INFERENCIA Y ESTIMACIÓN

AUTOR: GABRIEL PÉREZ LANCE

CONTENIDO

INTRODUCCIÓN *3*

1. INFERENCIA Y ESTIMACIÓN *4*

1.1. Variable aleatoria *4*

1.2. Distribución binomial *7*

1.3. Distribución de Poisson *8*

1.4. Distribución exponencial *9*

1.5. Distribución normal o gaussiana *10*

2. INTERVALOS DE CONFIANZA *13*

3. TEST DE HIPÓTESIS *16*

4. REGRESIÓN LOGÍSTICA *19*

5. ESTIMACIÓN BAYESIANA *23*

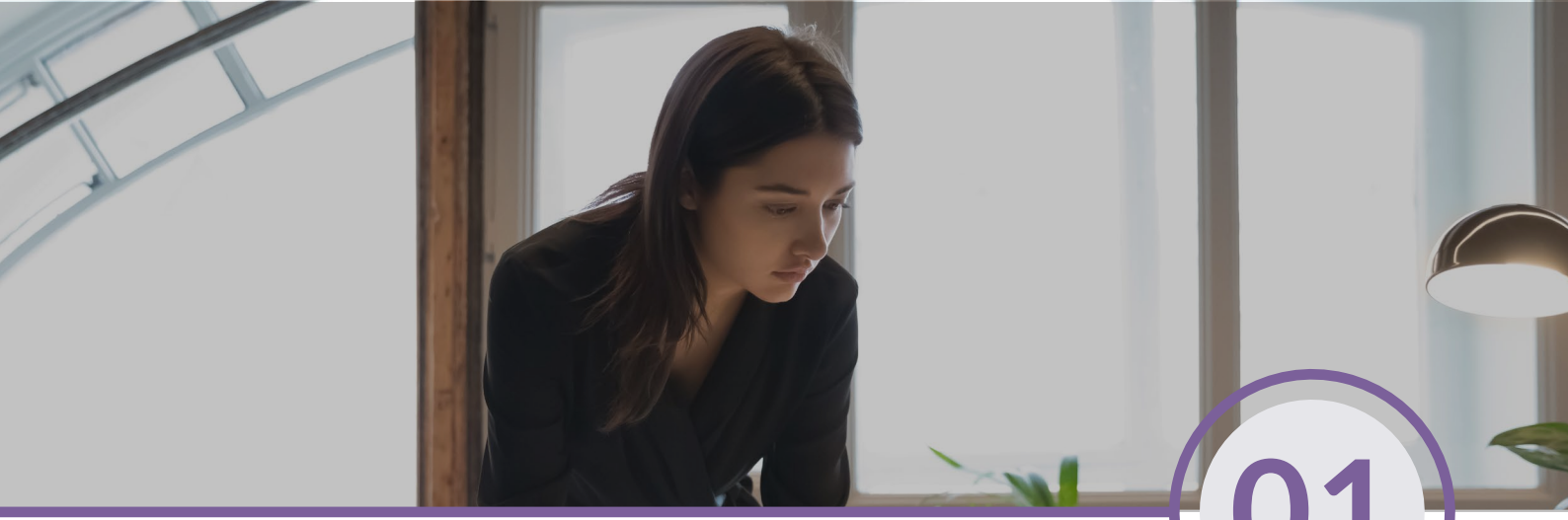
BIBLIOGRAFÍA *29*

INTRODUCCIÓN

Debido a la innumerable cantidad de datos, se hace imprescindible disponer de diversas metodologías que permitan analizarlos. En base a esto surgen las fórmulas estadísticas, las cuales permiten observar las probabilidades, variables y toda aquella información vital para los negocios.

La economía empresarial, entendida como las distintas formas en que se distribuyen los factores de producción, retoma lo que es necesario para establecer la relación existente entre la oferta y la demanda.

En el presente manual, proponemos visualizar aquellas fórmulas y modelos que posibilitan estudiar la información más relevante a la hora de emprender y mantener un negocio.



01

INFERENCIA Y ESTIMACIÓN

1.1. VARIABLE ALEATORIA

Para efectuar pronósticos y realizar estimaciones, es necesario contar con modelos que describan de modo adecuado a las situaciones que se están analizando.

A su vez, para elaborar los modelos, se toman los datos, y con ellos, tal como vimos en el apartado de bondad de ajuste, validamos (o no) si ajustan con determinada distribución.

Cuando hablamos de distribución, nos referimos a distribuciones de probabilidad, y, de esta manera, nos conduce al concepto de "variable aleatoria".

Una variable aleatoria (X), es una asignación entre cada uno de los resultados posibles de un experimento aleatorio, y un número real.

Por ejemplo, Si se realizara un experimento que consista en arrojar 3 monedas no cargadas y se observara el número de caras obtenido, entonces se podría definir a la variable aleatoria, como:

X: número de caras que se obtienen al arrojar 3 monedas

La función de probabilidad asociada a esta variable aleatoria, sería:

X	p (X)
0	1/8
1	3/8
2	3/8
3	1/8

El tercer renglón significa que la probabilidad de que salgan 2 caras es 3/8, pues, para que salgan dos caras, podría darse: CCX, CXC, XCC, es decir 3 alternativas, de un total de 8 posibilidades (CCC, CCX, CXC, ... , XXX).

Esto se puede leer como: P(X=2) = 3/8, es decir p(2)=3/8

La P *"mayúscula"* se aplica al *"evento"* X=2, mientras que la p *"minúscula"* es una función que toma un 2 y devuelve un 3/8. Es importante notar que X *"mayúscula"* es la variable aleatoria, y hace referencia a la frase *"número de caras que se obtienen al arrojar 3 monedas"*, mientras que x *"minúscula"* es un número en particular, por ejemplo, el 2.

Además de la función de probabilidad, se puede definir una *"función de distribución"*, que es la probabilidad acumulada.

La función de distribución F(x), se define como: $F(x) = P(X \leq x)$

La tabla, para este ejemplo, sería:

X	F (X)
0	1/8
1	4/8
2	7/8
3	1

Un concepto muy importante relacionado con las variables aleatorias, es el denominado *"valor esperado"* o *"esperanza matemática"*.

La esperanza matemática es el resultado promedio del experimento, si este se realizara una gran cantidad de veces (en teoría, infinitas veces).

Se calcula según:

$$E(X) = \mu = \sum_{\forall x} x.p(x)$$

Sumado a eso, para conocer el grado de dispersión de la distribución, se define la varianza de una variable aleatoria, como:

$$\sigma^2(X) = \sum_{\forall x} (x - \mu)^2.p(x)$$

Esta expresión es equivalente a:

$$\sigma^2(X) = \left[\sum_{\forall x} x^2.p(x)\right] - \mu^2$$

> Se define el desvío de una variable aleatoria, como la raíz cuadrada de su varianza.

Una variable aleatoria como la mencionada para el caso de las monedas, es una variable que toma valores aislados, es una *"variable aleatoria discreta"*.

En contraposición a esto, existen las *"variables aleatorias continuas"*, estas pueden tomar valores reales. Algunos ejemplos de variables aleatorias continuas son: *peso de una pieza fabricada, tiempo que dura un proceso de atención a un cliente, volumen que una máquina embotelladora cargó en el envase, longitud de una varilla.*

A diferencia de las variables aleatorias discretas, las variables aleatorias continuas tienen asociada una *"función de densidad de probabilidad"* (en lugar de una función de probabilidad).

El área debajo de esta función de densidad entre dos valores "a" y "b" representa la probabilidad de que la variable aleatoria sea mayor que "a" y menor o igual que "b".

También, para el caso del continuo, la función de distribución se define como $F(x)=P(X\leq x)$, y geométricamente, es el área debajo de la función de densidad, a la izquierda del valor "x".

Existen algunas distribuciones que por su naturaleza son de especial interés y muy utilizadas. Tal es el caso de las distribuciones:

- Binomial
- Poisson
- Exponencial
- Normal

NOTAS

Es importante no confundir el desvío y la varianza de un conjunto de datos, con el desvío y la varianza de una variable aleatoria. A pesar de que ambas denominaciones coinciden, se refieren a conceptos diferentes.

1.2. DISTRIBUCIÓN BINOMIAL

Una variable aleatoria sigue una ley binomial, cuando se cumplen las siguientes condiciones:

1. Se trata de n ensayos o repeticiones idénticas
2. Además son ensayos independientes
3. Cada una de las repeticiones tiene un resultado dicotómico (éxito o fracaso)

La probabilidad de éxito de un ensayo individual, se denomina "p", y la de fracaso "q".

La variable aleatoria se define como:

X: cantidad de éxitos obtenidos al realizar n repeticiones

Entonces, de acuerdo con estas definiciones, la función de probabilidad para la distribución binomial, es:

$$p(x) = \binom{n}{x} \; p^x \; q^{n-x}$$

La fórmula para el cálculo de la esperanza en una binomial, es: E(X) = n.p
y el desvío:

$$\sigma(X) = \sqrt{n.p.q}$$

A modo de ejemplo, considérese el caso de una empresa con una flota de treinta camiones que deben realizar diariamente, traslados de mercadería. Supongamos que todos los camiones son similares en cuanto a sus posibilidades de entregar la mercadería en tiempo y forma en el destino previsto. De acuerdo con datos históricos, se sabe que, el 85% de los camiones cumplen con las expectativas respecto de la puntualidad en la entrega. Se pretende determinar cuál es la probabilidad de que más de 25 camiones sean puntuales en la entrega de las mercaderías.

En este caso, se puede considerar que es una binomial, donde el número de ensayos es igual a 30, con una p=0,85 (probabilidad de éxito).

Se quiere calcular, P(X>25), entonces:

P(X>25) = p(26) + p(27) + p(28) + p(29) + p(30)

Aplicando la fórmula de p(x) para la distribución binomial en cada término, se obtiene:

P(X>25) = 0,20281 + 0,17026 + 0,10337 + 0,04040 + 0,00763 = **52,45%**

El valor esperado del número de camiones que serán puntuales en la entrega es:

E(X) = n.p = 30 . 0,85 = 25,5 es decir que, en promedio, considerando varios días, 25,5 camiones serían puntuales.

El desvío sería igual a 1,9557, o sea que es muy poco probable que el número de camiones puntuales se aparte de la media más/menos 2 desvíos, es decir estará comprendido entre (aproximadamente) 22 y 29 camiones.

1.3. DISTRIBUCIÓN DE POISSON

Cuando se quiere describir el número de clientes que llegan a una fila en el lapso de 15 minutos, o el número de accidentes laborales que hay en un período de 4 meses, o bien el número de llamadas que llegan a un conmutador durante media hora, se suele utilizar un modelo de tipo *Poisson*. Todos los casos mencionados, son eventos por unidad de tiempo.

También se aplica a veces, en casos donde se trata de eventos por unidad de espacio, tanto lineal, superficial, o volumétrico. Por ejemplo: el número de fallas que hay en un tramo de 10 metros de cable, o el número de partículas contaminantes por litro de aire.

Esto no significa que, por el sólo hecho de ser eventos por unidad de tiempo o eventos por unidad de espacio, sea Poisson, sino que es habitual que lo sea. Para validar si realmente Poisson es un modelo adecuado, hay que aplicar lo visto en bondad de ajuste anteriormente.

La variable aleatoria en este caso es:

X: número de eventos (llamadas, fallas, cliente, etc) en la unidad de tiempo o espacio considerada (media hora, 10 metros, 1 litro, 15 minutos).

La función de probabilidad en esta distribución, es:

$$p(x) = \frac{e^{-\lambda}.\lambda^{x}}{x!}$$

El valor esperado resulta igual al parámetro lambda (que es justamente la tasa media de eventos)

$$E(X) = \lambda$$

Por su parte, el desvío está dado por:

$$\sigma(X) = \sqrt{\lambda}$$

Por ejemplo, se sabe que en determinado *call center*, llegan a un conmutador (en promedio) 6,2 llamadas en el lapso de 15 minutos. La cantidad de llamadas se distribuye según una ley Poisson. Se quiere conocer cuál es la probabilidad de que en los próximos 10 minutos vengan más de 2 llamadas.

Entonces, se pretende determinar la P(X>2), o lo que es lo mismo, teniendo en cuenta la probabilidad complementaria P(no ocurra A) = 1 - P(ocurra A), resulta: 1 - P(X≤2)

Por lo tanto:

P(X>2) = 1 - P(X≤2) = 1 - [p(0) + p(1) + p(2)] = 1 - [0,01500 + 0,06298 + 0,13226]
= 1 - 0,21024 **= 78,98%**

Se ha utilizado la fórmula de probabilidad para la distribución Poisson mostrada más arriba, con una tasa media de eventos (lambda) igual a 6,2 . 10 / 15 = 4,2

Esto es porque el valor dado era para 15 minutos, y queremos aplicarlo para una ventana de tiempo de 10 minutos.

1.4. DISTRIBUCIÓN EXPONENCIAL

Es habitual que, para modelizar el tiempo de espera de un cliente en una caja, o el tiempo que demora un proceso de fabricación, o bien el tiempo de vida útil de una máquina, se utilice una variable aleatoria de tipo exponencial.

No significa que cualquier proceso temporal esté asociado a esta distribución, sino que, al igual que otras distribuciones y modelos, deben ser validados por el método de bondad de ajuste antes mencionado.

Pero es muy común que la exponencial sea una buena distribución para ese tipo de procesos.

La variable exponencial, a diferencia de la binomial o la poisson - que son discretas - es una variable aleatoria continua. Toma valores dentro de un intervalo real.

La función de densidad de probabilidad para la distribución exponencial, es (para todo valor de x positivo o cero):

$$f(x) = k.e^{-k.x}$$

Dicha función es cero para valores de x negativos.

La función de distribución (es decir la probabilidad acumulada) está dada por:

$$F(x) = 1 - e^{-k.x}$$

El valor esperado de una variable aleatoria exponencial es:

$$E(X) = \sigma(X) = \frac{1}{k}$$

Coincide con el desvío.

Para ejemplificar esta distribución, supongamos que la esperanza de vida útil de un determinado equipo es de 50000 horas. Evaluar cuál es la probabilidad de que el equipo dure al menos 60000 horas considerando que la exponencial es un modelo adecuado.

Si X: duración en horas del equipo hasta su primera falla, entonces:

P(X>60000) = 1 - P(X≤60000) = 1 - F(60000) = 1 - [1- e -50000 / 60000] = e -5 /6 **= 43,46%**

1.5. DISTRIBUCIÓN NORMAL O GAUSSIANA

Una gran cantidad de variables aleatorias sigue una ley denominada "normal". El teorema central del límite explica que la suma de una gran cantidad de variables aleatorias independientes con una dada distribución, aunque esta no sea gaussiana, da como resultado una variable que resulta tendiente a una distribución gaussiana.

Es una distribución de densidad de probabilidad, presente en innumerables situaciones.

Se trata de una variable aleatoria continua que tiene su dominio en todos los reales.

La función de densidad de la denominada campana de Gauss tiene como expresión:

$$f(x) = \frac{1}{\sigma.\sqrt{2\pi}} \; e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

La esperanza y el desvío de esta distribución, son:

$$E(X) = \mu \qquad \sigma(X) = \sigma$$

La obtención de la función de distribución implica resolver una integral con cierto nivel de complejidad, y por eso se suelen utilizar tablas o funciones de planilla de cálculo, o bien métodos numéricos para obtener el valor para un x dado.

Se suele trabajar con una variable aleatoria Z, denominada "variable aleatoria estandarizada" cuya expresión es:

Esta variable Z, tiene $Z = \frac{X - \mu}{\sigma}$ esperanza igual a cero y desvío igual a 1.

La tabla que se muestra a continuación, es la probabilidad acumulada desde Z=0, hasta un valor dado de z.

Normal (0;1) (z) Φ (z)

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,00000	0,00399	0,00798	0,01197	0,01595	0,01994	0,02392	0,02790	0,03188	0,03586
0,1	0,03983	0,04380	0,04776	0,05172	0,05567	0,05962	0,06356	0,06749	0,07142	0,07535
0,2	0,07926	0,08317	0,08706	0,09095	0,09483	0,09871	0,10257	0,10642	0,11026	0,11409
0,3	0,11791	0,12172	0,12552	0,12930	0,13307	0,13683	0,14058	0,14431	0,14803	0,15173
0,4	0,15542	0,15910	0,16276	0,16640	0,17003	0,17364	0,17724	0,18082	0,18439	0,18793
0,5	0,19146	0,19497	0,19847	0,20194	0,20540	0,20884	0,21226	0,21566	0,21904	0,22240
0,6	0,22575	0,22907	0,23237	0,23565	0,23891	0,24215	0,24537	0,24857	0,25175	0,25490
0,7	0,25804	0,26115	0,26424	0,26730	0,27035	0,27337	0,27637	0,27935	0,28230	0,28524
0,8	0,28814	0,29103	0,29389	0,29673	0,29955	0,30234	0,30511	0,30785	0,31057	0,31327
0,9	0,31594	0,31859	0,32121	0,32381	0,32639	0,32894	0,33147	0,33398	0,33646	0,33891
1	0,34134	0,34375	0,34614	0,34849	0,35083	0,35314	0,35543	0,35769	0,35993	0,36214
1,1	0,36433	0,36650	0,36864	0,37076	0,37286	0,37493	0,37698	0,37900	0,38100	0,38298
1,2	0,38493	0,38686	0,38877	0,39065	0,39251	0,39435	0,39617	0,39796	0,39973	0,40147
1,3	0,40320	0,40490	0,40658	0,40824	0,40988	0,41149	0,41308	0,41466	0,41621	0,41774
1,4	0,41924	0,42073	0,42220	0,42364	0,42507	0,42647	0,42785	0,42922	0,43056	0,43189
1,5	0,43319	0,43448	0,43574	0,43699	0,43822	0,43943	0,44062	0,44179	0,44295	0,44408
1,6	0,44520	0,44630	0,44738	0,44845	0,44950	0,45053	0,45154	0,45254	0,45352	0,45449
1,7	0,45543	0,45637	0,45728	0,45818	0,45907	0,45994	0,46080	0,46164	0,46246	0,46327
1,8	0,46407	0,46485	0,46562	0,46638	0,46712	0,46784	0,46856	0,46926	0,46995	0,47062
1,9	0,47128	0,47193	0,47257	0,47320	0,47381	0,47441	0,47500	0,47558	0,47615	0,47670
2	0,47725	0,47778	0,47831	0,47882	0,47932	0,47982	0,48030	0,48077	0,48124	0,48169
2,1	0,48214	0,48257	0,48300	0,48341	0,48382	0,48422	0,48461	0,48500	0,48537	0,48574
2,2	0,48610	0,48645	0,48679	0,48713	0,48745	0,48778	0,48809	0,48840	0,48870	0,48899
2,3	0,48928	0,48956	0,48983	0,49010	0,49036	0,49061	0,49086	0,49111	0,49134	0,49158

2,4	0,49180	0,49202	0,49224	0,49245	0,49266	0,49286	0,49305	0,49324	0,49343	0,49361
2,5	0,49379	0,49396	0,49413	0,49430	0,49446	0,49461	0,49477	0,49492	0,49506	0,49520
2,6	0,49534	0,49547	0,49560	0,49573	0,49585	0,49598	0,49609	0,49621	0,49632	0,49643
2,7	0,49653	0,49664	0,49674	0,49683	0,49693	0,49702	0,49711	0,49720	0,49728	0,49736
2,8	0,49744	0,49752	0,49760	0,49767	0,49774	0,49781	0,49788	0,49795	0,49801	0,49807
2,9	0,49813	0,49819	0,49825	0,49831	0,49836	0,49841	0,49846	0,49851	0,49856	0,49861
3	0,49865	0,49869	0,49874	0,49878	0,49882	0,49886	0,49889	0,49893	0,49896	0,49900
3,1	0,49903	0,49906	0,49910	0,49913	0,49916	0,49918	0,49921	0,49924	0,49926	0,49929
3,2	0,49931	0,49934	0,49936	0,49938	0,49940	0,49942	0,49944	0,49946	0,49948	0,49950
3,3	0,49952	0,49953	0,49955	0,49957	0,49958	0,49960	0,49961	0,49962	0,49964	0,49965
3,4	0,49966	0,49968	0,49969	0,49970	0,49971	0,49972	0,49973	0,49974	0,49975	0,49976
3,5	0,49977	0,49978	0,49978	0,49979	0,49980	0,49981	0,49981	0,49982	0,49983	0,49983
3,6	0,49984	0,49985	0,49985	0,49986	0,49986	0,49987	0,49987	0,49988	0,49988	0,49989
3,7	0,49989	0,49990	0,49990	0,49990	0,49991	0,49991	0,49992	0,49992	0,49992	0,49992
3,8	0,49993	0,49993	0,49993	0,49994	0,49994	0,49994	0,49994	0,49995	0,49995	0,49995
3,9	0,49995	0,49995	0,49996	0,49996	0,49996	0,49996	0,49996	0,49996	0,49997	0,49997
4	0,49997	0,49997	0,49997	0,49997	0,49997	0,49997	0,49998	0,49998	0,49998	0,49998

Figura 1: Probabilidad acumulada
(Elaboración propia, 2021)

Con la finalidad de mostrar el empleo de dicha tabla para calcular probabilidades en una variable normal, supongamos que el tiempo de permanencia de un visitante en una determinada página web es una variable aleatoria normal con media igual a 8,4 minutos y un desvío de 2,1 minutos. Se desea calcular el porcentaje de visitantes que permanezcan entre 5 y 11 minutos.

En primer lugar, calculamos la variable aleatoria estandarizada, correspondiente a los valores de x1=5 y x2=11.

$$z1 = (x1 - 8,4) / 2,1 \quad \Rightarrow \quad z1 = (5 - 8,4) / 2,1 = -1,62$$
$$z2 = (x2 - 8,4) / 2,1 \quad \Rightarrow \quad z2 = (11 - 8,4) / 2,1 = 1,24$$

En la tabla, para encontrar el valor correspondiente a “1,62”, se busca el en la fila del 1,6 y la columna del 0,02, y vemos que en la celda se encuentra el 0,44738.

Análogamente, para encontrar el valor correspondiente a “1,24”, se busca el en la fila del 1,2 y la columna del 0,04, y vemos que en la celda se encuentra el 0,39251.

Esto significa que la probabilidad acumulada desde el centro z=0, hasta z=-1,62 es 0,44738, y la probabilidad entre z=0 y z=1,24 es igual a 0,39251.
Por lo tanto, la probabilidad entre z=-1,62 y z=1,24 es igual a 0,44738 + 0,39251 = 0,83989.

Entonces la probabilidad de que un visitante permanezca entre 5 y 11 minutos es igual a 83,989%.

De este modo, la estimación es que el **83,989%** de los visitantes permanecerán entre 5 y 11 minutos en la mencionada página web.

02

INTERVALOS DE CONFIANZA

Una de las maneras de realizar estimaciones y pronósticos es mediante los denominados *"intervalos de confianza"*.

Esto se basa en tomar una muestra al azar, de tamaño "n", de una población de tamaño "N". Con los datos de esta muestra se calcula la media y el desvío insesgado, y con estos valores se puede elaborar un intervalo que tendrá una cierta probabilidad de contener al verdadero valor de la media poblacional.

La probabilidad de que ese intervalo contenga a la media de la población, se denomina *"nivel de confianza"* NC.

Se define *"nivel de significación"* (α) como α = 1 - NC

El intervalo de confianza se calcula con la expresión:

$$Int = \overline{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

zα/2 es el valor en la tabla de Gauss que deja a la derecha una probabilidad igual a α/2. En general se considera un nivel de confianza del 95%, por lo tanto, el nivel de significación es del 5%, es decir α=0,05. Por lo tanto, en la tabla mostrada anteriormente, habría que buscar cuál es el valor de z que acumule desde el centro un 0,475 de probabilidad. Se puede observar que ese valor es el z=1,96.

El desvío poblacional (sigma) se aproxima mediante el desvío insesgado (S), que se obtiene a partir de los valores de la muestra utilizando la fórmula vista en estadística descriptiva.

La fórmula dada, es válida para calcular un intervalo de confianza cuando se trata de *"muestras grandes"* (típicamente n=30 o más), si esto no se cumple, entonces en lugar de utilizar la distribución de Gauss, habría que emplear la distribución denominada "t de Student" con "n-1 grados de libertad", que contempla el caso en que la muestra es pequeña y se desconoce la varianza de la población.

Otra hipótesis que se ha considerado para elaborar la fórmula del intervalo de confianza, es que se trata de *"población infinita"*, esto es, la población es mucho más grande que el tamaño de la muestra, o sea N es mucho mayor que n. Si esto no fuera cierto, habría que aplicar lo que se conoce como *"factor de corrección por población finita"* y está dado por:

$$fc = \sqrt{\frac{N-n}{N-1}}$$

Incorporando este factor de corrección, el intervalo de confianza es:

$$Int = \overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

En algunos casos, se busca conocer una cota de la media poblacional, es decir cuánto es como máximo o cuánto es como mínimo. Para eso se pueden utilizar *"intervalos de confianza unilaterales"*.

Las fórmulas para dar estimaciones de la cota superior y de la cota inferior de la media poblacional, son:

$$Cota\ sup = \overline{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

$$Cota\ inf = \overline{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

En estos casos, se requiere zα (en lugar de zα/2). Para una confianza del 95%, zα es igual a 1,64.

A modo de ejemplo, tomemos el siguiente caso:

En una línea de producción, en la etapa de embotellado, se necesita estimar cuál es el valor medio del líquido que se carga

en cada botella. Para eso, se toma una muestra de 40 botellas al azar y se observa que la media muestral es igual a 1020 cc y el desvío insesgado es de 60 cc. Determinar un intervalo de confianza para estimar la media poblacional de todas las botellas de la producción, basándose en los datos de la muestra tomada.

Entonces en el caso planteado es: **n=40, media=1020 y S=8,5**

Por lo tanto, el intervalo de confianza para la estimación de la media poblacional es:

int = 1020 +- 1,96 . 60 / sqrt(40) = 1020 +- 18,6 = [1001,4 ; 1038,6] con un 95% de confianza. (sqrt representa la función "raíz cuadrada").

Esto significa que hay una probabilidad de 95% de que el intervalo [1001,4 ; 1038,6] contenga el verdadero valor promedio del contenido de las botellas de toda la producción.

En algunos casos, es necesario estimar el parámetro *"proporción poblacional"* (p) correspondiente a una *"población dicotómica"*.

Es decir, podría ser el caso de tener un lote de una gran cantidad de piezas, y se pretende estimar el porcentaje de aquellas defectuosas dentro de ese lote. Para eso, se toma una muestra al azar de tamaño "n", y se observa que dentro de la muestra hay una cantidad "x" de piezas defectuosas.

Entonces, el intervalo de confianza para la estimación de la proporción poblacional "p", basado en la *"proporción muestra"* x/n, está dado por:

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

$$Int = \overline{p} \pm z_{\alpha/2}\sqrt{\frac{\overline{p}.(1-\overline{p})}{n}}$$

Si en el caso anterior, en el muestreo de n=400 piezas, se encuentran 36 defectuosas, entonces el intervalo de confianza para p, sería:

proporción muestral = 36 / 400 = 0,09
int = 0,09 +- 1,96 . sqrt [0,09 . (1-0,09) / 400] = 9% +- 2,65% = [6,35% ; 11,65%]

Es decir que este intervalo tiene una probabilidad del 95% de contener el verdadero porcentaje de piezas defectuosas del lote recibido.

NOTAS

En los intervalos de confianza, el valor que sigue al "+-" se denomina "margen de error". En este caso es igual a 2,65%. En el ejemplo anterior es 18,6.

03

TEST DE HIPÓTESIS

Otro modo de realizar inferencia estadística, es mediante lo que se conoce como *"test de hipótesis"*.

Un test de hipótesis se basa en partir de una suposición inicial (hipótesis nula), y luego de realizar un ensayo basado en una muestra que aportan los datos, ver si es posible sostener la creencia original, o si la evidencia empírica muestra que esa creencia original debe ser rechazada, y entonces, en su lugar, considerar la "hipótesis alternativa".

Formalmente, un test de hipótesis tiene cuatro elementos:

1. **Hipótesis nula H0:** es la afirmación acerca de un valor poblacional. Podría ser, por ejemplo, que la media poblacional fuera igual a cierto valor, o que la proporción poblacional fuera de determinado porcentaje.
2. **Hipótesis alternativa Ha:** se refiere a cuál sería la realidad en el caso de rechazarse la hipótesis nula. Es habitual que sea: media poblacional > valor establecido por H0, o media poblacional < valor establecido por H0 (si se trata de un test a una sola cola); o bien: media poblacional distinto al valor establecido por la H0 (si es el caso de un test a dos colas).
3. **Regiones de aceptación y rechazo:** son las zonas correspondientes a los valores que puede tomar el estadístico de prueba, y que conducen a aceptar o rechazar H0. En general, la zona de rechazo, es: estadístico > valor límite, o estadístico < valor límite.
4. **Estadístico de prueba:** es una variable aleatoria que es función de los datos de la muestra y que tiene una distribución específica (por ejemplo: normal, t de Student, chi cuadrado, F, etc.).

Para el caso de test de hipótesis a dos colas relacionados con la media poblacional, es:

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0 \text{ (dos colas)}$$
$$H_a : \mu > \mu_0 \text{ (una cola derecha)}$$
$$H_a : \mu < \mu_0 \text{ (una cola izquierda)}$$

$$z_{test} = \frac{\overline{x} - \mu}{\frac{S}{\sqrt{n}}}$$

Para la proporción poblacional:

$$H_0 : p = p_0$$
$$H_a : p \neq p_0 \text{ (dos colas)}$$
$$H_a : p > p_0 \text{ (una cola derecha)}$$
$$H_a : p < p_0 \text{ (una cola izquierda)}$$

$$z_{test} = \frac{\overline{p} - p}{\sqrt{\frac{\overline{p}\,(1-\overline{p})}{n}}}$$

Para determinar el tamaño muestral correspondiente al caso de una proporción, se asume el peor caso (p=0.5) y en base a eso, se puede determinar el valor mínimo de n necesario para un margen de error (ME), dado un determinado nivel de significación alfa, según la expresión:

$$n = \frac{0{,}25 z^2_{\alpha/2}}{ME^2}$$

Tomemos el caso de una empresa que fabrica y vende cables de acero para elevadores. Según la especificación los cables soportan una tensión de corte media de 5000 Kg. Si los cables tuvieran una tensión media menor, sería un problema, pues habría quejas por parte de los clientes y riesgos de acciones legales por las consecuencias debido a los accidentes que podrían ocurrir. Tampoco sería adecuado que los cables soporten una tensión que esté por encima de la especificada, porque entonces significa que se está generando un producto que debería tener una especificación de mejor calidad, y por tanto, un precio mayor. Debido a todo esto, es necesario que se analice si los cables fabricados cumplen con esta especificación.

Para eso, se toma una muestra de 64 cables, se los somete a tensión, y se observa para cada uno de ellos, cuál es la tensión a la que se cortan. Se obtiene una media muestral de 4900 Kg y un desvío muestral insesgado igual a 360 Kg.

$$H_0 : \mu = 5000$$
$$H_a : \mu \neq \mu_0 \text{ (dos colas)}$$

$$z_{test} = \frac{\overline{x} - \mu}{\frac{S}{\sqrt{n}}}$$

El planteo mediante un test de hipótesis sería el siguiente:

Es un test a dos colas, pues nos importa que la media no sea mayor y que no sea menor.

Si reemplazamos los valores de la media poblacional supuesta (5000), n=64, media muestra igual a 4900, y S=360 en la fórmula del estadístico de prueba, se obtiene: ztest = 2,22

Si tomamos un nivel de confianza para el test igual a 95%, entonces el nivel de significación alfa es de 0,05. Entonces, la zona de rechazo es para z mayor que 1,96 o z menor que -1,96.

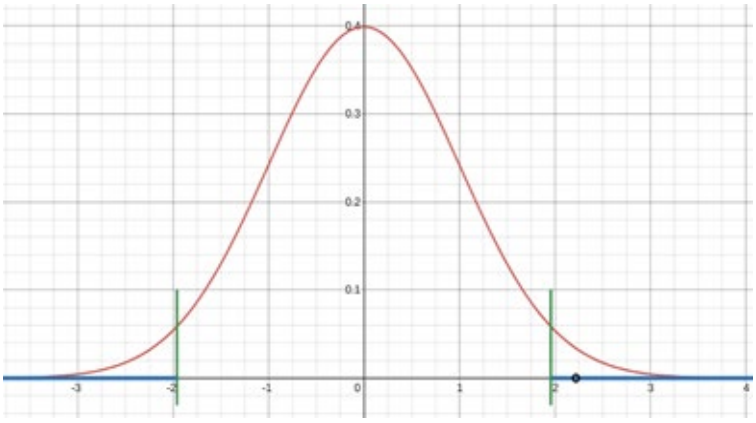


Figura 2: Zona de rechazo
(Elaboración propia, 2021)

Por lo tanto, el z del ensayo igual a 2,22 cae en la zona de rechazo, y esto significa que, a partir de los datos, hay evidencia de que los cables tienen su tensión de corte por debajo del valor nominal, y entonces debería revisarse la materia prima o el proceso de fabricación, para corregir el problema.

04

REGRESIÓN LOGÍSTICA

Con frecuencia nos podemos encontrar con la necesidad de predecir una variable binaria. Esta podría ser éxito - no éxito, percepción positiva-percepción negativa, curado-no curado, etc.

Como nuestro objetivo es predecir, utilizaremos la siguiente estrategia, primero conseguir datos y con estos, generar un modelo que permita, basándose en lo que ocurrió y en alguna variable explicativa, estimar lo que podría pasar.

Las herramientas que utilizamos para tomar la data pasada y usarla para generar una predicción son los llamados modelos y existe una gran cantidad de estos. Una subcategoría son los ya vistos en un módulo anterior, bajo el nombre de modelos de regresión, y su atractivo era su relativa simpleza respecto de los otros modelos. Vimos que, en la regresión lineal, básicamente consiste en ajustar una recta a la data de modo de minimizar el error cuadrático cometido.

La regresión logística hace algo similar, no obstante, con algunas diferencias. Estas diferencias vienen motivadas por el hecho de que la variable explicada ahora no es una variable cuantitativa, sino que es una variable binaria.

¿Cómo se modifica el modelo de regresión lineal para explicar variables binarias?

En principio se requiere *"mapear"* el rango de valores de la componente "y" (que en principio podría tomar cualquier valor real) al intervalo [0, 1], para que pueda ser interpretado como una probabilidad.

Para eso, se utiliza la denominada *"función sigmoide"*:

$$p = \sigma(x) = \frac{1}{1 + e^{-x}}$$

y su inversa, que es la "*función logit*":

$$logit(p) = \log\left(\frac{p}{1-p}\right)$$

En ambas fórmulas, la p simboliza la probabilidad de que la variable binaria toma el valor correspondiente al "éxito" por así decirlo, mientras que 1 - p simboliza la probabilidad de que la variable binaria tome el valor asociado al *"no éxito"*.

Veamos cómo funciona esta modelización mediante la regresión logística, a través de un ejemplo paso a paso:

Digamos que tenemos una tabla de datos con dos columnas:

X: miles de pesos invertidos por diferentes empresas en el packaging de su último producto

Y: variable binaria que es 1 si el producto es exitoso y 0 si no

X	5	10	16	19	20	22	24	26	30
y	0	0	0	0	1	1	1	1	1

En el siguiente gráfico vemos esa hipotética tabla:

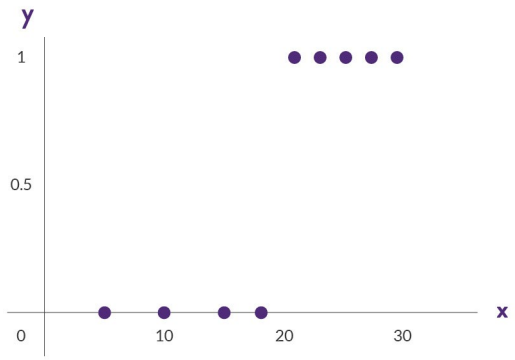


Figura 3: Variables hipotéticas
(Elaboración propia, 2021)

Lo que hacemos ahora es tomar la función sigmoide e ir desplazándola, de forma tal de lograr una versión, que ajuste lo mejor posible a la data. Se puede observar que ese sería el caso de la curva roja en la siguiente imagen:

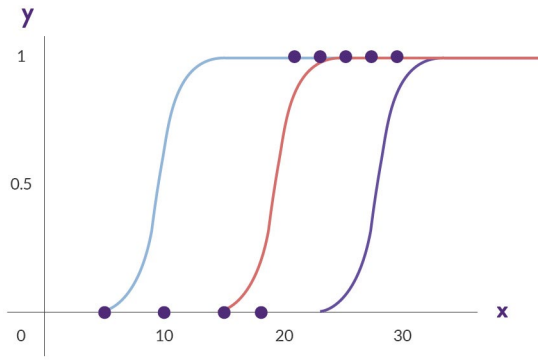


Figura 4: Desplazamiento de la función sigmoide
(Elaboración propia, 2021)

En términos técnicos, esta sigmoide roja es la que maximiza la denominada "función de verosimilitud" de la data considerada. El significado de la función de verosimilitud, es la probabilidad de que obtenga los datos que obtuve. Maximizar esta función, equivale a buscar los parámetros que hagan que lo que ocurrió, sea lo que tenía mayor probabilidad de ocurrencia. Si se plantea esto como un objetivo, surgen los parámetros alfa y beta, correspondientes a la sigmoide graficada en color rojo.

Utilizar el criterio de "*máxima verosimilitud*" (en lugar de minimizar los cuadrados de los residuos), es justamente la diferencia central entre la regresión logística con respecto a la regresión lineal.

La implementación matemática de estos cálculos es relativamente compleja, y no es la idea realizarlos manualmente. Para esto, justamente hay software que lo implementa (por ejemplo, en R).

Si se ingresa en ese software la data correspondiente a la tabla mostrada anteriormente, se obtiene (para un nivel de confianza del 95%):

α = −751.97
β = 38.56
Valor p = 0.998

Por lo tanto, el modelo es:

$$P = \frac{1}{1 + e^{-(-751{,}97 + 38{,}56\,x)}}$$

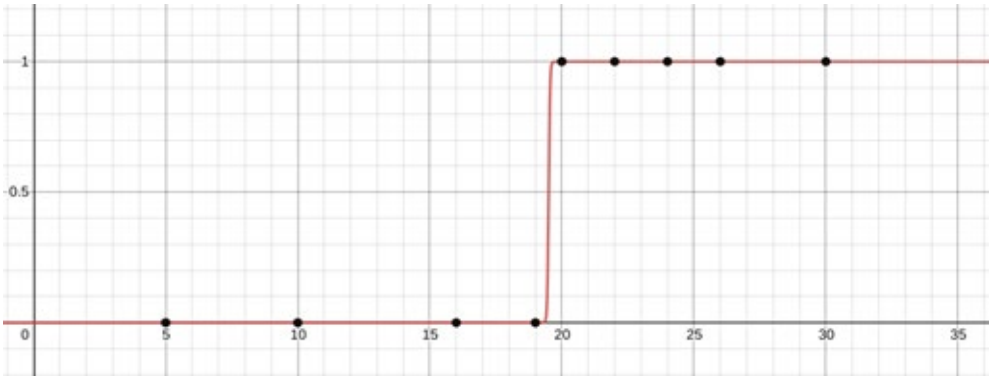


Figura 5: Implementación de software
(Elaboración propia, 2021)

Para observar con mayor detalle la transición de 0 a 1:

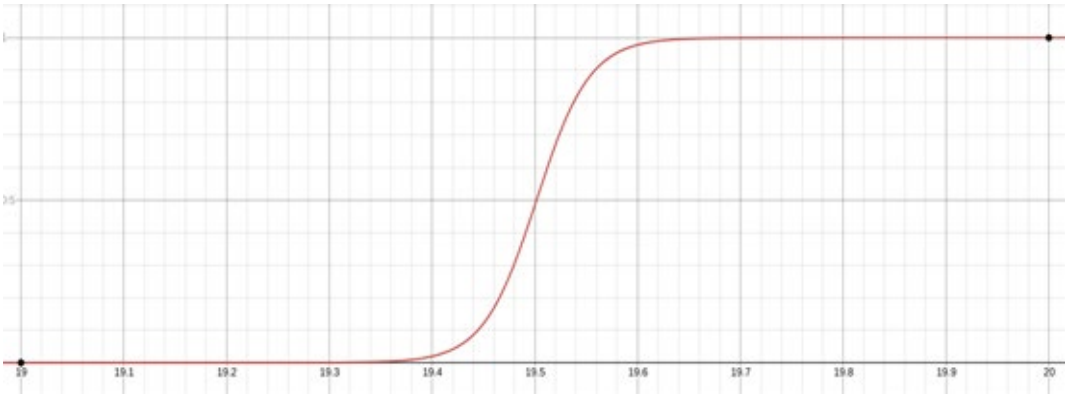


Figura 6: Segunda implementación
(Elaboración propia, 2021)

Para los datos dados, esta resulta ser la sigmoide óptima, y con ella, para los valores futuros de x que tengamos, podremos estimar cuál es la probabilidad de que ocurra el éxito en cada caso.

05

ESTIMACIÓN BAYESIANA

El núcleo filosófico de la estimación bayesiana se basa en la idea de que uno no habla de probabilidades *"absolutas"* o de resultados *"absolutos"* sino que se considera que hay probabilidades o resultados en relación a la data recolectada hasta el momento. Esto es, se comienza con alguna suposición previa acerca de cómo se comporta algún proceso aleatorio, y a partir de la recolección de evidencia empírica, se corrige la creencia previa que se tenía con respecto al comportamiento estadístico del proceso aleatorio.

Ahora bien, el punto es cómo, dada una creencia previa y la data recolectada, ajustar la creencia inicial en función de los datos que se fueron obteniendo. La respuesta a esta pregunta es el eje de la denominada *"estimación bayesiana"*, y tiene su origen en el Teorema de Bayes. La fórmula matemática de este teorema es la siguiente:

$$P(A\,|B) = \frac{P(B|A\,)P(A\,)}{P(B)}$$

Si tomamos A como un evento de interés y B como información que tenemos disponible, cada parte de esta ecuación tiene un significado:

P(A) es la llamada probabilidad "a priori" de A y representa lo que creemos que va a pasar, o nuestras expectativas, sin tomar en consideración la información disponible.

P(B|A) representa la probabilidad de tener la información que tenemos suponiendo que los procesos son como creíamos que eran "a priori".

P(B) es la probabilidad de tener la data recolectada teniendo en cuenta "todos los estados posibles".

Finalmente, P(A|B) representa la creencia actualizada por los nuevos datos obtenidos.

Precisemos esto con un ejemplo:

Tomemos como A al evento “el paciente tiene la enfermedad X” y tomemos como B a “el test dio positivo”.

Supongamos que vamos a hacernos un chequeo para ver si tenemos una enfermedad X, y el test nos da positivo ¿Significa esto necesariamente que tenemos la enfermedad X? Es posible, pero también existe la posibilidad de que el test haya fallado y nos haya arrojado un falso positivo. Ante esto nos interesa saber entonces cuál es la probabilidad de estar enfermos con X, sabiendo que el test nos dio positivo. Es decir, nos interesa conocer P(A|B).

Para utilizar el teorema de Bayes necesitamos una creencia a priori, es decir un valor de P(A). Digamos que hacemos un breve estudio y vemos que 10 de cada 100 personas están enfermas con X. En este caso P(A) = 10/100 = 0.1.

También requerimos conocer la cantidad P(B|A) ¿Qué representaba esto? Esto era la probabilidad de observar B suponiendo que ocurre A. Esto se traduce acá en: la probabilidad de que el test diera positivo si se tiene que la persona tiene la enfermedad X. Esto, de alguna manera, habla de lo bueno que es el test. Supongamos que este P(B|A) = 0.95.

Sólo resta conocer P(B). Esto sería: la probabilidad de tener la data recolectada teniendo en cuenta “todos los estados posibles” ¿Cómo se interpreta esto? ¿Cuáles son los estados? La respuesta es: o se está enfermo o no se está enfermo. Esto es lo mismo que decir A o no A. Matemáticamente podemos expresar a P(B) entonces como:

$$P(B) = P(B|A)\,P(A) + P(B|\text{ no } A)P(\text{no } A)$$

Esto se interpreta del siguiente modo: la probabilidad de que el test de positivo es igual a la probabilidad de que de positivo ante un paciente con la enfermedad (teniendo en cuenta qué tan probable es que un paciente esté enfermo), más la probabilidad de que dé positivo ante un paciente sin la enfermedad (teniendo en cuenta qué tan probable es que un paciente no esté enfermo).

De acuerdo con los datos planteados en el ejemplo, surge:

$$P(B) = P(B|A)\,P(A) + P(B|\text{ no } A)\,P(\text{no } A) = 0{,}95 \,.\, 0{,}1 + 0{,}05.\, 0{,}9 = 0{,}14$$

Si reemplazamos esto en el teorema de Bayes, vemos que entonces:

$$P(A|B) = P(B|A) \,.\, P(A)/P(B) = 0{,}95 \,.\, 0{,}1 \,/\, 0{,}14 = 0.67$$

Es decir, hay un 67% de probabilidad de estar enfermos con X teniendo en cuenta el conocimiento a priori de tratarse de una enfermedad que afecta al 10% de la población y teniendo en cuenta la data, es decir el hecho de que nos dio positivo el test.

Pero entonces, ¿dónde está la actualización de la creencia?

Si nos hubiesen preguntado antes de testearnos que chances tenemos de estar enfermos hubiésemos respondido el a priori P(A), es decir, un 10%. Ahora que nos realizamos el test (recolectamos data), podemos responder que las chances, en realidad, eran de un 67%. Así, nuestra creencia fue actualizada por la información contenida en la data.

Lo que acabamos de ver entonces es un ejemplo que muestra la esencia de la estadística bayesiana. Cómo se relaciona esto con la estimación, es lo que veremos a continuación.

Hemos hablado en videos anteriores acerca de estadística inferencial, y del hecho de que, el foco estaba en conocer o estimar parámetros de distribuciones.

En esos videos, además habíamos mencionado que, en general, uno tiene cierta idea a priori de cómo podrían distribuirse dichos parámetros. Es decir, no se conoce exactamente el valor del parámetro, pero se sabe qué valores son más probables en algún sentido.

Entonces, este tipo de problemas se presta adecuadamente bien para ser tratado desde el enfoque bayesiano que hemos comentado. La utilización del teorema de Bayes, en el caso de funciones de densidad, para variables aleatorias continuas, es:

$$f_{\Theta|X=x}(\theta) = \frac{f_{X|\Theta=\theta}(x) f_{\Theta}(\theta)}{f_X(x)}$$

Lo que vemos en esta fórmula es una adaptación general del teorema de Bayes, pero para la manipulación de valores continuos de nuestro parámetro theta (θ). El denominador de esta división sería análogo al P(B) anterior. El numerador sería análogo al P(B|A) y al P(A) en ese orden, y el lado izquierdo de la ecuación sería análogo al P(A|B). Theta, tal como se dijo, sería análogo al A y X en este caso sería análogo al B.

Veamos ahora un ejemplo de aplicación de estimación bayesiana:

Digamos que tenemos una moneda y queremos ver si es una moneda equilibrada. Esto es, si su probabilidad de cara es 0.5. Nuestra creencia a priori, es que la probabilidad de cara de la moneda (θ) podría ser cualquiera. Esta suposición equivale a decir que su probabilidad θ estará distribuida en forma uniforme en el intervalo

(0,1), es decir todos los valores posibles de probabilidad, tienen igual chance a priori.

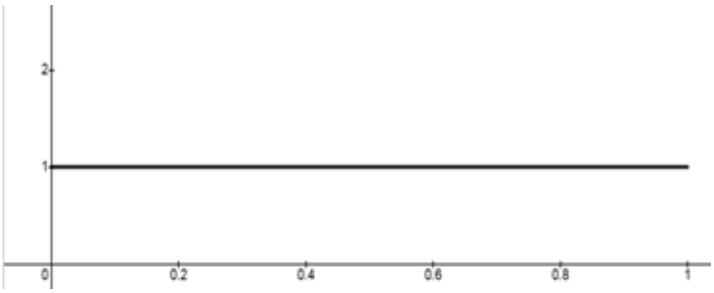


Figura 7: Distribución "a priori" de theta
(Elaboración propia, 2021)

Luego, con la finalidad de obtener información a partir de datos, tenemos que realizar un experimento. Supongamos que arrojamos la moneda 10 veces, y observamos que 7 veces sale cara. Entonces, en este momento correspondería calcular, parametrizado con un θ genérico, la función:

$$f_{X|\Theta=\theta}(x)$$

Dicha fórmula en este experimento resulta en una binomial (n=10, p=θ) evaluada en 7 y esta tiene la siguiente expresión:

$$\binom{10}{7}\theta^7(1-\theta)^{10-7}$$

Cuando hacemos el producto del numerador del teorema de Bayes, encontramos que debemos multiplicar esa binomial con una uniforme. Dado que multiplicar una distribución por una uniforme (0,1) tiene el efecto de no cambiar nada, volvemos a obtener la binomial como resultado. Lo interesante es que entonces obtenemos una distribución que sigue una ley denominada "beta" con parámetros (8, 4).

Una distribución beta, viene dada por la siguiente expresión:

$$Beta(\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

En base a lo dicho, podemos establecer la siguiente tabla:

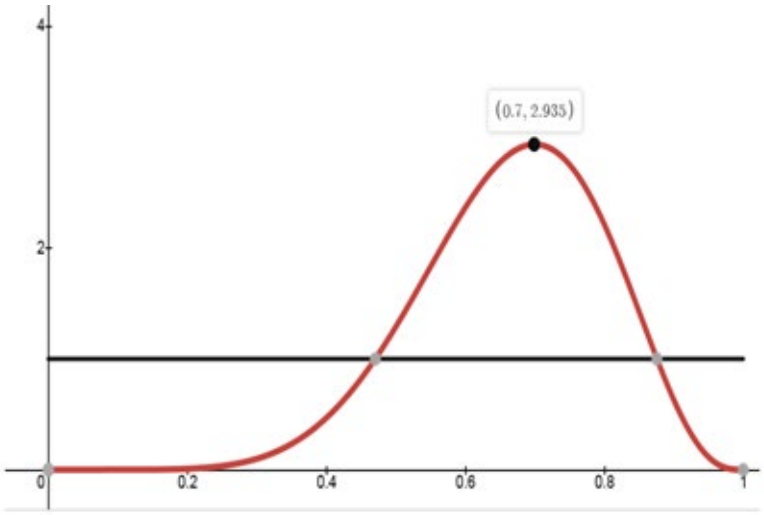


Figura 8: Distribución del parámetro de la probabilidad
(theta) a posteriori (Elaboración propia, 2021)

Vemos ahora, que la distribución tiene un pico centrado en torno a 0,7, justamente esto es por la información que aportaron los datos.

Si ahora volvemos a realizar experimentos para obtener más información y para eso, arrojamos 20 veces la moneda y esta vez obtenemos 16 caras. Entonces ahora, nuestro a priori es la beta (8,4) anterior, y, haciendo nuevamente el producto en el numerador obtenemos lo siguiente:

$$\theta^{16}(1-\theta)^{20-16}\theta^{7}(1-\theta)^{3} = \theta^{23}(1-\theta)^{7}$$

Lo cual conduce a una distribución que resulta en una beta (24,8):

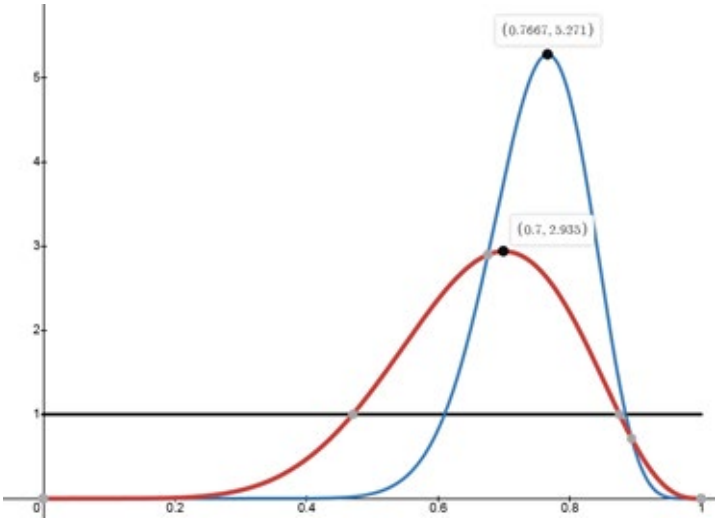


Figura 9: Nueva distribución
(Elaboración propia, 2021)

Vemos entonces, como nuestra creencia de que la moneda está cargada se vuelve cada vez más pronunciada. En la gráfica podemos observar que la curva azul (la creencia más actualizada de todas) tiene una moda de 0.76 y menor desvío que las demás. Esto se puede interpretar como que nuestra experimentación arrojando la moneda (es decir la data recolectada) va indicando y reforzando la creencia de que la moneda tiene una probabilidad de cara de 76%.

BIBLIOGRAFÍA

CANAVOS, GEORGE (1988) Probabilidad y Estadística. Aplicaciones y Métodos. 1ª Edición. México: Mc Graw-Hill.

CARLIN, BRADLEY P. (2009) Bayesian Methods for Data Analysis. CRC Press.

GELMAN, ANDREW B. AND JENNIFER HILL (2018). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

KRUSCHKE, JOHN K. (2015) Doing Bayesian Data Analysis: A Tutorial with R, Jags, and Stan. Academic Press.

MEYER, PAUL (1973) Probabilidad y Aplicaciones Estadísticas. México: Fondo Educativo Interamericano

REFERENCIAS

Imagenes de portada obtenidas de Shutterstock.