Parcial computacional sobre Test de Hipótesis

Métodos Estadísticos en Física Experimental

Verano 2019

1 Generalidades

1. Indicar nombre, apellido, libreta o DNI en el informe.
2. Dar una descripción clara y precisa de la metodología utilizada.
3. Incluir todos los gráficos como figuras con sus correspondientes leyendas.
4. Justifique las hipótesis en la que se sustenten sus resultados y discuta los resultados obtenidos.
5. Para considerar el informe aprobado hay dos opciones de mínima, realizar los ítems 2, 3, 4 y 7, o bien los ítems 2, 5, 6, 7. Pero si querés hacer más (todos por ejemplo) será considerado para cerrar tu nota final.
6. Enviar a todos los docentes de la práctica la versión digital del informe antes del 5 de abril de 2019. Nombrar al archivo de la siguiente manera: TH-SuApellido.pdf y adjuntar en el mismo mail los códigos desarrollados. Utilizar como asunto del mail: "Parcial TH-2".
7. Por consultas sobre interpretación de los enunciados escribir a: `mefe2019@googlegroups.com`, así todos tienen acceso a todas las respuestas.

2 Estadística en la calle

Desde el 1 de abril de 2016, los automóviles de nuestro pais se patentan con un nuevo sistema, este es de la forma L_1L_2 $N_1N_2N_3$ L_3L_4 donde L son letras y N son números. Así, esta nueva notación, que comenzó con AA000AA, luego AA000AB, ..., AA000ZZ, AA001AA, etc, terminará con ZZ999ZZ.

Hay muchas preguntas que uno podría pretender responder con sólo mirar patentes, como por ejemplo, ¿es uniforme la distribución de patentes? ¿puedo estimar la patente más nueva en circulación? ¿los autos de un barrio, son en promedio más nuevos que los de otro? lo que aprendimos en MEFE nos puede ayudar a responder esas preguntas, y muchas más.

Para eso te vamos a pedir, en primer lugar, un pequeño trabajo de campo, que consiste en sentarte en una plaza o un café y anotar patentes nuevas que veas en circulación (prohibido hacerlo manejando!). Cuantas más patentes anotes, mas representativos serán

tus resultados y más fácil la aplicación de los test porque, en muchos de ellos, podrás usar la aproximación gaussiana para la distribución del estadístico.

Una vez que tengas tu muestra, deberás transformar cada una de las k patentes observadas en un número natural, asignando el 1 a la patente AA000AA, el 2 a la AA000AB y así sucesivamente, hasta llegar a m, el natural correspondiente a la patente más nueva que hayas visto. Una vez que tengas la lista, estarás en condiciones de empezar el parcial.

Dato: Podés usar que la patente más nueva en circulación cuando tomaste la muestra era AD592MF.

Aclaración importante: el ejercicio de mirar patentes en busca de la más nueva posible puede volverse adictivo, tómeselo como un juego y nada mas!

3 ¿Uniformemente distribuidos?

Tenemos razones para pensar que las patentes observadas (o mejor dicho, los naturales que les asociaste) representan una variable aleatoria con distribución uniforme.

1. Aplicá el test de Kolmogorov-Smirnov sobre la muestra y presentá en una misma figura: la distribución teórica y la experimental, el estadístico observado, el p-valor, y decinos si con una significancia $\alpha = 0.05$ se puede rechazar la hipótesis nula. Sugerencia: No uses un test ya implementado en tu lenguaje de programación, codealo vos. De hecho, en el item siguiente no vas a poder evitar tener que hacerlo.

2. Considerá ahora, como hipótesis alternativa, que la distribución de patentes es exponencial de parámetro λ=4×10^{-7} y calculá la potencia del test del ítem anterior. Ayuda: No olvides que la potencia es una propiedad del Test y no de los datos. Y tampoco olvides que el test queda definido por su H_0 y su significancia. Entonces, si H_1 es verdadera, los datos provendrían de una exponencial y pero al aplicar el test, su distribución acumulada experimental sería igualmente comparada con la acumulada de la uniforme (de otro modo hubiera cambiado el test!). Averiguá entonces la distribución del estadístico cuando H_1 es verdadera haciendo una simulación y luego, usando el valor crítico del estadístico hallado en el ítem anterior (para α=0.05) podrás calcular la potencia del test.

4 La patente del auto más nuevo

Para motivar este ítem del parcial, podes leer sobre lo que en estadística se conoce como *El problema de los Tanques alemanes*[1]. Aquí, nos limitaremos a decir que, la distribución de probabilidad de m (el natural correspondiente a la patente más nueva observada), puede escribirse en función del número total de patentes nuevas observadas k y del número de autos con patentes nuevas en circulación n, de la siguiente manera:

$$P(m; k, n) = \frac{\binom{m-1}{k-1}}{\binom{n}{k}} \tag{1}$$

[1] `https://en.wikipedia.org/wiki/German_tank_problem`

1. Repetí N veces el experimento de generar k realizaciones de una variable aleatoria con distribución uniforme en $[1,\ n]$, y para cada experimento calculá el valor de m. Presentá una figura con la distribución $P(m; k, n)$ así obtenida y comparala con la dada por la ecuación (1). Para no tener problemas con los factoriales, calculá $\log(P(m; k, n))$ usando la aproximación de Stirling.

2. Usando inferencia bayesiana con un prior no informativo, encontrá la distribución $P(n; k, m)$. A continuación te sugerimos dos posibles caminos para para lograrlo:

   (a) Para k fijo, obtené una matriz P tal que la entrada (m, n) es $P_{mn} = P(m; k, n)$. Esto se puede realizar por simulación o bien mediante la ecuación (1). Elegí rangos adecuados para m y para n. Luego, para un dado m fijo, la curva de $P(n; k, m)$ para distintos valores de n la obtenés normalizando la fila que corresponda.

   (b) Para k y m fijos, evaluá $P(m; k, n)$ en algún rango de valores de n alrededor del de la patente más nueva. Esto se puede realizar por simulación o bien mediante la ecuación (1). Luego bastará normalizar esa curva para obtener la distribución de n. Dejalo expresado en función de k y m.

3. Ahora que tenés una distribución para n, usando tus valores observados en la calle para k y m, decinos cual es tu estimación bayesiana para n.

4. Si consideramos que al momento de este examen la patente más nueva en circulación es AD592MF[2], calculá cuanta suerte tuviste el día que armaste tu lista. Es decir, ¿cual es la probabilidad de que hagas lo mismo y obtengas una estimación igual o peor que la obtenida para la patente más nueva en circulación? Aclaración: Peor acá sería que tu estimación quede más lejos del valor verdadero de la patente más nueva, ya sea por exceso o por defecto.

5 ¿Independiente del barrio?

En este punto debes compartir tu muestra con otro estudiante de la materia. De este intercambio ambos tendrán dos muestras. Es importante que elijas para el intercambio a alguien que haya tomado los datos en un barrio distinto al tuyo, cuanto más alejado esté mejor.

1. Aplicá el test de Wilcoxon para testear si ambas muestras provienen de poblaciones con la misma esperanza y calculá el p-valor (p_w). Para más información sobre este test ver el libro de Frodesen página 450.

2. Ahora nos gustaría que apliques el estadístico propuesto en el problema 4 de la guía de Test de Hipótesis. Pero antes, notá que no es correcto usar que tiene distribución t-student. ¿Por qué? Tomando como hipótesis nula que la distribución de patentes es uniforme, encontrá computacionalmente la distribución del estadístico propuesto. Sugerencia: Nuevamente, la idea es que generes N veces dos set de datos con distribución uniforme y que calcules el valor del estadístico U. Así tendrás N valores

[2] `https://twitter.com/locosxpatentes?lang=es`

para U y conocerás su distribución. Luego podrás aplicarle el test a tus datos y usar esa distribución para calcular el p-valor.

3. Aplicá sobre tus observaciones el test construido en el ítem anterior (ahora que si conoces la distribución de su estadístico) y calculá el p-valor (p_t).

6 Combinando los tests

Una posible forma de combinar los p-valores obtenidos en los dos test aplicados en el punto anterior (p_w y p_t) es a través del estadístico $T = -2\log(p_w \cdot p_t)$. Notar el signo menos.

1. Si los test fueran independientes, ¿que distribución esperás para T? Acá podes hacer la cuenta analíticamente o bien simularlo.

2. Estudiá que distribución tiene T si se construye con los p-valores dados por los dos test aplicados en el ítem anterior cuando las muestras provienen de poblaciones con la misma esperanza (es decir, cuando la hipótesis nula es verdadera). Comparalo con la distribución que según el item anterior esperas para este estadístico. Usalo para calcular un nuevo p-valor que combine los resultados de ambos test ya aplicados.Notá que para calcular los valores de p_w y p_t para una realización del experimento, se usan los mismos datos.

7 Sé tu propio verdugo

Aunque muy pero muy poco probable, cabe la posibilidad de que alguno de ustedes decidiera inventar la lista de patentes y como no podía ser de otra manera, pensamos aplicar un test de hipótesis para testearlo. El test que decidimos usar tendrá al valor de m que informaste como estadístico y su distribución será la dada por la ecuación (1).

Tomaremos para n el natural correspondiente a la patente AD592MF, y para k, el número de patentes que hayas dicho que observaste. Con eso sabremos cuan probable es que hayas observado lo que decís haber observado o algo más raro.

La buena noticia, es que te damos a vos la oportunidad de elegir la máxima significancia de nuestro test. Es decir, queremos que nos indiques el máximo (o mínimo) valor de α antes de que debamos aceptar la hipótesis de que inventaste tus datos. Para hacerlo, esperamos que vos primero le apliques este test a tus datos y calcules el p-valor. Con eso sabrás informarnos que significancia tomar.