

Resolución del práctico sobre Test de Hipótesis

Juan Piccini

1. Los tiempos de secado de una marca de pintura “A ”siguen (según el fabricante) una distribución normal de parámetros $\mu = 5$ horas y desvío $\sigma = 45$ minutos. La competencia dice que su pintura tiene tiempos de secado también normales con el mismo desvío y media = 4 horas y media. Se toman 20 piezas iguales de madera (misma madera, mismas dimensiones, mismo tenor de humedad, etc.) se eligen al azar 10 de ellas y se les aplica una mano de pintura A, a las otras 10 se les aplica una mano de pintura B y si miden los tiempos de secado de cada pieza, obteniendo los siguientes datos:

A				
4.21	5.09	4.05	6.08	5.17
5.27	5.12	5.26	5.25	6.30

B				
4.29	3.10	4.28	5.21	3.37
3.15	5.07	4.03	4.47	4.50

- b) Plantee un T.H. bilateral para ver si existe una diferencia significativa entre los tiempos de secado, hallando el p-valor del estadístico elegido. (Sugerencia: Si $X \sim N(\mu_x, \sigma_x^2)$ e $Y \sim N(\mu_Y, \sigma_Y^2)$, entonces $X - Y \sim N(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$).

Llamemos D a la resta de ambas muestras, $D = A - B$. Si trabajamos en horas, un desvío de 45 minutos son $3/4$ de hora, de donde $\sigma_X^2 = \sigma_Y^2 = \frac{9}{16}$.

Si H_0 fuese cierta, entonces $D = \{d_1, \dots, d_{10}\}$ debería ser una muestra i.i.d proveniente de una distribución normal con $\mu_D = 0$ y $\sigma_D^2 = \frac{9}{16} + \frac{9}{16} = \frac{18}{16} = \frac{9}{8}$, por tanto $\sigma_D = \sqrt{9/8} = 1.0607$.

- c) Resuelva dicho test “a mano ”y compare el resultado con el del test del paquete statistics de octave.

Hacemos un test sobre la media de una población normal con varianza conocida. El estadístico del test es $\frac{\sqrt{10}(\bar{X}_D - \mu_D)}{\sigma_D} = 3.0798$, quien

debería portarse como una variable aleatoria $\sim N(0,1)$ si H_0 fuese cierta.

La probabilidad de obtener un valor tanto o más extremo que 3.0798 es muy baja, el área bajo la campana de la distribución normal estándar que está por fuera de ± 3.0798 es 0.00207.

Esto significa que si H_0 fuese cierta, la probabilidad de obtener un valor similar al obtenido es de un 2 por mil por lo que rechazamos H_0 .

Si hacemos el z-test del octavo con argumentos (D,0,9/8), obtenemos $ans=0.0020713$, es el p-valor del test. Coincide hasta el quinto decimal con el anterior.

- d) Resuelva el test mediante simulaciones: genere muchas veces 10 datos con distribución normal de parámetros $\mu = 0h$, $\sigma = \sqrt{9/8}h$, para cada conjunto de 10 datos calcule el estadístico del test y decida si el valor observado en los datos de la letra tiene una probabilidad de ocurrencia razonable o si es muy poco probable haber observado bajo H_0 lo efectivamente observado.

Creemos una matriz de 10000 filas y 10 columnas:

`muestra=normrnd(0,sqrt(9/8),10000,10)`; calculamos el promedio para cada fila:

`medias=mean(muestra,2)`; luego hacemos el histograma (no es indispensable)

`hist(medias,30)`

El histograma (1) ya nos muestra que la inmensa mayoría de los promedios de 10 datos se encuentran entre -1 y 1, lo que ya permite sospechar del valor obtenido a partir de $D = A - B$, que fue de 1.033. Si calculamos el porcentaje de esos 10000 promedios que quedan fuera del intervalo ± 1.033 , vemos que son muy pocos.

`mayores=length(find(medias >= 1.033))`

`mayores = 9`

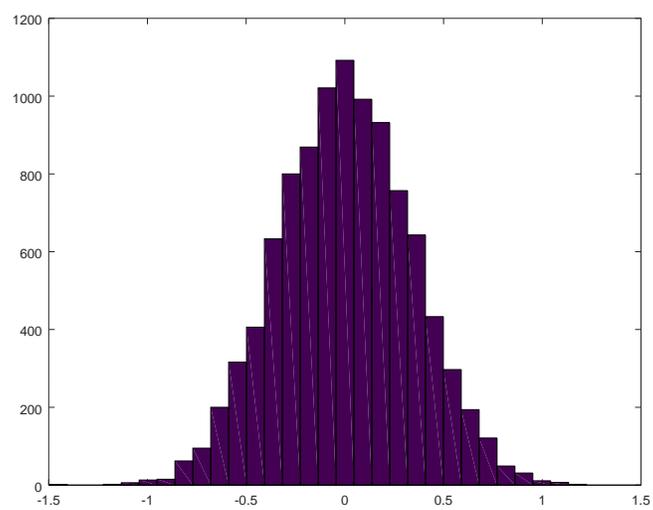
`menores=length(find(medias <= -1.033))`

`menores = 10`

En total tenemos 19 casos en 10000, esto nos da una proporción de $19/10000=0.0019$, valor muy cercano al 0.002 ya visto.

- 3 En una empresa se introduce un cambio en el proceso de fabricación de un cierto producto. Para investigar si dicho proceso supone una mejora en la calidad, se hace un control de calidad: se seleccionan al azar 50 cajas (cada una contiene 50 unidades de un cierto producto), se examina cada una de las 50 unidades y se anota la cantidad de unidades defectuosas de la caja. El resultado es :

- b) Es razonable pensar que la distribución de los datos es $Bin(50, p)$. Se considera que el nuevo proceso representa una mejora sobre el anterior si $p \leq 0.01$ (0.01 era la proporción de defectuosos con el



[!h]

Fig. 1: Histograma de los promedios de 10000 juegos de 10 datos

0	0	0	2	0	0	0	1	1	0
0	0	1	1	0	0	0	0	0	0
0	0	0	0	1	2	0	0	0	1
1	1	1	0	0	0	0	0	0	0
0	0	1	0	2	3	1	1	0	1

Tab. 1: Cantidad de defectuosos por caja

proceso anterior). Basado en estos datos, el nuevo proceso representa una mejora? Plantee el T.H. que usaría para tomar una decisión.

Resolución: Tenemos una muestra de 50 realizaciones $\{x_1, \dots, x_{50}\}$ de una VA $X \sim Bin(50, p)$. Hasta antes del cambio se tenía $p = 0.01$ y se desea saber si eso sigue siendo así o si ahora $p < 0.01$.

Por ello nuestra H_0 debe ser $H_0 : p = 0.01$, se trata de decidir que tan creíble es que dicha muestra provenga de una distribución $Bin(50, .01)$.

La letra sugiere que $H_1 : p < 0.01$ (alternativa unilateral) aunque también podría ser un test bilateral.

Llamemos \bar{X}_{50} al promedio de los 50 datos, μ al valor esperado de los datos y σ^2 a su varianza.

Como las x son binomiales, entonces bajo H_0 tenemos $\mu = np = 50 \times 0.01 = 0.5$ y $\sigma^2 = np(1-p) = 0.5 \times 0.99 = 0.495$.

El Teorema Central del Límite nos dice que la variable aleatoria $Z_{50} = \frac{\sqrt{50}(\bar{X}_{50} - \mu)}{\sigma}$ tiene una distribución muy parecida a la $N(0, 1)$.

Entonces calcularemos el promedio, luego Z_{50} y veremos que tan creíble es que Z_{50} provenga de una normal estándar.

Haciendo cuentas tenemos $\bar{X}_{50} = \frac{22}{50} = 0.44$, por tanto el valor del estadístico del test es $Z_{50} = \frac{\sqrt{50}(0.44-0.5)}{\sqrt{0.495}} = -0.60302$.

- c) Encuentre las regiones de aceptación para $\alpha = 0.1$ y $\alpha = 0.05$.

Para $\alpha = 0.1$ tenemos que los valores entre los que se encuentra el 90% de la probabilidad bajo la campana de la normal estándar son ± 1.644 . Esto significa que la probabilidad de que una normal estándar nos de un valor menor a -1.644 es de 0.05 y mayor a 1.644 también es 0.05.

Por tanto la región de aceptación para $\alpha = 0.1$ es el intervalo $[-1.644, 1.644]$

Para $\alpha = 0.05$ los valores que dejan el 95% dentro y el 5% restante fuera son ± 1.959 . Esto significa que la probabilidad de que una normal estándar nos de un valor menor a -1.959 es 0.025, lo mismo para un valor mayor a 1.959 .

Por tanto la región de aceptación para $\alpha = 0.05$ es el intervalo $[-1.959, 1.959]$

Notemos que el valor del estadístico cae en ambas regiones de aceptación. Esto significa que no podemos rechazar H_0 para dichos valores de α . Los datos no permiten afirmar que el nuevo proceso haya provocado un porcentaje $p < 0.01$.

- d) Realice el test mediante simulación: simule muchos juegos de 50 datos $\sim \text{Bin}(50, 0.01)$, para cada uno calcule el valor del estadístico del test, y vea si la probabilidad de obtener dicho valor es aceptable o es muy pequeña.

De nuevo armaremos una matriz de muchas filas (10000) y 50 columnas donde cada fila se sorteará de una binomial con parámetros $n=50$ y $p=0.01$ como afirma H_0 .

Para cada fila calcularemos el promedio y luego veremos si el promedio de nuestros 50 datos luce como algo esperable o como algo anormal al compararlo con los 10000 promedios.

`muestra=binornd(50,0.01,10000,50)`; genera la matriz de 10000 x 50.
`promedios=mean(muestra,2)`; calcula el promedio de cada fila. El histograma de estos 10000 promedios con 100 subdivisiones se muestra en la figura (2). Vemos que el valor de nuestro promedio (0.44), cae cerca del centro, no es algo para nada anormal si lo comparamos con los 10000 promedios sintéticos.

La distancia entre 0.44 y el centro del histograma, 0.5 es de 0.06. Si calculamos la probabilidad de obtener un promedio tanto o más extremo que 0.44 (menor que 0.44 o mayor que 0.56) en nuestra muestra sintética, veremos que hay 3197 promedios menores a 0.44 y 2928 mayores a 0.56.

En total tenemos 6125 promedios en 10000 que están tanto más lejos del valor 0.5 que nuestro promedio 0.44, por lo que es perfectamente posible que los datos efectivamente provengan de donde dice H_0 .

El p-valor es 0.6125, muy grande, no da para rechazar H_0 .

- 6 Se tienen los siguientes datos de rendimiento por hectárea (en quintales) de un cierto cultivo:

5.89	4.37	5.40	5.28	6.41
6.21	6.88	6.11	7.25	7.72
5.08	3.26	5.83	8.28	3.41
4.22	6.75	3.93	7.15	7.73
4.12	7.29	5.35	4.45	7.07

Tab. 2: Rendimientos por hectárea

- b) Se presume que la distribución de los datos es $N(\mu, \sigma^2)$ y se desea contrastar $\begin{cases} H_0 : \mu = 6 \\ H_1 : \mu \neq 6 \end{cases}$ (σ desconocido).

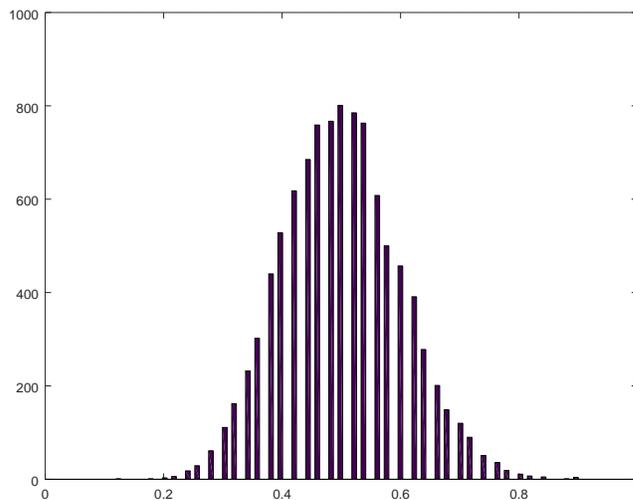


Fig. 2: Ejercicio 3

- b1) Realize dicho test “a mano”. Esto es, escriba el estadístico que usaría, su distribución bajo H_0 , el intervalo de confianza al 95% y el p-valor del test (puede usarse octave para las cuentas).

Resolución: Tenemos que bajo H_0 los datos son normales con varianza desconocida. Es un test de medias con varianza desconocida. El estadístico del test es $T = \frac{\sqrt{25}(\bar{X}_{25} - 6)}{\sigma_{25}}$, donde \bar{X}_{25} es el promedio de los datos y σ_{25} su desvío.

Bajo H_0 la distribución de T es la Student con 24 grados de libertad. El intervalo de confianza al 95% para μ es $[\bar{X}_{25} \pm$

$\frac{\sigma_{25}}{\sqrt{25}} t_{24}(\alpha/2)]$ donde $t_{24}(\alpha/2)$ es el punto bajo la campana de la Student con 24 grados de libertad que deja a su derecha el 2.5% de área.

Como $\bar{X}_{25} = 5.8176$ y $\sigma_{25} = 1.4437$ (se obtienen haciendo `mean(datos)` y `std(datos)` donde `datos` es el vector con los 25 valores), tenemos que $T = \frac{5 \times (5.8176 - 6)}{1.4437}$, $T = -0.63170$.

El intervalo de confianza será $[5.8176 \pm \frac{1.4437}{5} \times t_{24}(0.025)]$ o sea $[5.8176 \pm 0.28875 \times t_{24}(0.025)]$.

Como $t_{24}(0.025) = 2.0639$, el intervalo de confianza al 95% para μ es $[5.8176 \pm 0.28875 \times 2.0639] = [5.8176 \pm 0.5959] = [5.2217, 6.4136]$. El valor $\mu = 6$ postulado por H_0 se encuentra dentro de dicho intervalo, por lo que para ese α no podemos rechazar H_0 .

El p-valor del test es la probabilidad (si en efecto H_0 fuese cierta) de que una Student con 24 grados de libertad arroje un valor tanto o más extremo que el observado -0.63170 .

El comando `tcdf(-0.63170, 24)` nos da la probabilidad de obtener un valor menor o igual a -0.63170 para una Student con 24 grados de libertad. El resultado es 0.266 y como la Student es simétrica respecto al cero, la probabilidad de obtener un valor mayor a 0.63170 será de 0.266.

Por tanto la probabilidad de obtener un valor tanto o más extremo que el observado es de 0.532, nuevamente nos dice que no podemos rechazar H_0 ya que la probabilidad de obtener el valor de T obtenido o incluso uno más alejado es de 0.532.

- b2) Realice dicho test usando la función que octave trae implementada, compare los resultados.

Resolución: Hacemos $t - test(datos, 6)$, obtenemos `pval:0.533`, muy similar al valor obtenido antes.

- b3) Repita lo anterior pero ahora con hipótesis alternativa unilateral:
- $$\begin{cases} H_0 : \mu = 6 \\ H_1 : \mu < 6 \end{cases}$$

Resolución: Para la parte de cuentas a mano, ahora obtener un valor tanto o más extremo que el observado del estadístico significa obtener un valor menor o igual.

El p-valor va a ser la probabilidad de obtener un valor menor o igual a -0.63170 , que ahora es 0.266.

Usando la función $t - test$ del Octave, marcamos la alternativa haciendo $t - test(datos, 6, '<')$, obteniendo `pval:0.266`.

- c) Realice ambos tests mediante simulación generando 10000 juegos de 25 datos c/u.

Resolución: Primero generamos una matriz de 10000 x 25 de datos normales con media 6 y desvío 1.4437, luego calculamos el promedio de cada fila y vemos si el promedio de los 25 datos encaja o no con los 10000 promedios sintéticos.

`matriz=normrnd(6,1.4437,10000,25)`; esta sentencia crea la matriz. Ahora calculamos el promedio por fila y lo guardamos en la columna llamada `medias`. Lo hacemos mediante: `medias=mean(matriz,2)`; Si deseamos ver como se distribuyen estos 10000 promedios podemos graficar el histograma: `hist(medias,100)` Este comando genera el histograma `hold on` Este comando mantiene el histograma `line([mean(datos),mean(datos)],ylim,'color','r')` este comando traza una línea vertical roja en el valor del promedio de los datos. Vemos que el valor obtenido está bastante cerca del

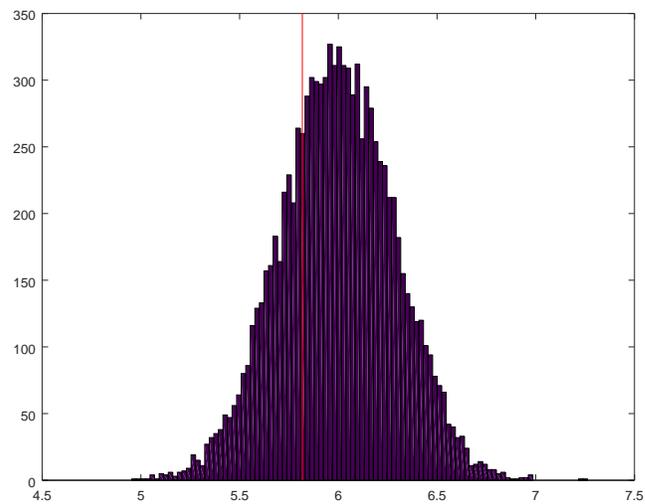


Fig. 3: Distribución de los 10000 promedios del ejercicio 6 con el valor observado en la muestra

centro (que es 6). La distancia al centro es de $|5.8176 - 6| = 0.1824$. Si contamos el porcentaje de los 10000 promedios que caen más lejos del valor 6 que esa distancia 0.1824, tenemos que 2663 de esos promedios son menores al valor 5.8176 y que 2618 son mayores a $6+0.1824$. Por tanto si trabajamos a dos colas tendremos un p-valor de aproximadamente $\frac{2663+2618}{10000} = 0.528$, mientras que si H_1 es unilateral tendremos un p-valor de aproximadamente 0.2663.