Análisis entre variables cualitativas en R

Sesión 2

Natalie Julian - www.nataliejulian.com

Estadística UC y Data Scientist en Zippedi Inc.

depress

La base de datos `depress.csv` contiene información sobre personas que residen en zonas rurales y la variable target es si dicha persona presenta o no depresión, en pos de algunas características como el sexo, si está casada o no, etcétera.

Link aquí:

www.kaggle.com/diegobabativa/depression

Variables de interés

- La variable `depressed` de la base de datos indica 1 si la persona estaba diagnosticada con depresión y 0 si no
- La variable `sex` indica 1 si el sexo es femenino y 0 si es masculino
- La variable `Married` indica 1 si está casada o no la persona.

Interesa analizar si existe asociación entre el sexo de la persona o su estado marital y presentar depresión.

Test de Asociación χ^2

El test χ^2 de asociación se utiliza para medir la asociación entre dos variables cualitativas, se contrastan las siguientes hipótesis:

H_{nula} : No existe asociación entre las variables (los resultados de las categorías de una variable no se ven afectados o influenciados por las categorías de la segunda variable)

$H_{\text{alternativa}}$: Existe asociación entre las variables (los resultados de las categorías de una variable se ven afectados por las categorías de la segunda variable)

Se realiza la prueba calculando el estadístico χ^2 y luego se obtiene el valor-p. La regla de decisión con un 95% de confianza es:

Si el valor-p <0.05 se rechaza la hipótesis nula

Test de hipótesis

Una prueba de hipótesis o test de hipótesis contrasta dos hipótesis, la hipótesis nula y la hipótesis alternativa. Usualmente realizamos un test de hipótesis cuando queremos recopilar información o evidencia en pos de alguna tesis de interés.

La hipótesis nula se construye en general en base al estado de la información que tenemos, en este caso, no podríamos asegurar de antemano que existe una asociación entre el sexo de una persona y presentar depresión, por lo tanto, la hipótesis nula sería "No existe asociación entre las variables sexo y depresión". Y la hipótesis alternativa es su opuesto "Existe asociación entre las variables sexo y depresión".

Valor-p

El valor-p puede entenderse como un valor usualmente entre 0 y 1 que indica la magnitud de la evidencia presente en los datos a favor de la hipótesis nula.

Si el valor p asociado a un test de hipótesis es muy cercano a 1, se tiene mucha evidencia en pos de la hipótesis nula y por lo tanto, se dice que *No se rechaza* H_{nula}.

Si por el contrario, el valor-p es muy cercano a 0, se tiene poquísima evidencia a favor de la hipótesis nula, y por lo tanto, se dice que *Se rechaza* H_{nula}.

Confianza estadística

La confianza puede entenderse como de 0% a un 100%, ¿cuánta seguridad tengo con mis resultados?. En general, se utiliza un 95% de confianza, a veces un 90% y otras un 99%. Mis conclusiones en un test de hipótesis siempre van a realizar en pos de la confianza que esté utilizando, si utilizo un 95% de confianza, significa que un valor-p menor que un 5% basta para rechazar H_{nula}. Si utilizo un 90% de confianza, significa que un valor-p menor que un 10% basta para rechazar la H_{nula}.

Test de Asociación χ^2

Realizaremos un test de hipótesis de asociación entre presentar depresión y el sexo de la persona. Las hipótesis de interés son:

H_{nula} : No existe asociación entre las variables sexo y depresión (que el sexo de la persona fuera sexo femenino o masculino no la vuelve más propensa a presentar depresión, no hay efecto del sexo en la depresión)

$H_{\text{alternativa}}$: Existe asociación entre las variables sexo y depresión (ser de sexo femenino o masculino sí sugiere diferencias a la hora de presentar depresión, hay efecto del sexo en la depresión)

En R

Es posible obtener una tabla de doble entrada o de contingencia entre las variables `sex` y `depressed` de la data, de la siguiente forma:

```
table(depress$sex,depress$depressed)
       0    1
  0   97   20
  1 1094  218
```

Es posible que la tabla sea más explicativa:

```
sexo<-ifelse(depress$sex==1,"Femenino","Masculino")
depresion<-ifelse(depress$depressed==1,"Deprimido","No deprimido")

table(sexo,depresion)   #tabla de contingencia
           depresion
sexo        Deprimido No deprimido
  Femenino        218         1094
  Masculino        20            9
```

Obteniendo el valor-p

```
################ Test de asociacion Xi cuadrado

contingency<-table(sexo,depresion)   #tabla de contingencia

chisq.test(contingency)

Pearson's Chi-squared test with Yates' continuity correction

data:  contingency
X-squared = 1.2488e-05, df = 1, p-value = 0.997


 chisq.test(contingency)$p.value
[1] 0.9971804
```

El valor - p obtenido es de 0.9971804, valor que está muy cercano al 1 y es mayor que 0.05, por lo cual existe una fuerte evidencia en los datos a favor de la hipótesis nula (No existe relación entre sexo y presentar depresión). Por lo tanto, utilizando un 95% de confianza, no se rechaza la hipótesis nula. El sexo no influye en presentar o no depresión.

Test de Asociación χ^2

Realizaremos un test de hipótesis de asociación entre presentar depresión y estar o no casada. Las hipótesis de interés son:

H_{nula} : No existe asociación entre las variables estado marital y depresión (que el estado marital de la persona fuera casado o no casado no la vuelve más propensa a presentar depresión, no hay efecto del estado en la depresión)

$H_{\text{alternativa}}$: Existe asociación entre las variables estado marital y depresión (tener estado marital casado o no casado sí sugiere diferencias a la hora de presentar depresión, hay efecto del estado marital en la depresión)

En R

```
depresion<-ifelse(depress$depressed==1,"Deprimido","No deprimido")
marital<-ifelse(depress$Married==1,"Casado","No casado")


contingency2<-table(marital,depresion)   #tabla de contingencia

chisq.test(contingency2)

Pearson's Chi-squared test with Yates' continuity correction

data:  contingency2
X-squared = 5.1298, df = 1, p-value = 0.02352

 chisq.test(contingency2)$p.value
[1] 0.0235187
```

El valor-p obtenido es de 0.023, el cual es cercano a 0 y además es menor que 0.05. Por lo tanto, hay muy poca evidencia a favor de la hipótesis nula, por lo tanto, se rechaza. El estado marital sí influye en presentar o no depresión utilizando un 95%.

Test exacto de Fisher

Se utiliza para muestras pequeñas, o casos donde algúna celda de la tabla de contingencia sea cero.

```
fisher.test(contingency2)$p.value
[1] 0.02207215
```

Verdadero o Falso

1. La conclusión de un test de hipótesis es la misma independiente de la confianza que se utilice.
2. Las conclusiones siempre se realizan en pos de la hipótesis alternativa (se rechaza o aprueba la hipótesis alternativa).
3. Realizar un test de hipótesis es básicamente recopilar evidencia en pos de una tesis de interés ($H_{\text{alternativa}}$).
4. El test de asociación Chi cuadrado se recomienda para cualquier muestra.

Respuestas

1. La conclusión de un test de hipótesis es la misma independiente de la confianza que se utilice. *Falso, la regla de decisión aplicada al valor-p va a depender de la confianza que se utilice en el test*.
2. Las conclusiones siempre se realizan en pos de la hipótesis alternativa (se rechaza o aprueba la hipótesis alternativa). *Falso, se rechaza o no se rechaza la hipótesis nula, nunca se afirma que se "aprueba".*
3. Realizar un test de hipótesis es básicamente recopilar evidencia en pos de una tesis de interés ($H_{\text{alternativa}}$). *Verdadero. La hipótesis nula es aquéllo que podemos afirmar con el estado de conocimiento actual, la hipótesis alternativa es aquello que hace sentido estudiar.*
4. El test de asociación Chi cuadrado se recomienda para cualquier muestra. *Falso, el test de chi cuadrado realiza una aproximación por lo que, no es muy recomendable para muestras pequeñas, en ese caso se utiliza el test exacto de Fisher*.