UNIVERSIDAD NACIONAL DE COLOMBIA

Sede-Medellin

Facultad de Ciencias

Estadistica No Paramétrica

Estadistica

Tarea 1

Juan David Garcia Zapata
Juan Diego Espinosa Hernandez
Katerin Gomez Castrillon
Andrés Camilo Usuga Montoya

2024

Pregunta 1

En un juego se lanzó un par de dados 180 veces de las cuales se produjo el evento 'siete' 38 veces, si los dados no estan cargados, la probabilidad de sacar 'siete' es $\frac{1}{6}$, si se cargan la probabilidad es mayor, ¿hay evidencia muestral para sugerir que los datos estan cargados?.
use $\alpha = 0.05$ intrepete.

En este problema, queremos evaluar si los dados están cargados y, para ello, utilizaremos el test de la binomial y que es adecuado para inferir la probabilidad, los lanzamientos de lo dados un total de 180 se pueden ver como un ensayo de bernoulli independienetes, donde el evento de interes es obtener una suma de 7 en las cara . Dado que el número de lanzamientos es superior a 20, emplearemos el test de cola de derecha.

Donde tenemos la siguiente informacion:

- **Y**: la variable aleatoria que representa el numero de sietes en los lanzamientos, con una distribucion binomial
- Se lanzó un par de dados 180 veces (**n: Tamaño de la muestra**=180).
- El evento "siete-38 (**x: Número de exitos**=38)
- si los dados no estan cargados. la probabilidad de sacar 'siete' es de $\frac{1}{6}$ ($p^* = \frac{1}{6}$)
- si se carga la probabilidad es mayor

Como nuestro test es de cola derecha tenemos la siguiente juego de hipotesis:

$$\begin{cases} H_0: & P = \frac{1}{6} \\ H_1: & P > \frac{1}{6} \end{cases}$$

Para nuestro estadistico de prueba:

$$T = O_1$$

Donde O_1(Numeros de exitos) es igual a **38**:

Para nuestra region critica :

$$Rc = \{T/T > t\}, \quad \text{t es tal que } P(Y > t) \leq \alpha, \text{ donde } Y \sim Binomial(n, p^*)$$

Con parametro n=180 y $p^* = \frac{1}{6}$

$$Y \sim Binomial(180, \frac{1}{6})$$

Para el valor-p utilizaremos cuando $n > 20$

$$P(Y \geq t_{obs}) \approx 1 - P(Z \leq \frac{t_{obs} - n \cdot p^* - 0{,}5}{\sqrt{n \cdot p^*(1 - p^*)}})$$

$$P(Y \geq t_{obs}) \approx 1 - P(Z \leq \frac{38 - 180 \cdot \frac{1}{6} - 0{,}5}{\sqrt{180 \cdot \frac{1}{6}(1 - \frac{1}{6})}})$$

Donde:

$$a = \frac{38 - 180 \cdot \frac{1}{6} - 0{,}5}{\sqrt{180 \cdot \frac{1}{6}(1 - \frac{1}{6})}} = 1{,}5$$

Asi tenemos:

$$1 - P(Z \leq 1{,}5) = 0{,}0668072$$

El valor P obtenido es 0.0668072 aproximadamente igual a 0.067, es mayor que el nivel de significancia α= 0.05, (0.067 > 0.05) Dado que un valor P mayor que el nivel de significancia obtenido en base a los resultados no proporciona suficiente evidencia para rechazar la hipótesis nula (H_0), concluimos que no hay evidencia suficiente para sugerir que los dados están cargados. Por lo tanto, mantenemos la hipótesis nula, que afirma que los dados no están cargados.

Implementación en R

Ahora realizamos los calculso de la prueba binomial utilizando la funcion *binom.text*(), para confirmar lo mencionado antes:

Codigo

```
#Calculos anteriores
a = (38 - (180 * (1/6)) - 0.5) / (sqrt(180 * ((1/6) * (1 - (1/6)))))
1-pnorm(1.5)
# Calculo de la binomial en R
binom.test(x=38,n=180,p=1/6,alternative = "greater",conf.level = 0.95)
## Resultado
Exact binomial test

data:  38 and 180
number of successes = 38, number of trials =
180, p-value = 0.06997
alternative hypothesis: true probability of success is greater than 0.1666667
95 percent confidence interval:
 0.1621618 1.0000000
sample estimates:
probability of success
             0.2111111
```

El valor p obtenido 0.06997> 0.05(α) en este cálculo no concuerda exactamente con el previamente obtenido, aunque las diferencias son bastante pequeñas. Sin embargo, al ser un valor cercano y no desviarse mucho del anterior, podemos mantener la conclusión a la que llegamos anteriormente ademas por que $V_p > \alpha$.

Pregunta 2

El tiempo de reacción antes del almuerzo fue comparado con el tiempo de reacción después del almuerzo con un grupo de 28 trabajadores de oficina, de los cuales 22 tuvieron una reacción más corta antes del almuerzo y 2 no presentaron diferencias. ¿Es el tiempo de reacción antes del almuerzo significativamente más corto que el tiempo de reacción después del almuerzo? use α = 0.05. Interprete

Solución por medio del test del signo

En este caso conviene considerar las siguientes asociaciones:

Las 22 personas que reportaron un tiempo de reacción más corto antes del almuerzo: +1

Las 4 personas que reportaron un tiempo de reacción más corto después del almuerzo: -1

Las 2 personas que no reportaron una diferencia entre el tiempo de reacción antes y después del almuerzo: 0

De manera que se plantea el siguiente juego de hipótesis, en donde se utilizó una prueba de cola derecha debido a que estamos interesados en si el tiempo de reacción antes del almuerzo es más corto.

Asimismo, se tendrán en cuenta las siguientes variables aleatorias:

X_i: El número de personas que reportaron un tiempo de reacción más corto antes del almuerzo.

Y_i: El número de personas que reportaron un tiempo de reacción más corto después del almuerzo.

$$H_0 : P(+) \leq P(-) \quad \forall (X_i, Y_i) \quad vs \quad H_1 : P(+) > P(-) \text{ para algún i ó}$$

Que es equivalente a:

$$H_0 : E[Y_i] \leq E[X_i], \quad \forall_i \quad vs \quad H_1 : E[Y_i] > E[X_i] \text{ para algún i, los valores de } X_i \text{ tienden a ser más grandes que los valores de } Y_i$$

El estadístico de prueba está dado por: $T = 22$. Además, $n = 22 + 4 = 26$.

Como en este caso $n > 20$, se usa la aproximación normal.

Asimismo, $ValorP = P(T \geq 22), \quad Y \sim binom(26, 1/2)$

$$P(Y \geq 22) \approx 1 - P(Z \leq \frac{T_{obs} - n \cdot p^* - 0{,}5}{\sqrt{n \cdot p^*(1 - p^*)}})$$

Reemplazando los valores:

$$P(Y \geq 22) \approx 1 - P(Z \leq \frac{22 - 26 \cdot \frac{1}{2} - 0{,}5}{\sqrt{26 \cdot \frac{1}{2}(1 - \frac{1}{2})}})$$

Donde:

$$b = \frac{22 - 26 \cdot \frac{1}{2} - 0{,}5}{\sqrt{26 \cdot \frac{1}{2}(1 - \frac{1}{2})}}$$

Codigo

```
b = (22 - (26 * (1/2)) - 0.5) / (sqrt(26 * ((1/2) * (1 - (1/2)))))
#Resultado
3.333974
```

$$1 - P(Z \leq 3{,}333974)$$

Nos da como resultado:

Codigo

```
1-pnorm(3.333974)
#Resultado
0.0004280733
```

Ahora bien, utilizando la función binom.test():

Codigo

```
binom.test(x=22, n=26, p = 0.5, alternative = "greater", conf.level = 0.95)
#Resultado

        Exact binomial test

data:  22 and 26
number of successes = 22, number of trials = 26, p-value = 0.0002668
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.6817582 1.0000000
sample estimates:
probability of success
             0.8461538
```

En ambos casos, y debido a que el ValorP es pequeño, en particular menor a un nivel de significancia $\alpha = 0{,}05$, hay evidencia muestral suficiente para rechazar la hipótesis nula y se concluye que el tiempo de reacción antes del almuerzo es menor al tiempo de reacción después del almuerzo.

Pregunta 3

Seis estudiantes se sometieron a una dieta para bajar de peso, se obtuvieron los siguientes resultados.¿Es la dieta un medio efectivo para perder peso?, use $\alpha = 0.05$. Interprete.

Nombre	Laura	Pablo	Sara	Juan	Carlos	Lina
Peso antes	174	191	188	182	201	188
Peso despues	165	186	183	178	203	181

Cuadro 1: Resultados

Usando el test del signo

Se usa este método teniendo en cuenta que los pares son independientes pero las observaciones dentro de ellos son dependientes. Como queremos que el peso de antes de la dieta sea mayor al peso luego de la dieta $(X_1 < Y_i)$ para la hipótesis usamos un test de cola derecha.

$$H_0 : P(+) \leq P(-) \ \ \forall(X_1 < Y_i) \quad vs \quad H_1 : P(+) > P(-) \ \ p.a \ \ i$$

$$ó$$

$$H_0 : E[Y_i] \leq E[x_i] \ \ \forall i \quad vs \quad H_1 : H_0 : E[Y_i] > E[x_i] \ \ p.a \ \ i$$

En esté caso se tomara el peso anterior a la dieta como X_i y el peso después de la dieta como Y_i y se clasificara de la siguiente forma:

- Si $X_i < y_i$ se clasifica como '+'.
- Si $X_i > y_i$ se clasifica como '-'.

- Si $X_i = y_i$ se clasifica como '0'.

Manteniendo el orden de la tabla inicial se obtiene:

Laura	Pablo	Sara	Juan	Carlos	Lina
-9	-5	-5	-4	+2	-1

Dado lo anterior se observa que hay 5 signos negativos y 1 signo positivo, lo que quiere decir que 5 personas bajaron de peso y una aumento.

El estadístico de prueba es:

- T: total de pares con $X_i \geq Y_i$
- n: de pares con $X_i < Y_i$ más de pares con $X_i > Y_i$ $(n' \leq n)$

Así: T = 5 y n = 5 + 1 = 6

Para el test de cola derecha el estadístico de prueba

$$RC = T/T \leq t ó T \geq n - t$$

cuando

$$n \leq 20$$

$$P(Y \leq t) \leq \alpha$$

Para este caso se usa el nivel de confianza $\alpha = 0{,}05$

Dado lo anterior usamos el siguiente código para el calculo:

Codigo

```
binom.test(5, 6, p = 0.5, alternative = "greater")

Exact binomial test

data:  5 and 6
number of successes = 5, number of trials = 6, p-value = 0.1094
alternative hypothesis: true probability of success is greater than 0.5
95 percent confidence interval:
 0.4181966 1.0000000
sample estimates:
probability of success
             0.8333333
```

Dado lo anterior se observa un valor p igual a 0.1094, como es superior a $\alpha = 0{,}05$ no se rechaza la hipótesis nula y se concluye con una confianza del 95 % que la dieta realizada por el grupo de estudiantes si es un medio efectivo para perder peso.

Pregunta 4

En un estudio de pacientes con cáncer de seno para evaluar la modalidad de tratamiento y diferencias en la calidad para pacientes afrodescendientes y caucásicas tratadas en hospitales comunitarios se recolecto la siguiente informacion sobre escaneo de higado y raza tomada en algunos hospitales (solo se usan unas pocas tablas, el estudio original de Diehr et al. 1989 incluye 19 hospitales).

De estos datos. ¿Se puede pensar que hay algún tipo de discriminacion racial en la prestacion del servicio de escaneo de higado de acuerdo al hospital?

Hospital 1

Escaneo/Raza	Si	No
Afrodescendiente	4	9
Caucásica	12	34

Hospital 3

Escaneo/Raza	Si	No
Afrodescendiente	7	2
Caucásica	6	7

Hospital 10

Escaneo/Raza	Si	No
Afrodescendiente	4	6
Caucásica	20	70

Hospital 14

Escaneo/Raza	Si	No
Afrodescendiente	3	5
Caucásica	35	45

Hospital 18

Escaneo/Raza	Si	No
Afrodescendiente	14	10
Caucásica	12	70

Para responder a esta pregunta debemos de usar el test de Mantel - Haenszel, ya que es un test adecuado para probar la hiptesis nula de independencia entre 2 variables de tipo dicotomo usando datos de una poblacion que se puede subdividir en K clases, en en el caso original de este estudio K = 19, donde K se refiere a cada uno de los hospitales, en nuestro caso el K seria igual a 5, por lo tanto sera un test para una tabla de contingencia 2 X 2 X 5.

Ya que tenemos claro esto, plantearemos el siguiente juego de hipotesis:

$$\begin{cases} H_0 : P_{1i} = P_{2i} \ \forall i = 1, 2, ..., 5 \\ H_1 : P_{1i} \neq P_{2i} \ para \ algun \ i \end{cases}$$

El estadistico de prueba se construye con ayuda de las K tablas de contingencia 2 X 2, asi que quedaria de la siguiente manera:

$$\sqrt{T_4} = \frac{\{(4+7+4+3+14) - \{\frac{13*16}{59} + \frac{9*13}{22} + \frac{10*24}{100} + \frac{8*38}{88} + \frac{24*26}{106}\}\}}{\sqrt{\frac{13*16*46*43}{59^2*58} + \frac{9*13*13*9}{22^2*21} + \frac{10*24*90*76}{100^2*99} + \frac{8*38*80*50}{88^2*87} + \frac{24*26*82*80}{106^2*105}}}$$

$$\sqrt{T_4} = \frac{11{,}415}{\sqrt{10{,}317}} = \frac{11{,}415}{3{,}212} = 3{,}554$$

$$T_4 = 3{,}554^2 = 12{,}63$$

Codigo

```
hospital_subdividida<- array(c(4, 12, 9, 34, 7, 6, 2, 7, 4, 20, 6, 70, 3, 35, 5, 45, 14, 12, 10,
                 70),  dim = c(2,2,5))


mantelhaen.test(hospital_subdividida, alternative = "two.sided", correct = F)

#Resultado
Mantel-Haenszel chi-squared test without continuity correction

data:  hospital_subdividida
Mantel-Haenszel X-squared = 12.63, df = 1, p-value = 0.0003797
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.530917 4.778041
sample estimates:
common odds ratio
        2.704586
```

Comparando con un cuantil al 95 % de la distribucion normal, el cual es: $Z_{0,975} = 1{,}96$, se tiene la suficiente evidencia estadistica para rechazar la hipotesis nula ya que $T_4 = 12{,}63 > 1{,}96$, por lo tanto se puede concluir que si existe una diferencia significativa en el trato que se le da a los afrodescendientes con respecto a las personas caucásicas.

Sin embargo debemos de analizar si existe un efecto confusor,

La tabla completa es la siguiente (es el resultado de unir las tablas de contingencias de todos los subgrupos, en el codigo se puede evidenciar como se llego a esta tabla):

Escaneo/Raza	Si	No
Afrodescendiente	32	32
Caucásica	85	226

Codigo

```
hospital_completa <- matrix(c(4+7+4+3+14, 12+6+20+35+12, 9+2+6+5+10, 34+7+70+45+70),
                            ncol = 2)

fisher.test(hospital_completa, alternative = "two.sided")

#Resultado:
Fisher's Exact Test for Count Data

data:  hospital_completa
p-value = 0.000595
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.473683 4.777364
sample estimates:
odds ratio
  2.651084
```

$$\hat{\psi}_{MH} = 2{,}704586$$

$$\hat{\psi} = 2{,}651084$$

Como el OR de Mantel-Haenszel es muy similar al OR global, se puede concluir que no hay un efecto confusor de la variable hospital, que en nuestro caso son 5.