«Sí o Sí»: La estadística oculta en las muletillas

Ibai Zubillaga Nogueira

Resumen

La estadística y las matemáticas pueden encontrarse en muchos aspectos de la vida. Casi siempre es posible llevar las cosas más simples al extremo, y darles un enfoque complejo. Eso es justo de lo que trata este trabajo, un estudio probabilístico a fondo sobre algo tan poco relevante como la cantidad de veces en que un profesor dice su frase favorita: «sí o sí».

Palabras clave: Distribución, Poisson, Erlang, muletilla.

Introducción

Existen numerosas palabras en el vocabulario español. Parte de ellas se emplean con muchísima más frecuencia que otras. Por una parte, este fenómeno es debido a la forma en la que está construido el lenguaje, haciendo que palabras como «de», «la» y «que» sean las más frecuentes[1]. Sin embargo, la frecuencia con la que cada palabra es utilizada también puede variar entre individuos.

Se define la muletilla como «voz o frase que alguien repite mucho por hábito». El objetivo de este trabajo es estudiar estadísticamente una de esas muletillas. El experimento se realizó con la frase «sí o sí», la cual fue notoriamente repetida por un profesor a lo largo de sus clases.

En este trabajo se presentan dos aproximaciones distintas a este fenómeno. La primera de ellas asume el suceso de que el profesor diga la muletilla como un proceso de poisson. De esta forma, se asume que el evento es totalmente aleatorio, y se distribuye uniformemente a lo largo de la duración de la clase. Esto, sin embargo, es falso. Realmente, los «sí o sí» se presentaban de forma agrupada, pues al ver el profesor que no había respuesta por parte de los alumnos, procedía a repetir la frase. Sin embargo, la implementación de una distribución probabilística que tenga en cuenta el agrupamiento de muletillas es demasiado compleja como para ser cubierta en este trabajo.

Existe una variante de la muletilla objeto de nuestro estudio: el «sí o no». Esto ofrece una oportunidad estupenda para aplicar la distribución binomial. Esta vez, la duración de la clase no tendría importancia, y se contabilizaría cada «sí o ... » como un experimento cuyos resultados pueden ser terminar en «sí» o «no».

Obtención de datos

Los datos (pág. 6) fueron tomados a lo largo de toda la clase, siempre por la misma persona. Inicialmente se llevó a cabo también la medición de la duración de la clase, pero finalmente se asumió dicha duración constante, de 50 min.

[1] https://corpus.rae.es/frec/1000_formas.txt

Fueron anotados a medida que transcurría la clase el número de «sí o sí» y «sí o no». Si la clase había finalizado, ninguna muletilla adicional era anotada.

El experimento duró **26** clases. A pesar de que un test de Grubbs indicó que no hay datos atípicos con una significancia del 0.05, fueron descartados los datos entre las clases $\mathbf{7-11}$. A lo largo de ese intervalo, el profesor estuvo enfermo. Apareció una tos, acompañada de un descenso drástico en el número de muletillas. La otra razón para descartar estos datos es que rompen por completo (1b) la forma de campana (1a) que sugiere el teorema central del límite (fig. 1).

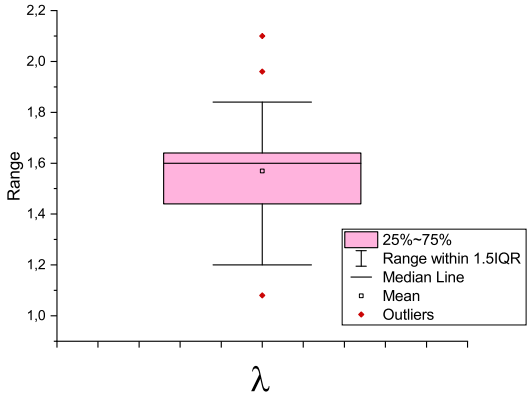


Figura 2: Gráfico de cajas para los valores de λ.

la figura 2 (en sucesos/min).

$$\lambda = 1{,}569\,52\,\text{suceso/min}$$

Aproximación I

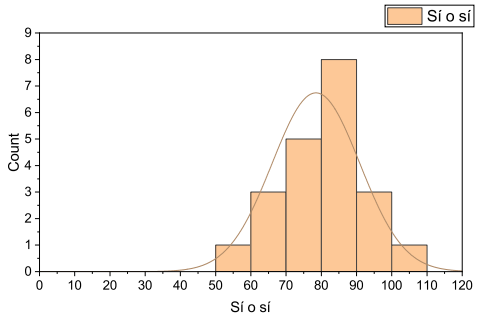
Proceso de Poisson

En este apartado trataremos el fenómeno de la muletilla como un proceso de Poisson, ignorando las agrupaciones de estas.

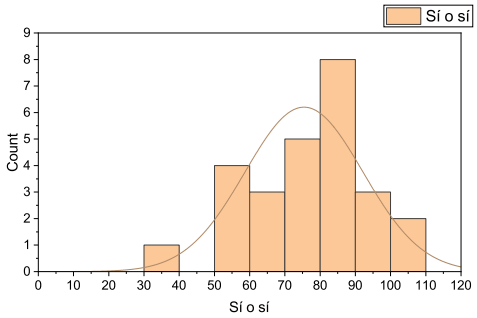
El único parámetro probabilístico a emplear es λ, el cual mide el promedio de sucesos por unidad de tiempo, y se muestra en

Con la distribución de Poisson se puede calcular la probabilidad de que se repita el suceso un número determinado de veces en un tiempo fijado por λ, mediante la siguiente expresión:

$$P(k) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{1}$$



(a) Sin datos atípicos.



(b) Todos los datos.

Figura 1: Histograma del número total de «sí o sí» en cada clase.

La probabilidad ofrecida por la ecuación (1) puede ser útil en otros procesos, pero no lo es mucho en este.

En nuestro caso, es más conveniente la distribución de Erlang. Con esta, es posible conocer la probabilidad de que se dé el suceso un número k de veces en un intervalo Ω de tiempo:

$$P(k, \lambda, \Omega) = \int_{\Omega} \frac{\lambda^k \, t^{k-1} \, e^{-\lambda t}}{(k-1)!} dt \qquad (2)$$

De esta forma, se pueden elaborar tablas como las tablas 1 y 2, donde se predice con una probabilidad preestablecida (95 % y 90 % respectivamente) cuánto tardará el profesor en decir «sí o sí» un número k de veces.

k	min.	k	min.	k	min.
1	1,91	11	10,81	21	18,52
2	3,03	12	11,61	22	19,27
3	4,02	13	12,39	23	20,02
4	4,95	14	13,17	24	20,77
5	5,84	15	13,95	25	21,51
6	6,70	16	14,72	26	22,25
7	7,55	17	15,49	27	22,99
8	8,38	18	16,25	28	23,72
9	9,20	19	17,01	29	24,46
10	10,01	20	17,77	30	25,19

Tabla 1: Tiempo (en minutos) en el cual existe un 95 % de probabilidad de que se dé el suceso k veces.

La elaboración de estas tablas se llevó a cabo con este `script` de *Python*, el cual resolvía numéricamente la integral de la ecuación (2) en el intervalo:

$$\Omega \in [0, a]$$

El valor de a que mejor aproximaba el resultado de la integral al valor de probabilidad preestablecido era seleccionado y tabulado después.

k	min.	k	min.	k	min.
1	1,47	11	9,82	21	17,23
2	2,48	12	10,58	22	17,96
3	3,39	13	11,33	23	18,68
4	4,26	14	12,08	24	19,40
5	5,09	15	12,82	25	20,12
6	5,91	16	13,57	26	20,84
7	6,71	17	14,30	27	21,56
8	7,50	18	15,04	28	22,27
9	8,28	19	15,77	29	22,99
10	9,05	20	16,50	30	23,70

Tabla 2: Tiempo (en minutos) en el cual existe un 90 % de probabilidad de que se dé el suceso k veces.

Aproximación II

Proceso Binomial

Si además del «sí o sí» (éxito) se tiene en cuenta el «sí o no» (fracaso), se puede estudiar el fenómeno de la variación en el final como una distribución binomial. El valor medio calculado para la probabilidad es:

$$P_{\text{Sí}} = 0,76387$$
$$P_{\text{No}} = 0,23613$$

Es posible calcular la probabilidad con la que suceden k éxitos tras un número n de intentos. Para ello se emplea:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (3)$$

Sin embargo, igual que sucedía en el apartado de Poisson, esta fórmula no es de gran interés en este caso. En su lugar, resulta muy interesante la distribución binomial **negativa**.

Esta distribución mide la probabilidad de obtener el r-ésimo éxito tras x intentos:

$$P(x, r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \qquad (4)$$

Se puede encontrar la derivación de esta expresión en el Anexo I (pág. 5).

Análogamente a la Aproximación I, se pueden construir tablas de resultados. En este caso, contienen por un lado un listado del número de éxitos, y por otro el número mínimo de intentos para que se alcance ese número de éxitos, acorde a una probabilidad preestablecida. Para generarlas, se suman los valores arrojados por la ecuación (4) al sustituir $x = r$, $x = r + 1$, $x = r + 2$, ... hasta alcanzar la probabilidad deseada.

Estos cálculos han sido realizados con otro `script` de *Python*, obteniéndose los resultados de las tablas 3 y 4.

k	Int.	k	Int.	k	Int.
1	13	11	70	21	120
2	20	12	75	22	125
3	26	13	80	23	130
4	32	14	85	24	134
5	38	15	90	25	139
6	43	16	95	26	144
7	49	17	100	27	149
8	54	18	105	28	154
9	59	19	110	29	159
10	64	20	115	30	163

Tabla 3: Número mínimo de intentos en los cuales hay un 95 % de que se diga «sí o no» k veces

Gracias a estos valores tabulados, se puede lanzar una rápida predicción de cuántos «sí o ...» harán falta hasta alcanzar el «sí o no» deseado.

Conclusiones

Antes de analizar los datos obtenidos, debe tenerse en cuenta que por mucho que se mida el valor de λ, esta sigue perteneciendo a una distribución normal con una desviación fija. Asumiendo la desviación estándar como

k	Int.	k	Int.	k	Int.
1	7	11	55	21	101
2	12	12	60	22	105
3	17	13	64	23	110
4	22	14	69	24	114
5	27	15	73	25	119
6	32	16	78	26	123
7	36	17	83	27	128
8	41	18	87	28	132
9	46	19	92	29	136
10	50	20	96	30	141

Tabla 4: Número mínimo de intentos en los cuales hay un 75 % de que se diga «sí o no» k veces

la desviación de la distribución, si quisiéramos abarcar el 95 % de los casos (2σ), tendríamos que manejar el siguiente intervalo:

$$\lambda_{\text{mín.}} = 1{,}07238$$
$$\lambda_{\text{MÁX.}} = 2{,}06666$$

Este es un intervalo muy grande. Será de esperar en el 95 % de las clases que λ se encuentre entre esos dos valores, pero no hay forma de saber cuál de todos ellos adoptará hasta que la clase termine. Es por ello que los resultados obtenidos no tienen utilidad práctica, la gran aleatoriedad del experimento con respecto a las pocas veces que se realiza inutiliza por completo cualquier tipo de predicción.

El teorema central del límite afirma que los datos siguen una distribución normal en torno al valor medio de λ. Eso significa que en el 50 % de los casos, el valor de λ en ese experimento en concreto será superior a la media, por lo que se puede afirmar con total seguridad que:

> «En la mitad de los casos, los tiempos tabulados en este documento serán iguales o mayores que los que corresponden a ese experimento y nivel de confianza».

Asimismo, la mitad de las veces, el tiempo necesario para que se den k sucesos será también superior al predicho con una probabilidad correspondiente a la confianza de la tabla.

Respecto a la parte binomial, se puede decir que la fiabilidad de los resultados de esta segunda aproximación es bastante dudosa, pues hace falta un número exageradamente grande de datos para obtener un valor preciso de la probabilidad mediante la regla de Laplace.

En resumen, un fenómeno tan banal como parece ser la cantidad de veces que un profesor dice su muletilla esconde tras de sí un patrón de aparición muy complejo. Si realmente interesa dominar las reglas que rigen la génesis de dicha frase, habría que aplicar distribuciones probabilísticas muy sofisticadas que tengan en cuenta cientos de factores que influyen en el proceso: estado de salud, estado de ánimo, respuesta por parte del alumnado, distracciones, etc. Queda comprobado de esta manera que las simplificaciones grotescas de procesos aparentemente simples como el «sí o sí», como la realizada en este trabajo, no llevan a resultados fructíferos.

Agradecimientos

Quisiera agradecer a María Vázquez Martínez (*la Gallega*) por la inestimable contribución que ha realizado a este trabajo mediante su gran aportación de datos, sin los cuales no habría tenido sentido realizar un estudio tan detallado de estos.

A. Anexo I

Distribución Binomial Negativa

La distribución binomial negativa mide la probabilidad de obtener un r-ésimo éxito en el intento número x. Esto implica que en el intento $x-1$, deben haberse dado $r-1$ éxitos. Esta probabilidad se puede calcular con la ecuación (3) de la siguiente manera:

$$\begin{aligned} P(X = r-1) &= \binom{x-1}{r-1} p^{r-1}(1-p)^{(x-1)-(r-1)} \\ &= \binom{x-1}{r-1} p^{r-1}(1-p)^{x-r} \end{aligned}$$

El siguiente suceso debe ser un éxito, por lo que la probabilidad de que este suceda es p. Multiplicando ambas probabilidades, se obtiene la fórmula de la probabilidad de la distribución binomial negativa (ec. 4).

B. Anexo II

Tabla de datos

Sí o sí	Sí o no	Total	Tiempo /min	λ /min	$P_{Sí}$	P_{No}
91	16	107	50	1,82	0,85047	0,14953
71	31	102	50	1,42	0,69608	0,30392
105	26	131	50	2,10	0,80153	0,19847
92	26	118	50	1,84	0,77966	0,22034
80	17	97	50	1,60	0,82474	0,17526
82	27	109	50	1,64	0,75229	0,24771
108	28	--	50	--	--	--
58	13	--	50	--	--	--
37	45	--	50	--	--	--
51	32	--	50	--	--	--
59	32	--	50	--	--	--
67	48	115	50	1,34	0,58261	0,41739
60	38	98	50	1,20	0,61224	0,38776
54	30	84	50	1,08	0,64286	0,35714
61	31	92	50	1,22	0,66304	0,33696
81	46	127	50	1,62	0,63780	0,36220
82	12	94	50	1,64	0,87234	0,12766
72	27	99	50	1,44	0,72727	0,27273
75	13	88	50	1,50	0,85227	0,14773
80	21	101	50	1,60	0,79208	0,20792
81	16	97	50	1,62	0,83505	0,16495
87	29	116	50	1,74	0,75000	0,25000
74	16	90	50	1,48	0,82222	0,17778
73	17	90	50	1,46	0,81111	0,18889
82	17	99	50	1,64	0,82828	0,17172
98	10	108	50	1,96	0,90741	0,09259

Figura 3: Recopilación de datos de las 26 clases, junto al cálculo de λ y probabilidades para cada una de ellas. Se incluyen en rojo los datos descartados.