Parcial computacional sobre test de hipótesis

Matias Cartelli Fernandez

Departamento de física, Universidad de Buenos Aires

Julio de 2021

1. Distribución:

Elegí dos libros en formato digital, uno en español y otro en alemán, y a partir de ellos construí la distribución de la variable aleatoria t_1: número de letras en una palabra para esos dos idiomas. Calculá su media, moda y mediana.

Para este trabajo elegí La divina comedia de Dante Alighieri. Antes de calcular la distribución para este estadístico t_1 veamos el histograma normalizado con la letras por palabra para todo el texto completo, en ambos idiomas.

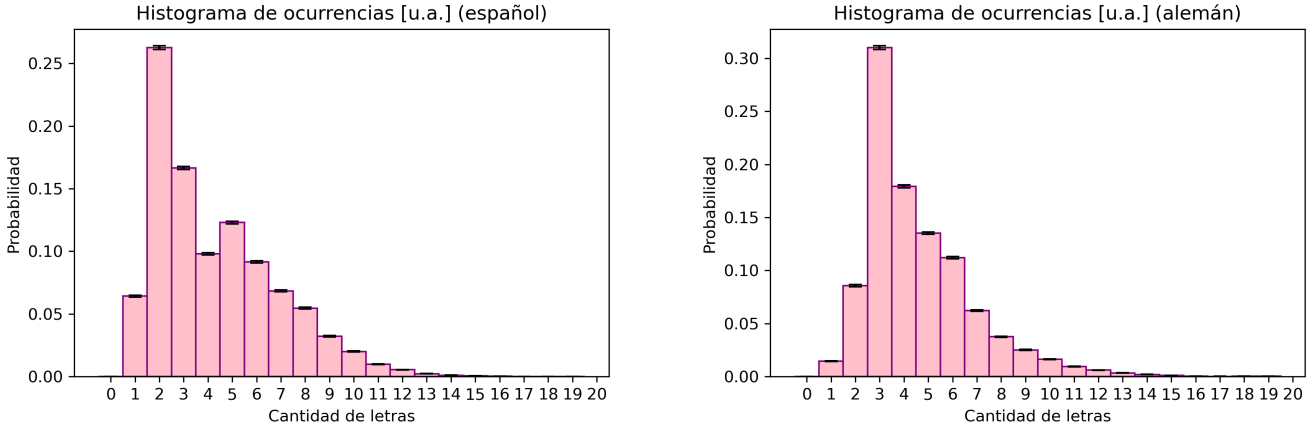


Figura 1: *Histogramas normalizados con la cantidad de letras por palabra para todo el texto, en su correspondiente idioma.*

Notamos que para el texto en español, hay una mayor cantidad de palabras con 2 letras, mientras que para el alemán, la mayor cantidad de palabras son de 3 letras. La forma que siguen ambos gráficos son muy parecidas entre sí.

Ahora nos construimos la distribución del estadístico t_1. Para ello tomamos una palabra al azar del texto y contamos la cantidad de letras, esto lo repetimos muchas veces (en nuestro caso, 1000 veces) con el fin de obtener la distribución de esta variable aleatoria.

Para el caso del texto en español, el histograma con la distribución de t_1 nos queda

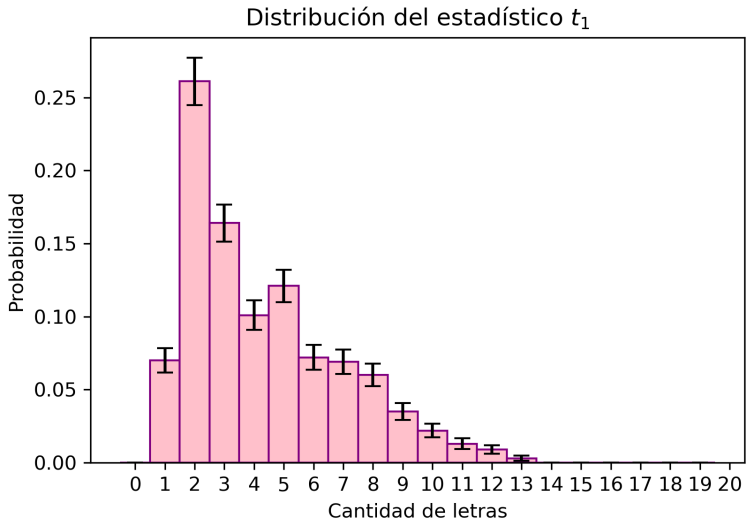


Figura 2: *Distribución del estadístico t_1 para el texto en español.*

Notamos que el gráfico obtenido es prácticamente idéntico al histograma de ocurrencias (es decir, la distribución real) salvo por los errores que son mayores, lo cual se debe obviamente a que para realizar el gráfico se tomaron muchas menos palabras que las que contiene el texto. Esto es lo que se espera, ya que en definitiva para calcular este estadístico estamos tomando palabras al azar del texto y, como éste tiene muchísimas palabras en comparación al número que estamos agarrando, la probabilidad de que tomemos la misma palabra más de una vez es muy baja. Ahora calculando su moda, media y mediana obtenemos

$$\begin{cases} Moda = 2.0 \\ Media = 4.3 \\ Mediana = 4.0 \end{cases}$$

Ahora para el caso del texto en alemán, el histograma con la distribución de t_1 nos queda

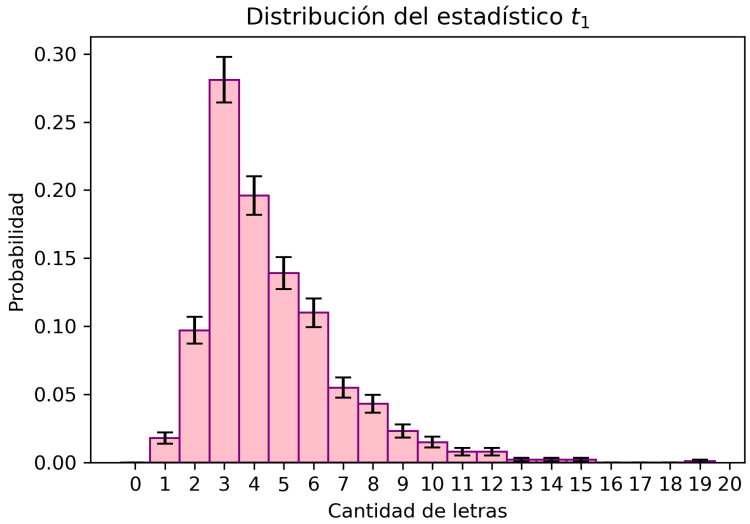


Figura 3: *Distribución del estadístico t_1 para el texto en alemán.*

Nuevamente notamos que es prácticamente idéntica a la distribución real salvo por sus errores. Calculamos su moda,

media y mediana obteniendo

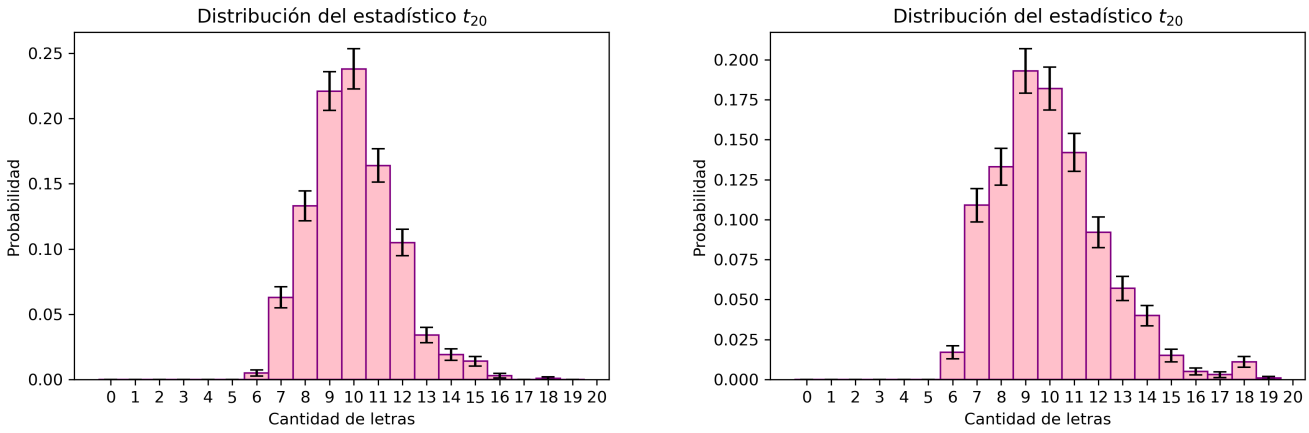
$$\begin{cases} Moda = 3.0 \\ Media = 4.6 \\ Mediana = 4.0 \end{cases}$$

Comparando ambos idiomas, vemos que sus medianas coinciden y las medias son muy parecidas entre sí. Esto lo pudimos prever mirando las distribuciones reales que tienen los textos en cada idioma.

2. El estadístico:

A partir de esas distribuciones calculá, para los dos idiomas, la distribución del estadístico t_{20}: número de letras en la palabra más larga en una muestra de $n = 20$ palabras.

Ahora para armar el estadístico tomamos 20 palabras al azar del texto, y nos quedamos con la palabra que contenga el mayor numero de letras, esto mismo lo repetimos 1000 veces. Con esto obtenemos el histograma para la distribución del estadístico para ambos idiomas.



(a) *Texto en español.*

(b) *Texto en alemán.*

Figura 4: *Distribuciones del estadístico t_{20} para ambos idiomas*

Lo primero que notamos es que ambos gráficos se encuentran corridos respecto de la distribución del estadístico t_1, lo cual es esperable ya que ahora estamos quedándonos con la palabra que tiene mas letras en un conjunto de 20 palabras. Vemos que para ambos idiomas, el pico se encuentra en alrededor de 10 letras. Para el caso del español la palabra con mayor cantidad de letras que aparece es de 18 letras, mientras que para el alemán es de 19 letras, cosa que no se pudo apreciar en los histogramas de ocurrencias debido a la diferencia entre alturas de los bins. Además las palabras más largas aparecen con mayor frecuencia en el texto en alemán.

3. El test:

Determiná cual es el valor crítico t^c_{20} del estadístico t_{20} a partir del cual rechazarías la hipótesis nula H_0: el libro está escrito en español con una significancia de a lo sumo $\alpha = 0.05$.

El valor crítico es el valor para el cual la integral en la distribución desde ese valor en adelante sea igual a α. En este caso nos dicen que la hipótesis nula será que el libro está en español, por ende usamos la distribución del estadístico t_{20} para el texto en español. Debido a que las distribuciones son discretas (histogramas de ancho de bin igual a 1) la integral

no será más que la suma de las alturas de los bins, además imponemos la condición de que la suma de las alturas desde t^c en adelante sea menor o igual a 0.05. Con esto obtenemos entonces que el valor de t^c_{20}

$$\boxed{t^c_{20} = 14}$$

con un valor $\alpha = 0.04$, probabilidad de cometer un error de tipo 1 (descartar H_0 cuando la hipótesis era cierta). De manera que si sumamos la altura del bin número 13, la suma da por arriba de 0.05.

4. Potencia:

Calculá la potencia del test propuesto con t_{20} cuando la hipótesis alternativa es H_1: el libro está escrito en alemán ¿Cómo cambia el resultado si utilizas directamente t_1 como estadístico? ¿Y si usas t_{100}?

Tenemos que la potencia es la probabilidad de rechazar la hipótesis nula cuando la hipótesis alternativa es cierta, en este caso la hipótesis nula será que el libro está en español y la alternativa es que el libro está en alemán. Utilizamos las distribuciones para el estadístico t_{20} (Figuras 4a, 4b). Anteriormente calculamos el valor para t^c bajo la hipótesis nula, entonces para calcular la potencia bajo la hipótesis alternativa debemos integrar la distribución de esta hipótesis desde ese valor de t^c en adelante. Teniendo en cuenta que en este caso, integrar es sumar las alturas obtenemos

$$\boxed{Potencia_{20} = 0.08}$$

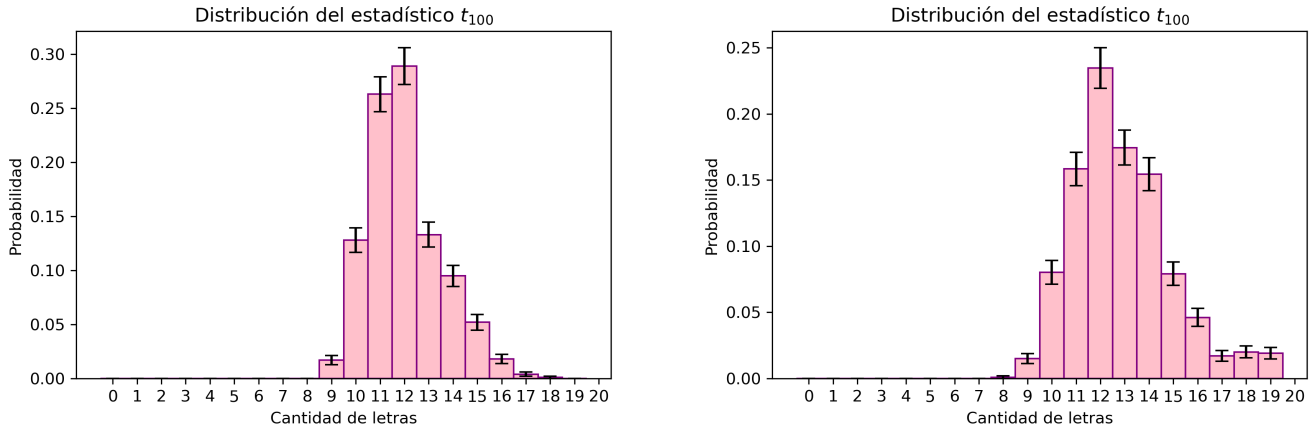
por ende tenemos un 8 % de probabilidad de rechazar H_0 siendo que H_1 es verdadera, lo cual es bastante bajo. Esto parece indicar que el test no es bueno.

Ahora realizamos el mismo procedimiento para los estadísticos t_1 y t_{100}. Para el estadístico t_1 (distribuciones en las figuras 2 (español) y 3 (alemán)) tenemos que el valor de $t^c_1 = 10$ bajo la hipótesis H_0 con un $\alpha = 0.05$ (calculados análogamente como con t_{20}) y con esto obtenemos una potencia integrando la distribución de t_1 para el texto en alemán

$$\boxed{Potencia_1 = 0.04}$$

esto es aún peor que la potencia para t_{20} y se debe a que las distribuciones entre ambos idiomas son mas parecidas cuanto menor es el largo n del conjunto de palabras que agarramos para comparar el largo y tomar el máximo. Por ende esperamos que cuanto más parecidas sean las distribuciones entre los idiomas, menor sea la probabilidad de descartar la hipótesis nula teniendo como verdadera a la alternativa.

Por último calculamos la potencia para el estadístico t_{100} análogamente. Para ésta esperamos que sea mayor que las calculadas anteriormente. Primero hallamos las distribuciones



(a) *Texto en español.*

(b) *Texto en alemán.*

Figura 5: *Distribuciones del estadístico t_{100} para ambos idiomas*

Notamos que se hacen mas pronunciadas las alturas de las palabras largas, y obviamente los picos se corren aún más para este estadístico. Vemos también que las palabras mas largas aparecen con mayor frecuencia en el alemán, al igual que como vimos en la distribución para el estadístico t_{20}, salvo que en este caso es mas pronunciado.

Calculamos el valor crítico para t_{100} bajo la hipótesis H_0 obteniendo $t^c_{100} = 16$ con un $\alpha = 0.02$ (que a mi parecer es bastante chico, pero si sumo una altura más, α sube bastante por encima de 0.05). Integrando en la distribución bajo la hipótesis H_1 (libro en alemán) obtenemos una potencia de

$$\boxed{Potencia_{100} = 0.10}$$

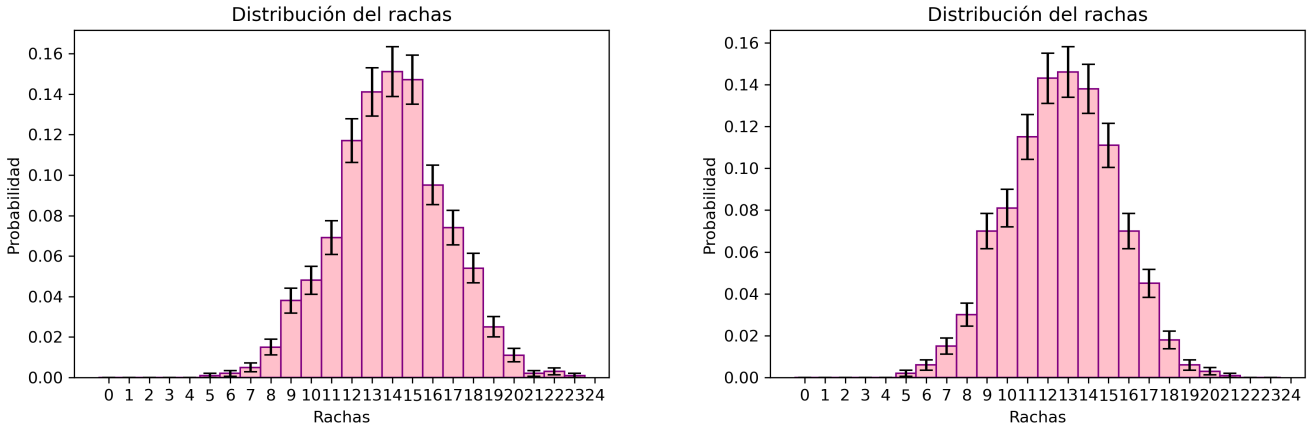
que como esperábamos, es mayor a las calculadas anteriormente (aunque difiere mucho de $Potencia_{20}$) debido a que las distribuciones t_{100} difieren en mayor medida entre ambos idiomas en comparación a las distribuciones para t_1 y t_{20} . Aún así la probabilidad de descartar H_0 tomando como verdadera H_1 sigue siendo bastante baja.

Aunque ya podía verse comparando las distribuciones de los estadísticos para cada idioma (ya que son parecidas), con esto podemos concluir que los estadísticos t_1, t_{20} y t_{100} son malos para testear las hipótesis planteadas (al menos para el valor de significancia pedido). Si en vez de tomar un $\alpha < 0.05$ tomásemos al primer valor para el cuál $\alpha > 0.05$ como significancia la potencia del test aumenta un poco. Aún así siguen siendo malos.

5. Rachas:

Para cada idioma, calculá la distribución del estadístico propuesto en el test de runs aplicado a la variable aleatoria t_1, para una muestra de tamaño 30. En este caso las rachas serán estar por encima o por debajo de la mediana. Si el valor coincide con la mediana entonces tomá uno de los siguientes tres criterios según tu fecha de nacimiento: si cumplís años en enero, febrero, marzo o abril: excluilo de la muestra; si cumplís en mayo, junio, julio o agosto: asumí que está por encima y en otro caso: asumí que está por debajo.

Para este caso nos armamos un nuevo estadístico a partir de t_1. Tomamos una muestra de 30 palabras, contamos su cantidad de letras. Con esto contamos cuantas veces el valor de letras de cada palabra cruza la mediana del estadístico en cuestión. Repitiendo esto para muchas muestras nos generamos un nuevo estadístico que da cuenta de las veces que la variable aleatoria t_1 cruza la mediana del estadístico t_1. A esta nueva variable aleatoria la llamamos *Rachas*. Con esto tomando 1000 muestras obtenemos la distribución para ambos idiomas. Debido a que nací en abril, si el valor cae sobre la mediana lo descarto.



(a) *Texto en español.*

(b) *Texto en alemán.*

Figura 6: *Distribuciones del estadístico Rachas para ambos idiomas.*

Notamos que las distribuciones son bastante parecidas, salvo por que la distribución en español parece estar mas centrada en 14 rachas, mientras que la distribución en alemán parece estar centrada en 13. Además para el caso en español, las rachas mas altas se alcanzan con mas frecuencia.

6. Quedate en casa:

Tomá una muestra de 30 palabras de un libro que tengas en tu casa y aplicale el test del ítem anterior para testear la hipótesis nula de que el libro está escrito en alemán. Para que la elección de las 30 palabras sea efectivamente al azar, usá la que aparece en la posición X de las primeras 30 carillas, donde X es justamente el número de página. Por ejemplo, si mirás la página 9, contá cuantas letras conforman la novena palabra de esa carilla. Calculá el p-valor. ¿Con que nivel de confianza dirías que el libro que agarraste no está escrito en alemán? sin intentar leerlo!

Ahora tomamos una muestra de 30 palabras de algún libro de mi casa. Tenemos que la cantidad de letras por cada palabra es

$$Muestra = [4, 6, 3, 6, 6, 2, 4, 2, 2, 3, 3, 2, 3, 11, 1, 6, 2, 3, 2, 9, 8, 3, 10, 3, 15, 2, 2, 4, 2, 5],$$

Realizamos el test para contar la cantidad de rachas que hay en esta muestra sabiendo que la mediana para el estadístico de t_1 en alemán es de 4.0. Con esto tenemos que para la muestra $Rachas = 15$.

Para calcular el p-valor debemos integrar desde el de rachas medido, en este caso 13, en adelante sobre la distribución bajo la hipótesis que estemos tomando, en este caso H_1. Entonces sumamos las alturas de los bins para la distribución de rachas en alemán desde 15 en adelante. Obtenemos

$$\boxed{\text{p-valor} = 0.25}$$

Sabemos que el nivel de confianza se calcula como $CL = 1 - \alpha$ y queremos hallarlo en base a lo medido bajo la hipótesis de que el libro está en alemán. Para esto podemos tomar $alpha =$ p-valor, es decir, fijamos la significancia en base al valor medido para poder rechazar la hipótesis de manera que el valor medido se encuentre dentro del rango de valores para los cuales rechazaríamos la hipótesis. Con esto obtenemos un nivel de confianza tal que

$$\boxed{CL = 75\,\%}$$

es decir que estamos un 75 % seguros de que el libro que agarramos no está en alemán. Parece ser que este test es mucho mejor que el test del ejercicio **(3)**.