Apuntes [esbozo] de Estadística para Marina

Pedro Fortuny Ayuso

2009-2010

Correo electrónico: fortunypedro@uniovi.es

Índice

Correlación y regresión lineal

1. Regresión y correlación

Por lo general, los datos que uno recibe o están agrupados en atributos cualitativos, o están agrupados en atributos cuantitativos o una mezcla de ambos, que suele ser lo general. Tomemos una tabla "al azar", parte de cuyos datos es la siguiente (la tabla está tomada del programa de cálculo estadístico **R**):

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
12	16	256	9.7	69	5	12
13	11	290	9.2	66	5	13
14	14	274	10.9	68	5	14
15	18	65	13.2	58	5	15
16	14	334	11.5	64	5	16
17	34	307	12.0	66	5	17
18	6	78	18.4	57	5	18
19	30	322	11.5	68	5	19
20	11	44	9.7	62	5	20
21	1	8	9.7	59	5	21
22	11	320	16.6	73	5	22
23	4	25	9.7	61	5	23
24	32	92	12.0	61	5	24
28	23	13	12.0	67	5	28
29	45	252	14.9	81	5	29
30	115	223	5.7	79	5	30

TABLE 1. Calidad del aire en NY

Es una tabla de calidad del aire (contenido de ozono, que es uno de los indicadores básicos). Los datos son de Nueva York, de mayo a septiembre

de 1973. Solar.R es la "radiación solar" en una medida estándar, Wind es la velocidad del viento, Temp la temperatura ambiente y Month, Day son obvios.

Si uno dibuja una gráfica de la cantidad de ozono contra la radiación solar, encuentra algo como la figura 1. Mientras que si uno dibuja una gráfica

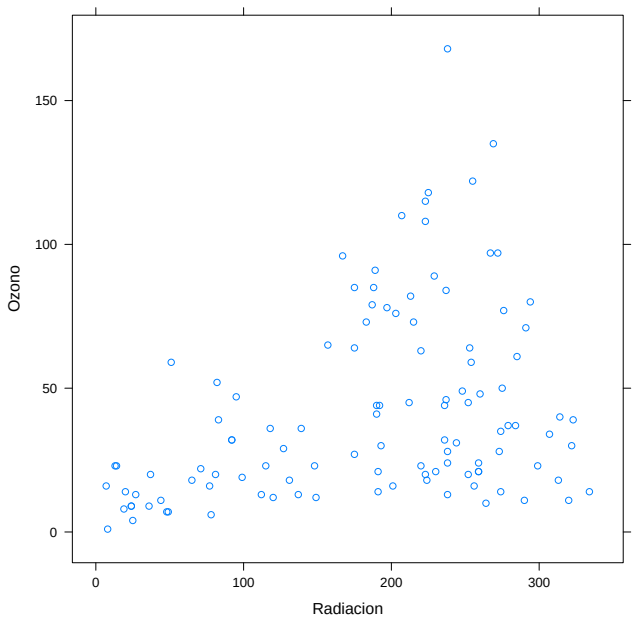


FIGURA 1. Cantidad de ozono contra radiación solar

de la cantidad de ozono contra la velocidad del viento, obtiene la figura 2.

Parece, según los gráficos, que la radiación solar influye en la cantidad de ozono de la atmósfera de alguna manera y que lo mismo le pasa a la velocidad del viento. Cuanta más radiación hay, más parece que es la cantidad de ozono, y cuanta mayor es la velocidad del viento, menor es la cantidad de ozono. De todos modos, no está claro. Pero son dos preguntas que parece natural hacerse.

Cuando uno solo tiene una variable, el "dato" importante que resume todo es la media. Pero la media es un dato que puede engañar. Para saber si la media es o no relevante, se utiliza la desviación típica. En los datos que tenemos, fijándonos en los diagramas de caja y bigotes (de Tukey), se observa que

- La distribución de la concentración de ozono está bastante centrada respecto de la media (la caja es pequeña y los bigotes también), aunque hay algún dato extraño.
- La distribución de la radiación solar es bastante "mala": la caja es enorme (es la mitad del recorrido), los bigotes llegan hasta los

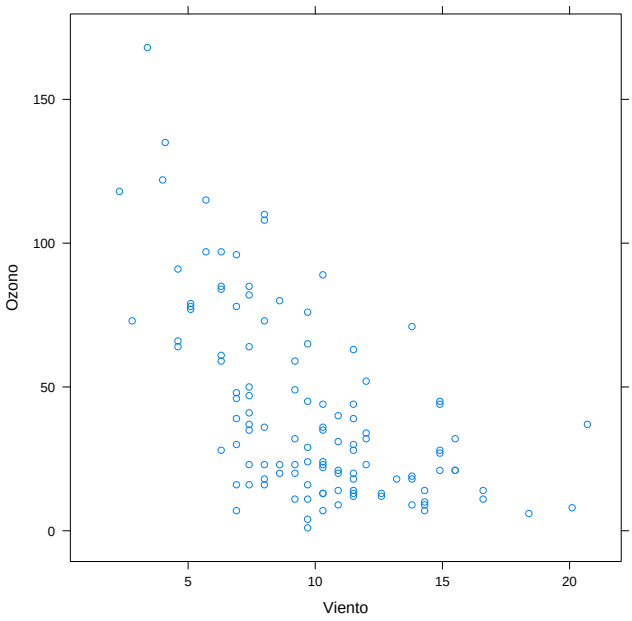


FIGURA 2. Cantidad de ozono contra velocidad del viento

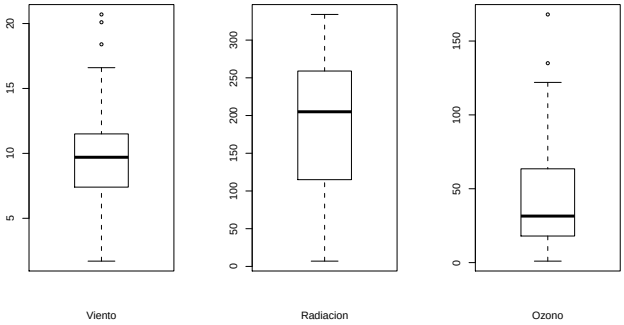


FIGURA 3. Diagramas de Tukey de todas las variables

extremos de la muestra y no se puede decir que la media sea relevante.

- La distribución de la velocidad del viento está bastante centrada, aunque al haber unos cuantos datos altos singulares, los bigotes se alargan mucho. De todos modos uno podría decir que la media es relevante en este caso.

Lo que queremos comprobar ahora es si la radiación solar "influye" o no en la cantidad de ozono y, por otro lado, si la velocidad del viento lo hace o no.

Desde el punto de vista "racional" (sentido común), parece que el ozono ha de ser mayor si hay más sol (pues el calor provoca fenómenos relacionados con la polución que hacen incrementar el ozono en la atmósfera). El hecho de que la velocidad del viento haga disminuir o incrementar la concentración de ozono podría ser "de sentido común" o una casualidad. Quizás al haber más viento se "acerca más aire limpio" (ya sea del mar, en el caso de NY o del "campo"), a la vez que se "libera" aire contaminado. Pero quién sabe.

Necesitamos una manera de calcular "cómo de relacionadas" están dos familias de observaciones. Esa manera se llama co-relación (relación conjunta). Y se expresa mediante un coeficiente: el coeficiente de correlación.

Lo que queremos estudiar son preguntas del tipo:

¿cuánto cambia una variable (p.ej. la concentración de ozono) si cambio otra? ¿cuánto sube una variable (el ozono) al subir o bajar otra (la velocidad del viento)? ...

Variación respecto de variación. "Cuánto varía"... Cuánto varía una variable es una cantidad que se mide con ... la varianza (precisamente la varianza mide "la variación (cuadrática, pero esto da igual) media" de un dato. Si tenemos dos datos, tenemos dos varianzas (la del ozono y la de la radiación solar, p. ej. o la del ozono y la de la velocidad del viento). En realidad, igual que la varianza, podríamos usar la desviación típica, que es lo que vamos a tomar.

Uno podría calcular el cociente de las dos variaciones, pero eso no nos hablaría de la "variación a la vez", sino simplemente de lo mucho que varía uno respecto de lo mucho que varía otra. Necesitamos "juntar" las dos variables... Necesitamos calcular variaciones a la vez... ¿A la vez? Uno podría pensar en sumar las variables, pero (por la razón que sea, que hay una) la operación correcta es multiplicarlas. De hecho, como nos intersa como "varían a la vez", hacemos los cálculos "respecto de sus medias":

Cov(X,Y) = Var((X-x)(Y-y))

Definimos la covarianza de dos tablas de datos como la varianza de "la tabla de datos producto de las diferencias respecto de las medias". Se fabrica una tabla de datos muy grande formada por los productos de cada dato menos la media de ese dato, y se calcula la varianza. A este número se le denomina "covarianza" de los dos datos. En nuestro ejemplo, tenemos:

Cov(ozono, Solar.R) = 630.1028 Cov(ozono, wind) = -34.85336

Ahora bien: estos números dicen bastante poco. Porque tienen unidades y con las unidades no se puede hacer mucho.

Este fenómeno es el mismo que ocurre al estudiar la desviación típica y la media. Que una muestra de pesos tenga una desviación típica de 2Kg no significa nada si no se sabe la media (2Kg en conejos es mucho, en elefantes es poco). Con la covarianza pasa lo mismo: 630 puede ser mucho o poco, dependiendo de cuánto varíen tanto la luz solar como el ozono. Lo mismo para el ozono y el viento. Por eso, el número realmente interesante es el

coeficiente de correlación:

$$r = \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)}$$

La covarianza (que tiene como unidades el producto de las unidades de las tablas) dividida por el producto de las desviaciones típicas. Así se obtiene un término adimensional (sin dimensiones). Este número mide cuánto "tiende a cambiar" uno de los datos si cambia el otro: para el caso del ozono respecto del sol se obitene $r = 0.3483$ y para el caso del ozono respecto del viento, se obtiene $r = -0.6124$. Un coeficiente positivo significa que "en la muestra, al aumentar un dato aumenta el otro", mientras que un coeficiente de correlación negativo indica "al aumentar un dato disminuye el otro".

Pero eso es "en media" y "en la muestra". ¿Podemos concluir que las cosas son así en la realidad? Y la segunda pregunta importante, ¿podemos concluir que un dato depende realmente de otro a la vista de una correlación?

Por ejemplo, si hacemos la nube de puntos del sol respecto de la velocidad del viento: se percibe que no hay ninguna tendencia. ¿Cuál es el coeficiente

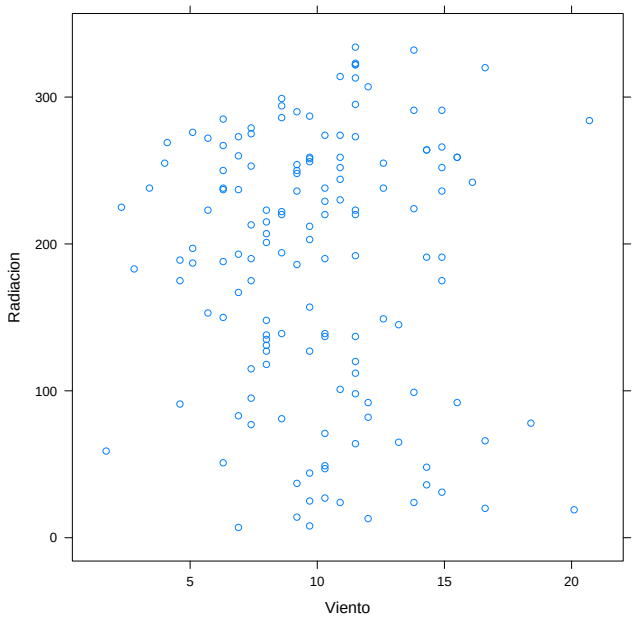


FIGURA 4. Cantidad de radiación solar contra viento

de correlación? La respuesta es $r = -0.127$, que significa que "parece que" los días que el viento es fuerte puede hacer un poco menos de sol. Pero no parece que esto explique mucho la gráfica tan "desordenada" que aparece. Esto es debido a que "la variación" en la cantidad de sol y la "variación" en la cantidad de viento influyen poco una en otra.

Por el contrario, aunque poco, el hecho de que el coeficiente de correlación entre la radiación solar y la cantidad de Ozono sea 0.34 significa que "el cambio en una de las variables coincide con el cambio en otra de las variables", digamos que una de cada tres veces (por decirlo de alguna manera). La correlación en el caso del ozono y el viento es mucho más representativa, del orden de -0.65, lo cual dice que en los datos observados, a mayor viento, menor ozono y viceversa (como ya dijimos, el signo del coeficiente indica que la correlación es inversa).

El dato de la correlación es importante a la hora de dirimir si "una variable depende de otra".

Pero no lo explica todo.

Por otro lado, el *dato que se suele usar* no es r, sino su cuadrado r^2, que en los casos de arriba es:

ozono-viento: $r^2 = 0.375\ldots$
ozono-sol: $r^2 = 0.121\ldots$
viento-sol: $r^2 = 0.016$

y se suelen tomar como relevantes datos para los cuales r^2 es grande. ¿Cómo de grande? Depende. Pero como está entre 0 y 1 (esto es algo que debería haber dicho), cuanto más cerca de 1 está r, más "relevante" parece es la relación entre los dos datos.

2. Correlación y causalidad

Un riesgo que se corre al hacer estudios de correlaciones es pensar que la correlación genera causalidad. Es decir, vistas las tablas de arriba y los resultados, uno podría tender a decir que "el viento es causa" de que haya menos ozono mientras que "el sol se puede decir que no es causa del aumento o disminución del ozono".

Eso es **mentira**. O más bien *no se deduce del estudio que hemos hecho.*

Para ejemplificarlo, tomemos la siguiente tabla de datos:

9.95	1930082	9.83	2126342
9.95	1921990	9.92	2171196
9.95	1922966	9.92	2153859
9.95	1929476	9.92	2151890
9.95	1983001	9.86	2173060
9.95	1975126	9.86	2179187
9.95	1936476	9.86	2271947
9.95	1895135	9.85	2282288
9.95	1911907	9.85	2315255
9.95	1902106	9.84	2318212
9.95	1930627	9.84	2317586
9.95	1916776	9.84	2340708
9.95	1993137	9.79	2394871
9.95	1981309	9.79	2407193
9.95	1977961	9.79	2419960
9.93	2022190	9.78	2446796
9.93	2042304	9.78	2452154
9.93	2089378	9.78	2401400
9.93	2108384	9.77	2452506

TABLE 2. Récord de 100m y muertes en USA de 1968 a 2005

Si dibujamos el diagrama de dispersión[1], obtenemos la figura 5 y se

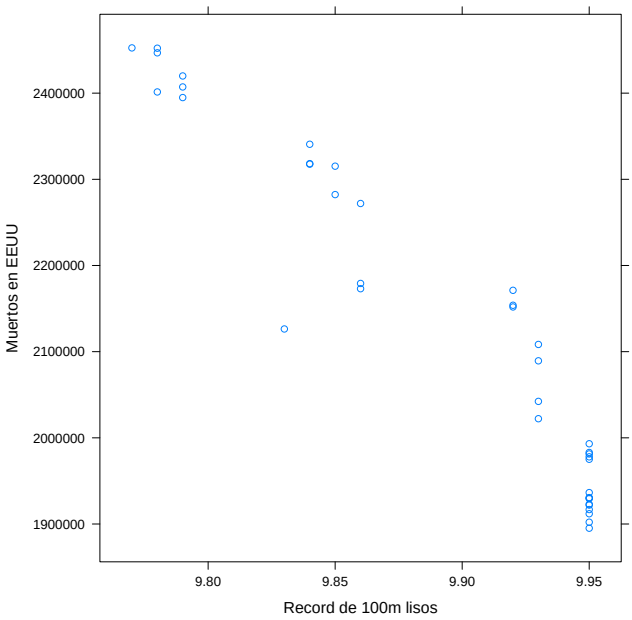


FIGURA 5. Diagrama de dispersión: récord mundial de 100m y muertes en USA

observa claramente que, según aumenta un dato disminuye otro. De hecho, el parámetro r^2 es, para esta tabla, grandísimo: $r^2 = 0.9806$, lo que indica que "eso que se ve", es decir, que al tiempo que aumenta el número de muertes en los Estados Unidos decrece el récord del mundo de velocidad. ¿Hay alguna relación causal? ¿Es quizás que la velocidad de Carl Lewis produce muchas muertes pero la de Usain Bolt todavía más? O ¿quiás es que la mayor cantidad de gente muerta hace que los velocistas tengan más miedo y por eso corran más rápido?

En fin: algo tiene que estar ocurriendo. Que no nos oculten la verdad.

Lo que está ocurriendo no es más que... Que los datos han sido escogidos con intención: primero de todo, *el récord del mundo de velocidad solo puede bajar*. Y no va a variar mucho con el paso del tiempo (alrededor del 0.02% cada vez que baja). Un dato que solo puede disminuir no será un buen dato para una estadística, pues su evolución es unidireccional.

Segundo: el número de muertes en los USA durante una temporada (la que hemos escogido) tenderá a subir, también poco a poco, pues la población aumenta y *el número de muertes anuales* es más o menos creciente con la población (no así la tasa de mortalidad, pero los datos han sido

[1]Los diagramas en que se muestran dos variables en coordenadas cartesianas

escogidos voluntariamente). Su evolución, por tanto, será prácticamente unidireccional durante una época (como ocurre en nuestra tabla).

De ahí que la correlación sea tan grande: las poblaciones son deterministas y han sido "escogidas". Esto se expresa diciendo que *los datos están cocinados*.

Si en lugar de tomar el récord del mundo tomamos otra variable un poco más "aleatoria", como el ratio de variación entre el récord del año anterior y el presente, se obtiene una gráfica como sigue:

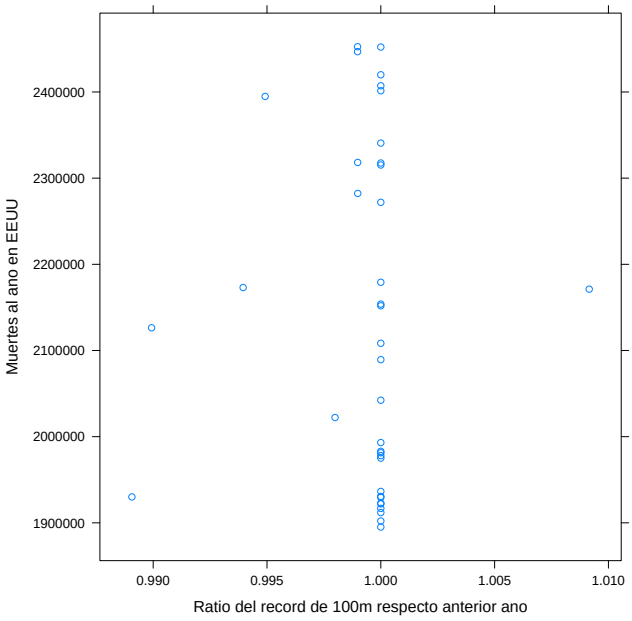


FIGURA 6. Número de muertos en USA y *ratio* entre récords del mundo, de 1968 a 2005

Que muestra una dispersión rara: los récord del mundo cambian muy poco (cambian muy pocas veces y por eso el ratio es 1 casi siempre) mientras que el número de muertos desciende siempre. El coeficiente r^2 es 0.01: no hay prácticamente correlación entre los dos datos. Esto es más natural.

La conclusión más clara es: el coeficiente de correlación r y su valor asociado r^2 son útiles para indicar si *cuánto cambia una variable* al cambiar otra, en media. Pero *no sirven para indicar* si una variable depende o no de otra. Para esto hace falta una explicación externa. Porque bien puede deberse a la casualidad, a que los datos han sido preparados ó a que hay una explicación externa (que *habitualemente* consiste en que "el tiempo pasa" y con el tiempo unos valores tienden a crecer y otros valores tienden a decrecer).

En general, uno debe tender a desconfiar de los datos que presentan altos coeficientes de correlación (desde luego, si el coeficiente r^2 es más de 0.9, hay que sospechar que los datos están preparados o bien que se está enunciando una tautología).

3. Regresión

Pero el mero hecho de que dos datos presenten correlación no es suficiente para las aplicaciones que uno espera. Se supone que, vista una correlación y *supuesto que es razonable*, se ha de poder *predecir* de alguna manera el comportamiento de una variable respecto de la otra. Esto es lo que se conoce como *regresión*.

En el ejemplo de la dependencia del ozono respecto del viento, nos gustaría poder predecir (obviamente con algo de error, pero no demasiado) cuánto ozono va a haber en función de la velocidad del viento. Para eso está la *recta de regresión*: una recta (la aproximación más simple) que minimiza el error de aproximación a una tabla de datos.

DEFINICIÓN 3.1. Dadas dos tablas de datos X, Y, la *recta de regresión* es la recta $y = a + bx$ que minimiza la suma de errores cuadráticos medios entre los puntos (X_i, Y_i) y los valores $(X_i, a + bX_i)$.

Por decirlo de alguna manera, la *recta de regresión* es la *ley lineal* que mejor aproxima la nube de puntos respecto del error cuadrático medio.

Para calcular la recta de regresión hacen falta los coeficientes a y b, que se denominan, respectivamente, *intercepción* y *pendiente*. Los nombres vienen de que a es el valor que toma la $y = a + bx$ cuando $x = 0$ (punto en que corta al eje OY) mientras que la pendiente es la inclinación de la recta $y = a + bx$ (la tangente del ángulo que forma con el eje OX): si $b > 0$, la recta es "creciente", si $b < 0$, la recta es decreciente.

Pues bien, se tienen las siguientes fórmulas:

pendiente: Para calcular a, se hace

$$b = r_{XY} \frac{\sigma_Y}{\sigma_X},$$

donde r_{XY} es el coeficiente de correlación de X e Y y σ es la *desviación típica* correspondiente.

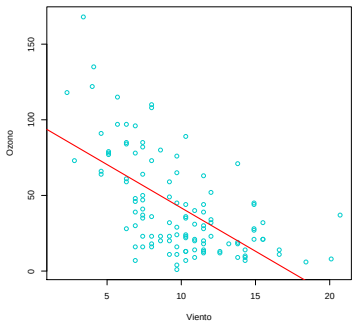
intercepción: Se calcula una vez calculada la pendiente:

$$a = \overline{Y} - b\overline{X},$$

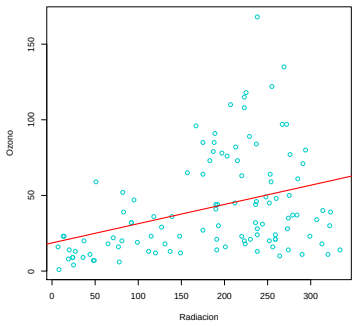
donde $\overline{Y}$ y $\overline{X}$ son las medias muestrales de Y y X, respectivamente.

Para terminar, se muestran los tres ejemplos principales de rectas de regresión en cada nube de puntos.

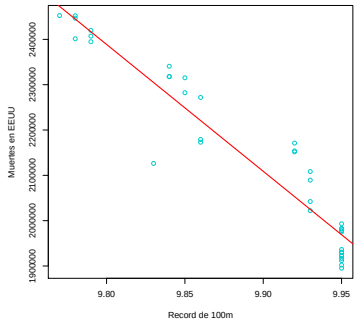
Con la última recta de regresión, se podría predecir que para que el récord del mundo llegara a ser 9.58 (marca actual de Usain Bolt en octubre de 2009), haría falta que murieran este año unos 3.050.000 americanos, cosa



$a = 99.041, b = -5.729, r^2 = 0.375$



$a = 18.598, b = 0.1272, r^2 = 0.121$



$a = 29816007, b = -2798695, r^2 = 0.986$

FIGURA 7. Rectas de regresión de los tres ejemplos principales

que *no va a suceder*, ni de lejos: las inferencias estadísticas hay que tomarlas primero de todo *con un grano de sal*.

Probabilidad. Funciones de distribución

4. Introducción a la probabilidad

La probabilidad es una manera de medir "cuánta posibilidad hay de que algo ocurra". Para hablar con propiedad, se establece siempre un *conjunto* —el espacio de probabilidad— en el que se contienen los *eventos* (cada evento es un subconjunto del conjunto principal). Por lo general, nosotros vamos a trabajar siempre con $\mathbb{R}$, aunque alguna vez haremos referencia a espacios de probabilidad finitos y a los números naturales $\mathbb{N}$.

Una función de probabilidad es una manera de asignar a cada subconjunto del espacio total ("todo lo que puede ocurrir") un valor entre cero y uno, que indica "las posibilidades que hay de que ocurra esto". La asignación es *arbitraria*, o más bien, es la más razonable (no hay matemáticas en esta asignación), pero se han de cumplir unas reglas de sentido común. Digamos que X es es espacio de probabilidad ("el espacio que describe todo lo que puede pasar") y sean A y B dos subconjuntos ("eventos" que pueden ocurrir). Para cada subconjunto $A \subset X$, $p(A)$ es "la probabilidad de que A ocurra". Entonces:

(1) La probabilidad de que ocurra algo es 1. Es decir, $p(X) = 1$.
(2) La probabilidad de que *no ocurra nada* es 0. Es decir, $p(\emptyset) = 0$.
(3) Si A y B son disjuntos, $p(A \cup B) = p(A) + p(B)$. Es decir, si dos eventos son tales que si pasa uno no pasa el otro y al revés, entonces la posibilidad de que pase alguno es la suma de las dos posibilidades.

4.1. Ejemplo: dos dados. Tomemos dos dados hexaédricos (normales) y pongámonos en el experimento "tirar los dados y sumar los valores". El conjunto X de "lo que puede pasar" es el conjunto de números del 2 al 12 (el 1 no puede salir). La tabla de todos los posibles sucesos, si la entrada superior es un dado y la izquierda el otro es la siguiente:

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Razonamiento erróneo: ¿Cuál es la probabilidad de que salga un número par? Pues *los casos favorables partido por los casos posibles*, claro, esto es la vida real. Los casos favorables son $\{2, 4, 6, 8, 10, 12\}$, es decir 6, mientras que los casos posibles son $12-2+1 = 11$. Por tanto, la probabilidad que hay de que salga un número par es $6/11$. ¿Y de que salga uno impar? Pues $5/11$ (los restantes casos). Es mejor apostar a par que a impar, como todo el mundo sabe al tirar dos dados. **Fin del razonamiento erróneo**

¿Por qué lo anterior está mal? Porque, a la vista de la tabla (que es la descripción precisa de la realidad), uno capta que la probabilidad de que salga un 2 es muchísimo menor que la de que salga un 6 y todavía menor que la de que salga un 7 (que es la más alta). Hay que distribuir las probabilidades según la tabla y, por tanto, de la siguiente manera:

2→1/36	7→6/36
3→2/36	8→5/36
4→3/36	9→4/36
5→4/36	10→3/36
6→5/36	11→2/36
	12→1/36

Como cada "suceso" individual (se denominan *sucesos puntuales*) es disjunto con cualquier otro, resulta que el evento "sale par" tiene como probabilidad la suma de todos los eventos relativos a números pares:

$$p(\text{par}) = \frac{1+3+5+5+3+1}{36} = \frac{18}{36} = 0.5$$

mientras que los eventos relativos a los impares suman:

$$p(\text{impar}) = \frac{2+4+6+4+2}{36} = \frac{18}{36} = 0.5$$

es decir, hay las mismas probabilidades de que salga un número par que de que salga un número impar.

Por otro lado, ¿cuál es la probabilidad de que salga un múltiplo de cuatro ó un múltiplo de tres? Tenemos en este caso dos conjuntos:

- **múltiplos de cuatro:** son los números $4, 8, 12$. A cada uno le corresponden probabilidades $3/36, 5/36, 1/36$, así que la probabilidad de que salga un múltiplo de cuatro es es $(3+5+1)/36 = 1/4$.
- **múltiplos de tres:** son $3, 6, 9, 12$, con probabilidades respectivas de $2/36, 5/36,\ 4/36, 1/36$, así que saldrá un múltiplo de tres con una probabilidad de $12/36 = 1/3$.

¿Se puede decir que la probabilidad de que salga múltiplo de cuatro ó múltiplo de tres es $1/4 + 1/3 = 7/12$ (es decir, algo más que la mitad? No. Porque estamos contando el suceso "sale 12" dos veces:

$$\text{múltiplo de } 4 = \{4, 8, 12\}\,,\ \text{múltiplo de } 3 = \{3, 6, 9, 12\}$$

El 12 está en ambos. De alguna manera hemos de quitarlo. Utilizando el evento "intersección": ¿cuándo es un número a la vez múltiplo de 3 y de 4? En nuestro caso, cuando vale 12. Así pues, para calcular la probabilidad del evento "múltiplo de 4 ó de 3", hacemos:

$$p(\text{múltiplo de 3 ó 4}) = p(\text{múlt. de 4}) + p(\text{múlt. de 3}) - p(\text{múlt. de 4 y de 3})$$

que es:

$$p(\text{múlt. de 3 ó 4}) = \frac{1}{4} + \frac{1}{3} - \frac{1}{36} = \frac{5}{9}$$

que es algo más que la mitad, pero menos que 7/12.

Esto es general. Dados dos eventos o sucesos, dos subconjuntos A y B del espacio X, se tiene que

$$p(A \cup B) = p(A) + p(B) - p(A \cap B).$$

Si los sucesos son disjuntos, es decir, si no tienen intersección, ocurre que $p(A \cap B) = 0$ (pues la probabilidad de que no pase nada es 0, como hemos dicho), y en ese caso, y solo en ese caso es $p(A \cup B) = p(A) + p(B)$.

5. Funciones de distribución habituales

Pasamos rápidamente —pues nuestro interés no es en absoluto la teoría de probabilidades sino la distribución de probabilidades— a estudiar las funciones de distribución más habituales.

5.1. Binomial. Antes de entrar en los casos de funciones de distribución sobre $\mathbb{R}$, vamos a explicar la más sencilla de todas: la distribución binomial, que corresponde a una serie de experimentos de Bernouilli independientes. No hace falta saber qué significa esto porque lo vamos a decir a continuación.

Definición 5.1. Un *experimento de Bernouilli* es un experimento en que se pueden obtener dos valores y solo dos valores.

Por ejemplo, al tirar una moneda puede salir cara o cruz. Un test de la gripe puede dar positivo o negativo. Un alumno puede sacar más de 5 o menos de 5. Una bombilla puede encenderse o no al apretar el interruptor. Todo esto son experimentos de Bernouilli. Está claro que *la probabilidad de que ocurra una cosa u otra* no es la misma en todos los casos. Uno espera, por ejemplo, que se equiprobable sacar cara que cruz, pero no espera, por ejemplo, que el test de la gripe de positivo en más del 1 por ciento de la población (si no, estaríamos buenos). Lo mismo con las bombillas: uno no espera que una bombilla se encienda con un 50% de posibilidades. Para entender esto, es siempre mejor pensar en *qué apostaría uno*, *dónde pondría el dinero*: ¿en que se funde o en que no se funde? (suponiendo que es un experimento no sesgado, claro).

Entre los dos sucesos, uno se toma como "éxito" y el otro como "fracaso" (aunque no tienen por qué ser buenos y malos, es para diferenciarlos). La

probabilidad de éxito se denomina p y la de fracaso (que es obviamente $1-p$) se denomina q.

Pero los experimentos de Bernouilli se suelen hacer en sucesiones: se tira una moneda varias veces, se toman varias muestras de personas, se prueban varias bombillas... Si estos experimentos son *independientes*, se dice que se tiene una *sucesión de Bernouilli independiente* (una colección de experimentos de Bernouilli independientes dos a dos, para ser precisos).

Supongamos que se realiza n veces un experimento de Bernouilli, de modo que son todos independientes (monedas, p.ej.). ¿Cuál es la probabilidad de que salgan exactamente k caras?

Pues, si uno lo mira con cuidado, el espacio muestral (el espacio de probabilidad) es el conjunto de las sucesiones de n elementos donde cada elemento puede ser "cara" ó "cruz". Es como el conjunto de números en base dos que tienen longitud n: hay 2^n elementos.

x_1	x_2	x_3	...	x_{n-1}	x_n

donde x_i es "cara" ó "cruz". La posibilidad de que entre estos n haya *exactamente* k caras está descrita por una función que se llama *función de densidad de Bernouilli*, o *función de distribución binomial* (binomial por las dos posibilidades). De hecho, conocemos el valor porque viene dado por el binomio de Newton: si p es la probabilidad de éxito (salir cara, por ejemplo), entonces $q = 1 - p$ y

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

Donde $K = k$ significa "ocurre *exactamente* k veces el suceso exitoso". El par de números entre paréntesis es el número binomial:

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

que es *el número de combinaciones de n elementos tomados de k en k*.

Pero por lo general lo que a uno no le interesa es la probabilidad de que *ocurra exactamente* un suceso puntual (pues, como se ve, cada suceso puntual es altamente improbable), sino la probabilidad bien de que ocurran al menos k sucesos, bien de que ocurran k o menos sucesos.

En el caso binomial, por ejemplo, la probabilidad de que ocurran k o menos sucesos:

$$p(K \leq k)$$

es la suma de todas las probabilidades de los sucesos anteriores:

$$p(K \leq k) = \binom{n}{0} p^0 (1-p)^n + \binom{n}{1} p(1-p)^{n-1} + \cdots + \binom{n}{k} p^k (1-p)^{n-k},$$

que es más difícil de calcular, pero más útil (imaginaos una apuesta: alguien estaría dispuesto a apostar que salen 23 o menos caras al tirar 50 veces la misma moneda, pero *nadie* apostaría a que salen 23 exactamente, o por lo mismo, 25 exactamente.

La probabilidad de que salgan k o menos caras es la suma de todas las probabilidades de que salgan exactamente i para $i \leq k$, pues todos estos eventos son independientes.

Si por otro lado, lo que se quiere saber es la probabilidad de que salgan *entre k y l caras*, con $l \geq k$, ¿qué podemos hacer? Utilizar el hecho de que

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

donde A es "salen l o menos" y B es "salen más de k". Para empezar, la probabilidad de que salgan *más de k* caras es:

$$p(K > k) = 1 - p(K \leq k)$$

pues la probabilidad del "suceso complementario" es uno menos la probabilidad de un suceso. Ahora, como $l > k$, resulta que $A \cup B = X$ (donde X es el espacio entero). Por tanto,

$$p(A \cup B) = p(X) = 1 = p(A) + p(B) - p(A \cap B),$$

por la explicación que dimos arriba. Pero precisamente $A \cap B$ es el suceso "salen al menos k y no salen menos de l", es decir, salen *entre k y l*: lo que buscamos:

$$p(k \leq K \leq l) = 1 - p(K < k) - p(K < l)$$

y ambas las sabemos calcular.

La figura 8 representa dos funciones: una es la distribución binomial para $p = 0.3$ y $n = 50$, es decir, para cada valor de la x, la probabilidad de que al tirar cincuenta veces una moneda "cargada" de manera que la probabilidad de cara sea 0.3 salgan *exactamente* x caras (como se ve, lo más probable es que salgan unas 50/3 17 caras, pero la probabilidad es poco más del 10%. La otra (en azul, como una curva, pero es igual que antes un valor para cada número entero) indica la probabilidad de que *salgan al menos* x caras al tirar cincuenta veces esa moneda "cargada". Como se ve, aproximadamente en 17 se alcanza la mitad de la probabilidad, como es lógico. La curva punteada se llama *distribución binomial* de parámetros $n = 50, p = 0.3$. La n indica "el número de intentos", la p la probabilidad de *éxito*. La curva continua se llama función de masa de la distribución binomial. Para otros n y p se obtienen curvas distintas.

Es claro (vista la definición) que la función de masa en x no es más que "el área que está debajo de la función de distribución desde 0 hasta x". Esto es lo que mide la función de masa.

Por ejemplo, si queremos saber la probabilidad de que en una población de 180 estudiantes haya entre 50 y 60 que saben inglés si la probabilidad de saber inglés es del 37% hemos de usar:

- Como $p = 0.37$.
- La distribución binomial de $n = 180$, y $p = 0.37$.

La probabilidad de que haya menos de 50 es exactamente el valor de la función de *masa* en 50, que es 0.00572+. La probabilidad de que haya

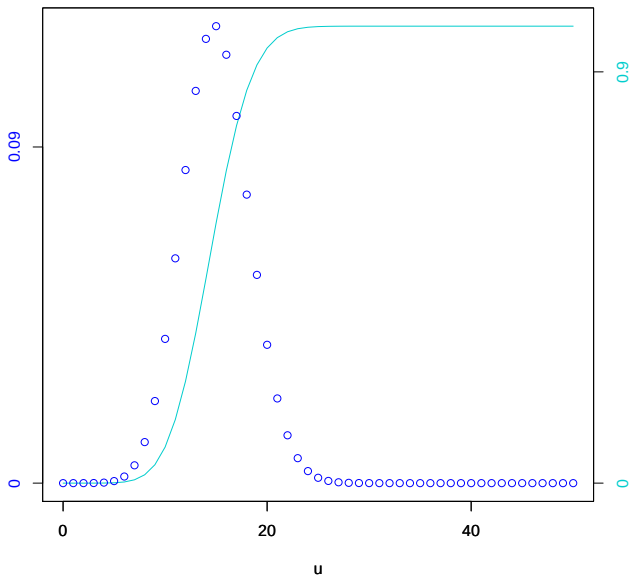


FIGURA 8. Distribución binomial y distribución de masas asociada

menos de 60 es el valor de la función de masa en 60, que es 0.17336+. La diferencia es exactamente el valor que estamos calculando: 0.16763+.

6. Brevemente: funciones de distribución

En general, los experimentos de medición (medida de longitudes, temperaturas, masas, volúmenes,...) se realizan utilizando como "espacio muestral" no un conjunto finito (como para los experimentos de Bernouilli), sino que se trabaja directamente con los números reales. El ejemplo típico es la medida de un error de fabricación en la longitud de una pieza. El error puede ser positivo o negativo, y es preferible pensar que "puede ser cualquier número real" a poner cotas artificiales (que luego se muestran irreales).

Es decir, para modelar la "probabilidad de que un experimento de medición de un resultado" se utiliza como espacio muestral usualmente $\mathbb{R}$ ó bien, si el experimento solo puede dar valores positivos, $\mathbb{R}_{\geq 0}$ (los números reales positivos).

La manera de *distribuir la probabilidad* es lo que se llama "función de distribución". Una probabilidad es una manera de asignar *pesos* a cada suceso. El problema es que, en el caso de las mediciones, cada *suceso puntual* tiene "probabilidad cero" (piénsese en la probabilidad de que un error sea exactamente 0.122881883 milímetros). Lo que sí tiene sentido es pensar en la "probabilidad de que la medición dé entre a y b", ó la "probabilidad de que la medición dé menos que a", ó sucesos similares. Por eso, en lugar de

asignar una *probabilidad* a cada número real, se le asigna una *densidad* de probabilidad:

DEFINICIÓN 6.1. Una *función de densidad o de distribución* de $\mathbb{R}$ en $\mathbb{R}$ es una función $f : \mathbb{R} \to \mathbb{R}$ tal que

- Es no negativa: $f(x) \geq 0$ para todo $x \in \mathbb{R}$.
- Su integral es 1:
$$\int_{-\infty}^{\infty} f(t)dt = 1.$$

La segunda condición *significa* que "la probabilidad de que pase algo" (es decir, de que el experimento de como valor un número) es 1.

Como se ha dicho, la función de densidad asigna a cada número real una *densidad*, de manera que la *probabilidad* de que ocurra un suceso es *la suma de las densidad por los diferenciales de longitud*: la probabilidad se entiende como "un peso", que es comparado con el "peso total", que es 1 (de ahí la condición sobre la integral).

DEFINICIÓN 6.2. Dada una función de distribución f, la *función de masa* asociada es la función $F : \mathbb{R} \to \mathbb{R}$ definida como sigue:
$$F(x) = \int_{-\infty}^{\infty} f(t)dt.$$

Que representa "el área debajo de f a la izquierda de x". Si se interpreta f como una *densidad*, entonces $F(X)$ es la *masa* de la parte a la izquierda de x del eje horizontal.

De la función de masa se puede obtener ahora la probabilidad de un suceso:

DEFINICIÓN 6.3. Si f es una función de distribución y F es su función de masa asociada, entonce se dice que $F(x)$ es la *probabilidad de que el experimento correspondiente dé menor o igual a* x, y se escribe $P(X \leq x)$, donde X es "la medida resultado del experimento".

(Toda esta discusión es vaga e imprecisa de intento, no queremos sobrecargar el lenguaje).

Como la probabilidad, como hemos dicho, de los sucesos puntuales es 0, $P(X = x) = 0$, resulta que:

PROPOSICIÓN 6.4. *La probabilidad de que un experimento que sigue una función de distribución f dé un resultado entre a y b es $F(b) - F(a)$, es decir:*
$$P(a \leq X \leq b) = \int_a^b f(t)dt = F(b) - F(a).$$
Por otro lado, la probabilidad de que dé mayor o igual *que b es $1 - F(b)$:*
$$P(X \geq b) = 1 - P(x \leq b) = 1 - \int_{-\infty}^{b} f(t)dt = 1 - F(b).$$

FIGURA 9. Función de distribución y de masa asociada.

7. Distribuciones comunes

7.1. La distribución normal. El movimiento browniano es el movimiento de partículas en un fluido, que surge por la propia naturaleza del fluido. Es un movimiento aleatorio: las partículas reciben "empujones" de manera equiprobable en cada dirección y se desplazan de acuerdo con esos empujones. Uno de los problemas fundamentales de la física resueltos por Einstein es la descripción matemática del movimiento browniano.

Este movimiento conviene pensarlo primero como si ocurriera en una sola dimensión (empujones hacia la derecha o hacia la izquierda). En cada instante, una partícula se desplazará una cierta distancia, en una de las dos direcciones. Pues bien, cualquier descripción que se haga medianamente razonable de este movimiento lleva a la conclusión de que *los desplazamientos en cada instante* siguen una distribución específica. Es decir, que la probabilidad de que una partícula se desplaze x o menos en cada instante viene dada por una función *universal*: la curva normal de Gauss.

Por lo general, cualquier cantidad *continua* (no discreta) de la naturaleza que dependa de procesos aleatorios (altura de una especie, longitud del tallo de una especie de planta...) sigue esta distribución (con un par de parámetros que se han de especificar: la media y la desviación típica). Pero veamos la curva:

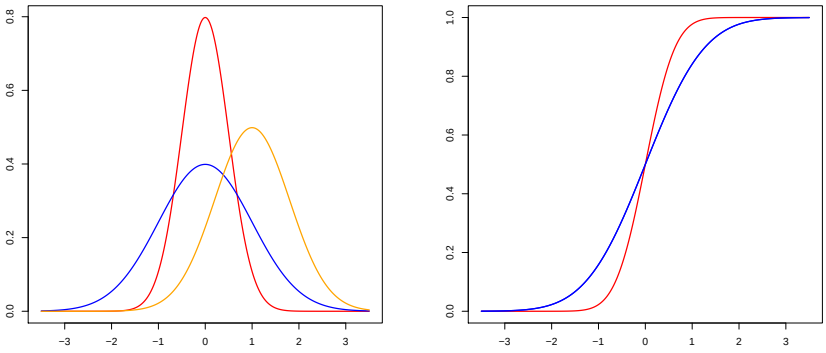


FIGURA 10. Curvas normales de varias medias y desviaciones típicas y sus funciones de masa

En la figura 10 se ven, a la izquierda, tres ejemplos de curvas de distribución normal (campanas de Gauss), con los siguientes parámetros:

- Una con media cero y desviación típica 0.4: la más alta. La distribución está muy centrada en la media porque la desviación típica es muy pequeña (y por tanto la media es muy relevante).

- Una con media cero y desviación típica 1: la centrada pero más achatada: se observa que la media es menos relevante que en la anterior. *Esta es* la normal "estandarizada".
- Una con media 0.7 y desviación típica 1.2: todavía más achatada. No está centrada en 0 sino en 0.7 (claro, está centrada respecto de la media).

En la misma figura, a la izquierda, se ven las funciones de masa correspondientes.

Dada una distribución normal $N(m, \sigma)$, con media m y desviación típica σ, se tiene que: la probabilidad de que un "experimento" que sigue esa distribución salga menor o igual que un número l es precisamente el área que hay debajo de la curva hasta l, que es precisamente *el valor de la función de masa* en l (por ejemplo, la probabilidad de que un experimento que sigue una normal $N(0.7, 1.2)$ obtenga un resultado menor que 0 es 0.27 aproximadamente, es 0.5 para un resultado menor o igual que 0.7 y es aproximadamente 0.69 para un resultado menor o igual que 1.3 (como se ve, está muy achatada, la curva). Por el contrario, la normal $N(0, 0.4)$ en seguida llega a masa próxima a 1.

En la figura 11 Se muestra el área que hay bajo la curva normal $N(0, 1)$ para $x < -0.4$, aproximadamente 0.344 (sobre un máximo de 1).

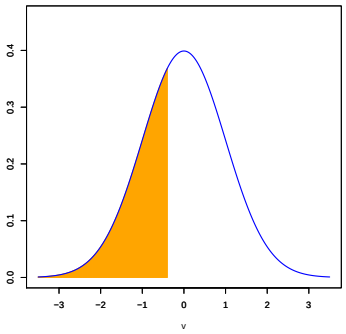


FIGURA 11. Probabilidad de $k \leq -0.4$ para la distribución $N(0, 1)$

En la figura 12 se muestra el área que cae debajo de la $N(0.7, 1.2)$ para $x \geq 2.17$, aproximadamente 0.11.

Finalmente, se muestra en la figura 13 la probabilidad de que x esté entre -0.2 y 0.45 para la normal $N(0, 0.4)$, que es más de la mitad, 0.532.

7.1.1. *Estandarización. La z.* Todas las distribuciones normales pueden "reconvertirse" en una $N(0, 1)$. Esto significa que para calcular la probabilidad de que un evento que sigue una distribución normal $N(\mu, \sigma)$ "caiga" en una zona de la distribución, p.ej. $x \leq -3.2$, puedo "recalcular un valor", que se llama z a partir de esa x y utilizar *el área por debajo de* la $N(0, 1)$. En general, la fórmula es, si X es un evento de una distribución normal

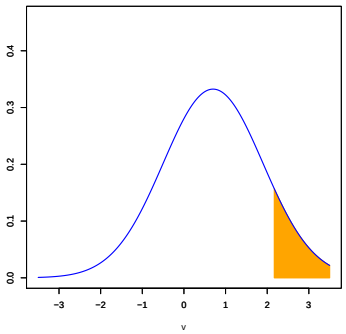


FIGURA 12. Probabilidad de que $k \geq 2.17$ para la distribución $N(0.7, 1.2)$

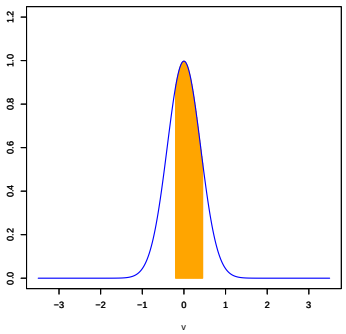


FIGURA 13. Probabilidad de que $-0.2 \geq x \geq 0.4$ para $N(0, 0.4)$

$N(\mu, \sigma)$, entonces

$$P(X \leq x) = P(Z \leq z), \text{ con } z = \frac{x - \mu}{\sigma} \tag{1}$$

y Z sigue ahora una distribución normal "estándar" $N(0, 1)$.

Hoy en día esto solo se hace "teóricamente", nadie hace la cuenta en realidad.

Lo mismo para eventos del tipo X *está entre* a *y* b. Se calculan la z_a y la z_b y el área por debajo de la $N(0, 1)$ entre z_a y z_b.

7.2. La distribución χ^2. Con frecuencia, más que estudiar un solo caso de un evento y su probabilidad, hace falta estudiar la probabilidad de que la variación cuadrática (*ojo: no la varianza*) de una muestra pueda ser grande comparada con la media. Veremos ejemplos más adelante.

Supongamos que tenemos una muestra $x_1, \ldots, x_n$ y sabemos que cada elemento (cada experimento) sigue una distribución normal. Por ejemplo: se toman n tornillos (mismo modelo) de una fábrica, y se supone que la

distribución de las longitudes sigue una curva normal de media la "altura especificada" μ y varianza σ^2. Se mide la "variación cuadrática muestral": $(x_1-\mu)^2+\cdots+(x_n-\mu)^2$ y se compara (es decir, se divide) entre la media:

$$\frac{(x_1-\mu)^2+\cdots+(x_n-\mu)^2}{\mu}.$$

La pregunta es: ¿cuál es la probabilidad de que esta variación cuadrática tome cierto valor? En otro lenguaje, ¿cuál es la distribución de probabilidad asociada a este experimento (tomar n muestras de una variable "normal" y medir su "variación cuadrática"). La respuesta es: la χ^2 con n grados de libertad.

En concreto:

DEFINICIÓN 7.1. Dados n sucesos independientes que tienen función de distribución normal de varianza 1 (lo que se ha llamado $N(0,1)$), la probabilidad de que la variación cuadrática de esos n sucesos (es decir, la suma $x_1^2+\cdots+x_n^2$) sea menor o igual que $r\in\mathbb{R}_{>0}$) viene dada por el área que hay debajo de la función χ^2_n (la función de distribución χ^2 con n grados de libertad).

Así pues, para la χ^2 no se especifica ni media ni varianza, sino los grados de libertad. En la figura 14 se muestran las gráficas de las distribuciones χ^2_k para $k=1,2,3,4,6$ y 10.

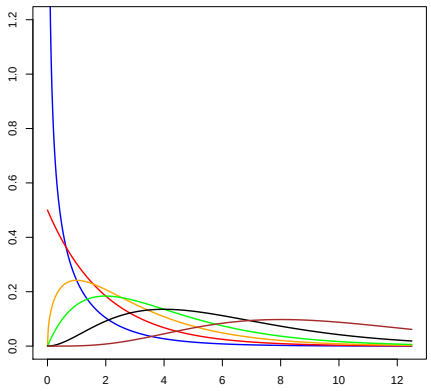


FIGURA 14. Distribución χ^2_k para $k=1,2,3,4,6$ y 10

Como se ve, la distribución "se va haciendo más plana" y adquiere una "joroba" a partir de los 3 grados de libertad. El aplanamiento es lógico: si sumamnos 10 números positivos, es más probable obtener un número grande cercano a 10 que un número pequeño (recuérdese que la probabilidad de que "salga" x o menos es el área por debajo de la curva).

La χ^2 está definida claramento solo para $x > 0$, pues es la probabilidad de que una suma de cuadrados tome cierto valor.

7.3. La t de Student. Muy sucintamente, pasamos por encima (literalmente) de la t de Student. Corresponde al siguiente experimento:

De una población que está distribuida según una distribución normal $N(\mu, \sigma)$ se toma una muestra de tamaño n. Llamemos $\overline{x}$ a la media muestral y s a la desviación estándard (desviación típica) muestral:

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

(nótese que *no es exactamente la desviación estándard*). Con esos datos, se obtiene un valor:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

(una especie de "variación de Fisher media", por decir algo). Pues bien, este "experimento" de tomar n muestras y calcular esa t tiene una probabilidad de que salga un valor u otro, como todo experimento.

Definición 7.2. La t de Student con n grados de libertad es la función de distribución asociada al experimento descrito (tomar n valores de la dicha población y calcular su t).

Así que la probabilidad de que t valga x ó menos es el área por debajo de la curva *hasta* x.

Como se ve, para cada número de grados de libertad hay una t. En la figura 15 se muestran las t para los grados de libertad $1, 2, 3$ y 7: según n va creciendo, la "joroba" se hace cada vez más empinada (hasta que cuando $n \to \infty$, la curva es lo mismo que una normal, que se muestra en negro y más fina).

La t se utiliza, entre otras cosas, para verificar la plausibilidad de ciertas hipótesis, como veremos más adelante (en esto es análoga a la χ^2): es más útil como instrumento de verificación que como distribución "de sucesos reales".

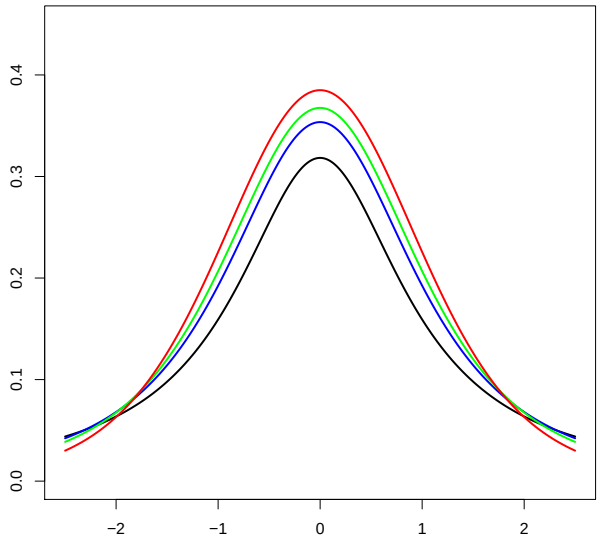


FIGURA 15. Gráficas de la t de Student para $n = 1, 2, 3, 7$ y la $N(0, 1)$ en negro. El vértice se hace mayor al crecer n.

La significatividad estadística: introducción

8. Introducción

La correlación es un dato interesante, pero la pregunta clave (aparte de la ya explicada de que correlación *no equivale* a causalidad), es ¿cuánto de relevante es?

Una de las grandes aplicaciones de la estadística es precisamente el discernimiento de lo posible que es que un suceso sea o no aleatorio.

Ejemplo 1. Tomemos el experimento de tirar una moneda al aire cien veces. Todo el mundo estará de acuerdo en que si en ese experimento salen 85 caras, hay bastante lugar para pensar que la moneda está trucada. Mucho más si salen solo 3 caras. Pero, ¿qué ocurre si salen 42 caras? ó ¿si salen 60? ?Tenemos en los dos últimos casos razones para pensar lo mismo? ¿O puede ser fruto del azar?

Esa y exactamente esa pregunta "¿puede ser lo que he obtenido en un experimento fruto del azar?", o más bien "¿cómo de probable es que lo que he obtenido en un experimento sea fruto del azar?" es lo que estudia la *significatividad estadística*.

Ejemplo 2. Sigamos con el ejemplo de la tirada de una moneda. Vamos a calcular cuál es la probabilidad de que salgan 79 caras ó más. Como ya vimos, la distribución que sigue la tirada una moneda 100 veces es una binomial y si suponemos (como hemos de suponer) que la moneda es justa, la p es 0.5 (y $q = 1 - p = 0.5$ también). La gráfica de la *distribución de masa* de la binomial se muestra en la figura 16 Para cada punto x, el valor de la función $P(x)$ es exactamente la probabilidad de que salgan *como mucho* x caras. Se aprecia que ya cerca de 60 se alcanza una probabilidad de casi uno. De hecho, $P(58) = 0.955+$ y $P(59) = 0.971+$, así que hay menos de un 5% de probabilidades de que salgan 59 caras ó más. Más aun, $P(79) = 0.99999999944+$ (es decir, hay una probabilidad contra mil millones de que salgan al menos 79 caras), lo que nos da una idea: si en un experimento aparecen 79 caras, haremos bien en pensar que la moneda no está *tan* equilibrada.

¿Dónde se pone el límite? Por razones de uso y de hábito, el límite para la *significatividad estadística* se pone en el 5%. Es decir, si se describe un modelo y ocurre un suceso que tiene menos del 5% de probabilidades de ocurrir, entonces se tiende a pensar que el modelo es erróneo. Este es el

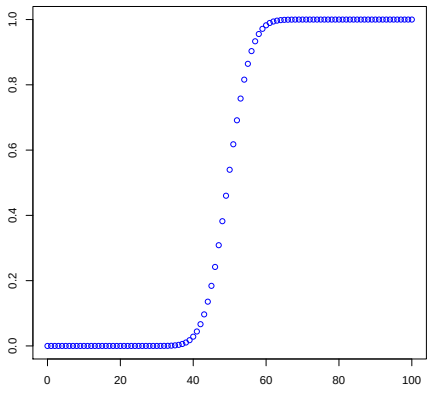


FIGURA 16. Función de masa de la distribución binomial $N = 100$, $p = 0.5$.

valor "mágico" que se utiliza en el contraste de hipótesis. Se podría utilizar otro, pero es este con el que se trabaja.

Pasa lo mismo con la correlación. El hecho de que alguien tire un par de dados y obtenga un doble "no es nada raro". Si eso ocurre muchas veces, "hay que pensar que algo pasa" (hay demasiada correlación entre los dados). ¿Cuántas veces? Ese es el problema. Depende de la correlación.

9. El test de la correlación de Pearson

Supongamos que tenemos una lista de pares de datos (x_i, y_i) de dos variables, X e Y y calculamos la correlación de la muestra, r:

$$r = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

que suponemos que es no nula.

El problema consiste en saber si la correlación es o no relevante (es decir, esa correlación que aparece en la muestra, ¿es razonable pensar que ocurre en la realidad o no?). En la realidad, las variables X e Y estarán o no correladas (es precisamente lo que no sabemos). Su correlación real será un número ρ, desconocido.

Los pasos para saber si es razonable suponer que r es relevante son los siguientes (se supone que ya se ha calculado r):

(1) Se enuncia la "hipótesis nula". Que consiste en asumir que lo que hemos encontrado en la muestra ha sido pura casualidad. Es decir, que la correlación real es $\rho = 0$ y la r que nos ha salido es por "azar". Todo el proceso consiste en saber si se puede o no desechar esta hipótesis razonablemente. La "hipótesis alternativa" es, en este caso: "ocurre que las variables X e Y sí están correladas".

(2) Se elige un nivel de significatividad, que *siempre* es 0.05 o el 5% (es lo mismo).
(3) Se calcula la probabilidad de obtener el resultado r "por puro azar". Es decir, si fuera cierta la hipótesis nula, ¿cuál es la probabilidad de que al tomar una muestra del mismo tamaño aparezca una correlación como r? Este es el paso más delicado, para el que antiguamente se usaban tablas. Ahora se utiliza un ordenador. Aquí se cuentan los *grados de libertad* como $N-2$, si la muestra tiene tamaño N.
(4) Se compara la probabilidad del paso anterior, p, con el nivel elegido de significatividad ($s = 0.05$). Si $p < s$ entonces *es razonable desechar la hipótesis nula.* Si no, no.
(5) Se enuncia el resultado (con precisión).

El tercer paso es el más elaborado, porque hay que convertir la r calculada en un número (denominado t):

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}.$$

Ahora se mira la tabla de la distribución *de masa* de la t de Student de $N-2$ grados de libertad, para el valor t *en las dos direcciones* y se obtiene un valor.

NOTA 9.1. *En las dos direcciones.* Estamos haciendo un ejemplo de mero estudio de si una correlación es o no relevante. Por eso, como no estamos mirando si el signo ha de ser positivo o negativo, hay que asumir que podría ser cualquiera. Por eso "las dos direcciones". Hablaremos más sobre esto, pero no hay que pensar en nada raro, simplemente compara los siguientes eventos:

(1) Tirar cien veces una moneda y que salgan al menos sesenta caras.
(2) Tirar cien veces una moneda y que salgan al menos sesenta del mismo tipo (sesenta caras o sesenta cruces).

Es más probable (de hecho, el doble de probable) el segundo que el primero. Por eso en el caso de arriba se toma el valor "bidireccional". Hay experimentos en que queremos saber si lo razonable es pensar si la correlación es positiva o negativa, y ahí hay que afinar más.

Hablaremos más adelante en detalle de la hipótesis nula y la direccionalidad.

Ejemplo 3. Para los ejemplos del Ozono, el viento y la radiación solar, habría que llevar a cabo las siguentes secuencias de pasos. Antes de nada, téngase en cuenta que la muestra tiene 111 datos:

Ozono-Viento: Se comienza enunciando la hipótesis nula: *la concentración de Ozono y la velocidad del viento no están* ***realmente*** *correladas, es decir* $\rho = 0$. Ahora
(1) Se calcula $r = -0.612+$.

(2) Se estipula el nivel de significatividad. El 0.05 (el usual).
(3) Se calcula la probabilidad de obtener una correlación $r = -0.612+$ por puro azar asumiendo que realmente la correlación es $\rho = 0$. Para ello:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{-0.612\sqrt{109}}{\sqrt{0.625}} = -8.082$$

La probabilidad de obtener $-|t|$ ó menos o bien $|t|$ ó más es el área que hay a la izquierda de -8.082 de una t de Student con 109 grados de libertad **más** la que hay a la derecha de 8.082, que es (tablas, ordenador, etc...) $p = 9.09 \times 10^{-13}$ (es decir, un cero seguido de unos 12 ceros y un nueve).
(4) Así que $p < 0.05$ y por tanto, *se puede rechazar razonablemente la hipótesis nula.*
(5) Se enuncia el hecho:

> Hay una relación estadísticamente significativa entre la velocidad del viento y la concentración de ozono (en NY, en tal año, etc...) *en los datos de la muestra*, con $r = -0.612$, $t(109) = -8.082$ y $p = 9.09 \times 10^{-13}$.

Como no se dice nada, se sobreentiende que la prueba es "sin dirección", es decir, para comprobar que la correlación *no es nula.*

Ozono-Sol: Se comienza enunciando la hipótesis nula: *la concentración de Ozono y la radiación solar no están* ***realmente*** *correladas, es decir* $\rho = 0$. Ahora
(1) Se calcula $r = 0.348$.
(2) Se estipula el nivel de significatividad (vamos a poner 0.03 por cambiar).
(3) Se calcula la probabilidad de obtener una correlación $r = 0.348$ por puro azar asumiendo que realmente la correlación es $\rho = 0$. Para ello:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{0.348\sqrt{109}}{\sqrt{0.878}} = 3.877$$

La probabilidad de obtener $-|t|$ ó menos o bien $|t|$ ó más es el área que hay a la izquierda de -3.877 de una t de Student con 109 grados de libertad **más** la que hay a la derecha de 3.877, que es (tablas, ordenador, etc...) $p = 0.00017$.
(4) Se tiene $p < 0.03$ (el nivel de significatividad) y, por tanto, *se puede rechazar razonablemente la hipótesis nula.*
(5) Se enuncia el hecho:

> Hay una relación estadísticamente significativa entre la radiación solar y la concentración de ozono en el ambiente (en NY, en tal año, etc...)

> *en los datos de la muestra*, con $r = 0.348$, $t(109) = 3.877$ y $p = 0.00017$.

Obsérvese como —y esto es importante— el estadístico *no habla de causalidad*, sino de *correlación en la muestra*. La realidad está ahí, pero nosotros hemos utilizado solo los datos. Es trabajo del científico o del experto tratar de desentrañar, si la hay, la relación de causalidad (buscar la explicación correcta para esa correlación significativa).

Ejemplo 4. Si hacemos todos los pasos en el ejemplo de los récord del mundo de 100m y el número de fallecidos en Estados Unidos en cada año, obtenemos $r = -0.953$, $t(36) = -18.9197$ (hay 36 grados de libertad) así que $p = 2.2 \times 10^{-16}$ (muchísimo menor que 0.05). Por tanto, el enunciado correcto, tomando como hipótesis nula que las muertes no están relacionadas con el récord de los 100 metros, es

> Hay una relación estadísticamente significativa entre el récord del mundo de cien metros lisos en un año y el número de fallecidos en los Estados Unidos de América entre 1968 y 2005, con $r = -0.953$, $t(36) = -18.9197$ y $p = 2.2 \times 10^{-16}$.

¿Alguien puede sacar alguna conclusión?

Test de significatividad: la hipótesis nula

10. Esperanza media aleatoria

Todos los tests de significación estadística consisten en una comparación. En concreto, en comparar un valor observado (p.ej. una correlación r) con el valor que uno esperaría encontrar si solo influyera el azar en el experimento (p.ej. una correlación *real* de $\rho = 0$). La tabla 3 muestra más ejemplos.

Qué se compara	Ejemplos
(1) **Un valor observado**	Un coeficiente de correlación observado, el número de caras al tirar N monedas, el número de pacientes que se ha recuperado por un tratamiento, la cantidad de piezas defectuosas en una cadena de montaje...
(2) **El valor que uno esperaría encontrar, si lo único que afecta al experimento es el puro azar**	Un coeficiente de correlación $r = 0$ asumiendo que la correlación en una población de pares X_iY_i es $\rho = 0$; 5 caras al tirar una moneda 10 veces (asumiendo que la moneda es justa); 400 recuperaciones en 1000 pacientes, si se asume que la probabilidad de curación es 4/10...

TABLE 3. Test de significación

Un nombre común en inglés para el segundo dato es *Mean Chance Expectation*, que yo he traducido como "Esperanza media aleatoria", pero que puede que no se llame así. En todo caso, lo vamos a denominar MCE por seguir una notación.

Por ejemplo, para un experimento binomial (tirar una moneda), la esperanza media aleatoria es exactamente la media (si la moneda es justa y se tira 10 veces, la MCE es "sacar 5 caras"). Por supuesto, esto *no tiene nada que ver con que ocurra siempre*: es la media, una manera de hablar del "resultado esperado", pero repito: no tiene nada que ver con que ocurra siempre.

Si la moneda estuviera cargada y la probabilidad de cara fuera 0.3, entonces la MCE del experimento de tirarla 10 veces sería la media, que es 3. De hecho, las posibilidades de que salgan exactamente i caras se resumen en la figura 17

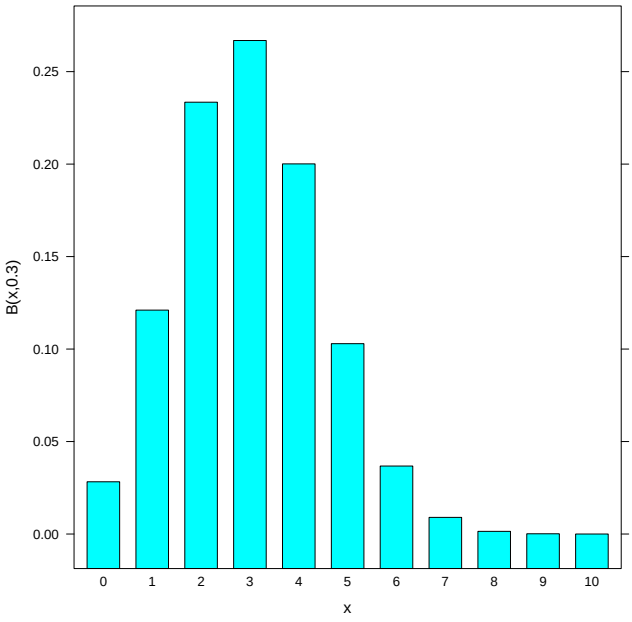


FIGURA 17. Distribución binomial para $N = 10$ y $p = 0.3$

11. La hipótesis nula y la hipótesis de trabajo

En este apartado (y solo en este, para enfatizar), vamos a utilizar en lugar de los símbolos $=$, $>$, $<$, $\geq$ y $\leq$, la siguiente colección de símbolos:

$a \simeq b$	a y b no difieren significativamente, es decir son iguales salvo diferencias aleatorias normales.
$a \not\simeq b$	a y b son significativamente diferentes, es decir, hay (o esperamos que haya) una diferencia que no se explica sin más por sucesos aleatorios.
$a \succ b$	a es significativamente mayor que b (más allá de lo explicable por puro azar).
$a \prec b$	a es significativamente menor que b (más allá de lo explicable por puro azar).

TABLE 4. Signos de relación estadística

Al realizar un test de significatividad estadística conviene distinguir claramente entre

i) La hiótesis específica que se está examinando: por ejemplo, que una máquina produce defectos con una varianza diferente que la especificada. Esto sería $a \not\simeq b$, donde a es la varianza de la máquina real y b la especificada.
ii) El enunciado lógicamente antitético: en el mismo ejemplo, que la varianza de los defectos es exactamente la especificada, es decir $a \simeq b$.

La segunda hipótesis es lo que se denomina la *hipótesis nula*. Nula en el sentido de "cero", y se denota siempre H_0. La hipótesis de trabajo se denomina *hipótesis alternativa* y se llama H_1 habitualmente. Por lo general, la hipótesis nula especifica que "lo observado, sea lo que sea, es fruto del azar", es decir, que "lo observado es igual al MCE, dentro de los límites naturales del azar". Por ejemplo,

$$H_0\text{: El valor observado} \simeq \text{MCE},$$

mientras que la hipótesis de trabajo especifica lo contrario: hay una diferencia significativa entre el resultado observado y la MCE:

$$H_1\text{: El valor observado} \not\simeq \text{MCE}.$$

Ahora bien, puede ocurrir que uno tenga una hipótesis un poco más específica. En el caso descrito arriba la hipótesis H_1 no hace referencia a si el valor observado es mayor (signficativamente) que la MCE ó menor: cualquiera de las dos posibilidades satisface la hipótesis.

Sin embargo, en algunos experimentos uno podría tratar de comprobar que el valor observado es $\succ$ MCE o bien $\prec$ MCE. Esto se denominan *hipótesis direccionales*.

Está claro que la hipótesis *no direccional* es una mera unión de dos hipótesis direccionales (un número real es distinto de otro si o bien es mayor o bien es menor, esto es lo que queremos decir en todo este texto).

La distinción es relevante, aunque a la hora de la práctica, está claro. El asunto es que la significatividad estadística *es diferente* para unas hipótesis que para otras.

11.1. Hipótesis direccionales y no direccionales y tests de significatividad. Supongamos que alguien aparece y dice:

$$H_1 : \text{Esta moneda está trucada y sale más una cosa que otra.}$$

y otra persona dice

$$H_1' : \text{Esta moneda está trucada y salen más caras que cruces.}$$

A la hora de hacer un test de significatividad estadística, ¿cuántas tiradas harán falta para rechazar la hipótesis H_1 y cuántas para rechazar la hipótesis H_1'? Es decir, ¿qué es más fácil que ocurra por azar, la H_1 o la H_1'?

Está claro que es mucho más fácil que ocurra por azar la H_1, pues como no especifica (es *no direccional*), tanto si salen muchas caras *como* si salen muchas cruces, se cumple; mientras que la H_1' solo se cumple en el caso de las caras.

Esto mismo es lo que pasa con los tests direccionales y no direccionales. A la hora de hacer un estudio de significatividad estadística, si la hipótesis de trabajo es *no direccional*, hace falta siempre *mirar los dos lados de la función de distribución correspondiente*, mientras que si la hipótesis de trabajo es *direccional*, solo hace falta mirar el lado correspondiente a la dirección.

Por eso, en todos los tests que hagamos a partir de ahora, la direccionalidad jugará un papel esencial: si el test es no direccional y tenemos pocos experimentos (datos), será difícil obtener una significatividad estadística relevante. Si el test es direccional, harán falta menos ensayos.

Pero esto es mejor verlo con ejemplos.

Test χ^2 para análisis de frecuencias en datos cualitativos

Del mismo modo que uno puede comprobar cuántas tiradas hay que realizar para asegurarse (con una probabilidad del 95% de que una moneda está cargada[2]) utilizando la distribución binomial, se utiliza la distribución χ^2 para estudiar la significatividad de un experimento en que los datos se agrupan cualitativamente en más de dos categorías.

También se utilizan, como veremos en capítulos posteriores, para analizar datos con más de dos dimensiones de categorización. Por ejemplo, piezas que pueden ser *fabricadas por X ó Y* y a la vez *utilizadas en máquinas A y B*: tendríamos una tabla con 4 categorías pero mutuamente relacionadas. Lo veremos.

12. El procedimiento χ^2 para una dimensión cualitativa

Como ya se explicó, cada una de las distribuciones χ^2 (hay una para cada número de grados de libertad n) define la "probabilidad de que la varianza de una muestra de tamaño n de una población *normal* de media cero y varianza 1 tome cierto valor". Esto se puede utilizar para comparar la distribución de una muestra en una variable cualitativa con la "distribución teórica".

Ejemplo 5. En una cadena de producción hay tres máquinas distintas de fabricar rodaminetos, de las marcas Siemes, Tekra y Ultimate. Se supone que las tres producen el mismo número de rodamientos defectuosos. Los rodamientos defectuosos se distribuyen todos a un saco, donde se mezclan todos.

Se toma una muestra del saco al azar en un momento dado y se obtiene la siguiente distribución (todos los rodamientos de la muestra son defectuosos):

	Siemens	Tekra	Ultimate
Experimental	84	119	92
Teórico	33%	33%	33%

Table 5. Rodamientos defectuosos

y uno se pregunta ¿qué significan estos datos?

[2]¿Cómo se haría esto?, pues no lo hemos explicado.

Como se ve, la muestra es de $84+119+92 = 295$ rodamientos. Salta a la vista que la máquina Tekra parece que produce más elementos defectuosos que las otras, pero ¿es esto realmente significativo? O quizás es que la Siemens es mucho mejor de lo que nos decían.

Esa es la pregunta relevante. Otra manera de enunciarla (la manera precisa) es ¿es el resultado que hemos obtenido estadísticamente significativo?

Lo que nos estamos preguntando es, en fin, si es razonable suponer que la distribución teórica $33\%, 33\%, 33\%$ es la real, o si es más razonable pensar que es otra.

El procedimiento para responder a esta pregunta es *siempre el mismo*: utilizar la distribución χ^2. Y la manera de utilizarla es siempre la misma: la descrita a continuación.

Antes de empezar se enuncia la hipótesis nula, que en este caso es "la distribución de errores por máquina no difiere significativamente de $1/3, 1/3, 1/3$", mientras que la hipótesis alternativa es "la distribución de errores por máuina difiere significativamente de esa". Como se ve (y esto es inherente a este tipo de problemas), *la hipótesis alternativa es casi seguro no direccional*. Porque en este tipo de tests es difícil (puede ocurrir que sea posible) conseguir una hipótesis direccional.

- Se calcula el número de grados de libertad (df), que es *el número de casillas menos* 1. En este caso $df = 3 - 1 = 2$.
- Se calcula, para la muestra y para cada categoría, la variación cuadrática respecto de la media esperada. Esto es, en cada casilla, se calcula: (*frecuencia encontrada* - *frecuencia esperada*)^2 y se divide por la *frecuencia esperada*:

Siemens	Tekra	Ultimate
2.08	4.34	0.40

- Se suman los valores obtenidos:

$$2.08 + 4.34 + 0.40 = 6.84$$

- Se calcula el valor (en realidad el área por debajo de) de la χ^2 para df grados de libertad en ese número: $p = 0.032$, que es considerablemente menor que 0.05. Si el valor obtenido es menor que 0.05, se puede rechazar razonablemente la hipótesis nula.
- Se enuncia el resultado: con un nivel de significatividad mayor que el 95%, se observa que una muestra se ha desviado de la distribución teórica $1/3\%, 1/3\%, 1/3\%$ (ahora habría que diseñar otro experimento para calcular la nueva distribución).

Nótese (y esto es importante y muestra cómo est test es no direccional) que *Siemens* "influye mucho más que Ultimate" en el valor 6.94 pero es *la que "mejor" se ha comportado de las dos*. Uno tiende a pensar, vistos los números, que "habría que rebajar" el porcentaje atribuido a Siemens y habría que subir el de Tekra, posiblemente. Pero esto de momento no sabemos cómo hacerlo.

13. El test χ^2 para varias dimensiones

Por lo general no se tiene una mera "clasificación en grupos" de unos datos, sino que hay una familia de grupos interconectados y datos del comportamiento de un experimento respecto de dichas interconexiones.

Ejemplo 6. La fábrica de rodamientso *Astur-roda* recibe acero de dos proveedores, *Asturacero* y *Ferroastur*, y lo transforma en rodamientos utilizando tres máquinas, una *Siemens*, una *Tekra* y otra *Ultra*. La dirección supone que el tipo de acero que se use es indiferente a la máquina (es decir, el rendimiento de cada máquina es indiferente al tipo de acero). Un día, durante una auditoría, se toma una muestra de las piezas defectuosas de cada partida, y se obtiene una tabla como la que sigue:

	Siemens	Tekra	Ultra
Asturacero	55	41	37
Ferroastur	70	25	27

TABLE 6. Piezas defectuosas de la fábrica Astur-roda

A la vista de la tabla, parece que se ve que la máquina *Siemens* trabaja mejor con *Asturacero* pero que las otras dos lo hacen mejor con *Ferroastur*. La cuestión es "¿estamos seguros de que hay cierta relación o los resultados pueden deberse a la casualidad?". Esta *y exactamente esta pregunta* es la que responde el test χ^2. En este caso, para lo que se denomina una *tabla de contingencia*.

La manera de llevar a cabo el test en este caso es similar (pero claro, algo distinta) al anterior. Se ha de usar la idea de la "probabilidad producto", que no hemos introducido pero que es natural. Primero de todo, hacen falta los totales de las sumas por filas y por columnas: Así que hay 255 piezas

	Siemens	Tekra	Ultimate	**Total**
Asturacero	55	41	37	**133**
Ferroastur	70	25	27	**122**
Totales	**125**	**66**	**64**	**255**

TABLE 7. Totales para las piezas defectuosas

defectuosas en total, con las distribuciones de la tabla.

Al contrario que en el test para una dimensión, *las probabilidades en las tablas de contingencia no se especifican*, se calculan. En el ejemplo de la sección previa, utilizamos como hipótesis nula que la distribución de las piezas defectuosas era una prevista: $1/3, 1/3, 1/3$. En el caso de tablas de dos o más dimensiones, la distribución se calcula a partir de los totales, siguiendo la regla de probabilidades "casos favorables partido por casos posibles". En el ejemplo: los 255 rodamientos defectuosos se distribuyen como 125/255 de

Siemens, 66/255 de *Tekra* y 64/255 de *Ultimate*. Estos datos son *la primera parte*. Ahora la distribución de todas las probabilidades en todas las casillas:

- Si la distribución fuera realmente aleatoria, del total de 133 rodamientos con acero de asturacero, ocurriría (o uno tendería a pensar) que 125/255 de ellos deberían ser de *Siemens*, otros 66/255 de *Tekra* y 64/255 de *Ultimate*.
- Por lo mismo, del total de 122 rodamientos con acero de Ferroastur, deberían ser 125/255 de *Siemens*, 66/255 de *Tekra* y 64/255 de *Ultimate*.

Es decir, quedaría una distribución de pesos como sigue: que especifica

	Siemens	Tekra	Ultimate	**Total**
Asturacero	65.196	34.423	33.381	**133**
Ferroastur	59.803	31.576	30.621	**122**
Totales	**125**	**66**	**64**	**255**

TABLE 8. Distribución *teórica* de los rodamientos.

cómo sería una distribución *teórica* si los totales por filas y columnas se distribuyeran "homogéneamente" en la tabla de contingencia (si fuera independientes las fila y las columnas, por así decir). Como se ve, hay diferencias importantes en algunas casillas. Lo que hay que estudiar es "cuánto de importantes", es decir, cuánto de *significativa* es la diferencia respecto de lo "teórico". Es aquí donde, una vez más, entra el test χ^2, pues al fin y al cabo vamos a comparar una distribución real con una distribución teórica (cuánto nos separamos de la realidad).

Igual que antes:

- Se enuncia la hipótesis nula: la distribución que hemos observado se ha obtenido por puro azar: no hay una relación significativa entre las filas y las columnas. La hipótesis alternativa es "hay una relación signficativa entre filas y columnas".
- Se establece un nivel de significatividad, que *habitualmente* es 0.05, por debajo del cual se establece que se puede rechazar *razonablemente* la hipótesis nula.
- Se calcula el número de grados de libertad: $df = (f-1)(c-1)$, donde f es el número de filas y c es el número de columnas. En nuestro caso, $df = (2-1) \times (3-1) = 2$.
- Se calcula la variación cuadrática respecto de la media esperada en cada casilla (la *media esperada* es el número que se ha calculado para cada casilla), la variación cuadrática es la diferencia cuadrática entre lo observado O y la media esperada E partido por E. En nuestro caso (aproximadamente):

	Siemens	Tekra	Ultimate	**Total**
Asturacero	1.594	1.256	0.3923538	2.343
Ferroastur	1.738	1.369	0.4281911	3.536
Total	3.333	2.626	0.820	**6.779**

Atención: si la tabla es 2×2 *y solo en ese caso*, la fórmula *no es* $(O-E)^2/E$, sino que se resta 0.5, una *corrección por continuidad*. Haremos un ejemplo.

- El valor total, que se denomina χ^2 de la muestra, 6.779 se utiliza para computar el área que hay debajo de la distribución χ^2 con *df* grados de libertad *desde ese valor* hacia la *derecha*, $1 - 0.9662 = 0.0337$. Se compara este valor con el nivel de significatividad establecido: 0.05. Si es menor, *se rechaza la hipótesis nula*.
- Se enuncia con propiedad el resultado: se puede rechazar la hipótesis de independencia del uso de acero *Ferroastur* o *Asturacero* respecto de las máquinas *Siemens*, *Tekra* y *Ultimate* con un nivel de significatividad del 95% para el test χ^2.

¿Qué pasa si el área del último paso es mayor que el nivel de significatividad? **Que no se puede concluir nada**. Esto es muy importante: el hecho de que "la hipótesis nula no se puede rechazar" no significa que "la hipótesis nula es cierta", *ni mucho menos*.