOS POR **S-PLUS** (STATISTICAL SCIENCE INC., SEATTLE, WA) CON UN PC CLÓNICO 486. **S** ES UN SOFTWARE SOFISTICADO, DESARROLLADO POR **AT&T** BELL LABS PARA ANÁLISIS AVANZADOS Y REPRESENTACIONES GRÁFICAS.

RYAN, BARBARA; JOINER, BRIAN; RYAN, THOMAS, *MINITAB HANDBOOK* (PWS-KENT, BOSTON, 1985) Y *THE STUDENT EDITION OF MINITAB* (ADDISON WESLEY) SON INTRODUCCIONES A LA ESTADÍSTICA COMPUTACIONAL RÁPIDAS Y NADA CARAS. **MINITAB** FUNCIONA EN ORDENADORES CENTRALES, PC COMPATIBLES Y MACINTOSH.

SE VENDEN MUCHOS PAQUETES DE SOFTWARE DE ALTA CALIDAD PARA ORDENADOR PERSONAL, POR EJEMPLO:

DATADESK (DATA DESCRIPTION, ITHACA, NY), PARA MACINTOSH.

SAS (SAS INSTITUTE INC., CARY, NC), ***SPSS*** (SPSS INC., CHICAGO, IL) Y ***BMDP*** (BMDP STATISTICAL SOFTWARE, INC., LOS ANGELES, CA) FUERON DISEÑADOS ORIGINALMENTE PARA SISTEMAS DE ORDENADORES CENTRALES Y AHORA HAN PASADO A LOS PC. SE PRESENTA COMPLETO CON WINDOWS.

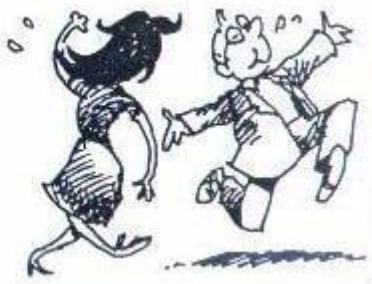
STATGRAPHICS (STATISTICAL GRAPHICS CORP., PRINCETON, NJ) PARA PC.

STATVIEW (ABACUS CONCEPTS, OAKLAND, CA) PARA MACINTOSH.

SYSTAT (SYSTAT, INC., EVANSTON, IL) CON SISTEMAS QUE FUNCIONAN EN TODOS LOS ENTORNOS.

ESTOS PAQUETES DIFIEREN EN DETALLES IMPORTANTES; TIENES QUE SER UN COMPRADOR INTELIGENTE. TE RECOMENDAMOS LA ELECCIÓN DE UN SISTEMA QUE YA HAYAN PROBADO TUS COMPAÑEROS. ENTRE NOSOTROS HAY POCOS PIONEROS NATOS DEL SOFTWARE ESTADÍSTICO. CUANDO SE APRENDE A USAR UN NUEVO SISTEMA, DEBE EXPERIMENTARSE CON CONJUNTOS PEQUEÑOS DE DATOS CONOCIDOS. RECUERDA: **LO MÁS CARO DE CUALQUIER SOFTWARE ES TU TIEMPO.** EL LEMA DEL CÓMIC PARA APRENDER ESTADÍSTICA COMPUTACIONAL ES:
FAMILIARIZARSE CON UN TEMA DA SUS FRUTOS.

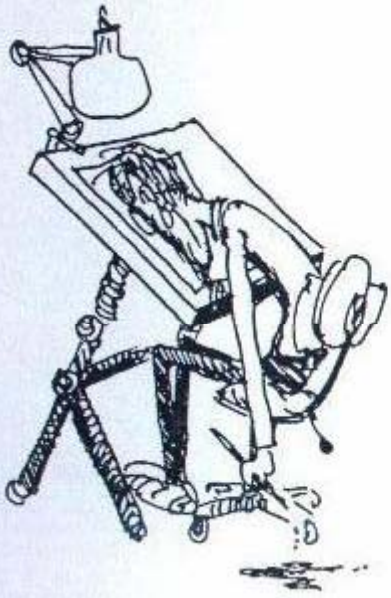
INTENTAR APRENDER TEORÍA ESTADÍSTICA Y ESTADÍSTICA COMPUTACIONAL AL MISMO TIEMPO ES CASI COMO INTENTAR ANDAR Y MASCAR CHICLE A LA VEZ. PARA REALIZAR CADA UNA DE ESTAS ACCIONES SE PONEN EN MARCHA DIFERENTES MECANISMOS Y PROCESOS MENTALES. RESERVA MOMENTOS DIFERENTES PARA APRENDER CADA UNA DE ESTAS COSAS, Y DESPUÉS JÚNTALAS. ASÍ, PODRÁS CONVERTIRTE EN UN ESTADÍSTICO RENACENTISTA, COMECHICLE, PASEANTE Y COMPUTACIONAL.



ÍNDICE

ACERCA DE LOS AUTORES

WOOLLCOTT SMITH ES PROFESOR DE ESTADÍSTICA EN LA UNIVERSIDAD DE TEMPLE. LICENCIADO Y MÁSTER EN CIENCIAS POR LA UNIVERSIDAD ESTATAL DE MICHIGAN, Y DOCTOR (PH. D.) POR LA JOHNS HOPKINS UNIVERSITY. ES AUTOR Y COAUTOR DE MÁS DE CUARENTA PUBLICACIONES DE ÁREAS TAN DIVERSAS COMO EL ESTUDIO DE DERRAMES DE PETRÓLEO, LA TEORÍA ESTADÍSTICA Y LA ESTADÍSTICA MEDIOAMBIENTAL. HA SIDO CONSEJERO DE VARIOS PROGRAMAS CIENTÍFICOS DE LOS ESTADOS UNIDOS. SUELE MANTENER INTERESANTES DISCUSIONES Y HACER PIRAGÜISMO CON SU MUJER, LEAH, Y SUS DOS HIJOS, KESTON Y AMELIA.



LARRY GONICK ES AUTOR Y COAUTOR DE UNA SERIE DE LIBROS DE NO FICCIÓN EN CÓMIC, ADEMÁS DE LA SECCIÓN *SCIENCE CLASSICS*, QUE APARECE CADA DOS MESES EN EL *DISCOVER MAGAZINE*. ABANDONÓ LOS ESTUDIOS DE DOCTORADO EN MATEMÁTICAS EN HARVARD, AUNQUE PARECE QUE HA VUELTO AL PUNTO DE PARTIDA. VIVE EN SAN FRANCISCO CON SU MUJER, LISA, Y SUS DOS HIJAS, SOPHIE Y ANNA, Y ESPERA LLEGAR A COMPRENDER LA VIDA MIENTRAS CONTINÚA ENCADENADO A SU MESA DE DIBUJO.

Si usted es de los que creen que la vida no es más que una variable aleatoria, y vaga por el espacio dimensional *n* en busca de más grados de libertad, *La estadística en cómic* le resultará imprescindible para iniciarse en el conocimiento de esta ciencia.

La estadística en cómic abarca todos los aspectos de la estadística moderna: el análisis y la descripción de datos, las probabilidades en el juego y la medicina, las variables aleatorias, las pruebas de Bernoulli, el teorema central del límite, el contraste de hipótesis, la estimación de intervalos de confianza, y mucho más. Todo ello explicado con ilustraciones simples, claras, y, sí, divertidas. ¡Nunca volverá a pedir la distribución de Poisson en un restaurante francés!

Larry Gonick es autor y coautor de otras cuatro guías en cómic. Woollcott Smith es un estadístico muy activo en el campo de la investigación y el asesoramiento empresarial, además de profesor de la Universidad de Temple.

«Gonick es único en su especie.» —*Discover*

editorial
Zendrera Zariquiey

ISBN 84-8418 041-7
9 788484 180418