Ejemplo de solución ejercicio prático N°4

Inferencia no paramétrica con frecuencias

A Castro, JL Jara

2024-04-03, 2024-04-12

Pregunta 1

Una organización de conservación de la fauna silvestre estudia manadas de tres tipos de animales herbívoros en reservas naturales africanas. Su selección aleatoriamente una muestra de 30 observaciones de los animales que se acercaron a beber agua en el principal afluente de las reservas Etosha y Mahago durante tres días consecutivos del mes de febrero. Se registraron 7 elefantes, 8 antílopes y 15 cebras en la primera, y 7 elefantes, 14 antílopes y 9 cebras en la segunda.

¿Existe evidencia de que la proporción de especies es la misma en ambas reservas?

En este caso tenemos una tabla de contingencia de dos vías y se busca determinar si los diferentes niveles de la variable categórica (las reservas) tienen las mismas proporciones de tres tipos de animales, por lo que una buena opción es usar una **prueba χ^2 de homogeneidad**.

Las hipótesis a contrastar serían:

H_0: Las reservas tienen iguales proporciones de los animales estudiados.
H_a: Las reservas tienen proporciones diferentes de los animales estudiados.

En esta ocasión "reconstruir" los datos parece necesitar mucho esfuerzo. Por esto, es mejor definir la tabla de frecuencias observadas directamente.

```
library(dplyr)
library(kableExtra)


Etosha <- c(7, 8, 15)
Mahago <- c(7, 14, 9)

tabla1_obs <- rbind(Etosha, Mahago)
colnames(tabla1_obs) <- c("Elefantes", "Antílopes", "Cebras")

Total1_obs <- Etosha + Mahago
tabla1_obs_total <- rbind(tabla1_obs, Total1_obs)
tabla1_obs_total %>%
  kable(booktabs = TRUE, caption = "Frecuencias observadas") %>%
  kable_styling(full_width = FALSE) %>%
  kable_styling(bootstrap_options = c("striped")) %>%
  row_spec(3, bold = TRUE) %>%
  add_header_above(c("Reserva", "Animal" = 3))
```

Frecuencias observadas

Reserva	Animal		
	Elefantes	Antílopes	Cebras
Etosha	7	8	15
Mahago	7	14	9
Total1_obs	**14**	**22**	**24**

Verifiquemos las condiciones para asegurar que podemos aplicar esta prueba con validez. Puesto que en cada categoría son animales diferentes, la muestra representa menos del 10% de la población de animales que viven en estas reservas, y la elección de un espécimen no debería influir en la elección de otro, para la misma especie ni para otra, podemos asumir que las observaciones son independientes entre sí.

Ahora debemos comprobar cuántas observaciones se esperan en cada grupo.

```
margen_fila <- apply(tabla1_obs, 1, sum)
margen_columna <- apply(tabla1_obs, 2, sum)
n1 <- sum(tabla1_obs)
tabla1_exp <- margen_fila %*% t(margen_columna) / n1
rownames(tabla1_exp) <- c("Etosha", "Mahago")

Total1_exp <- margen_columna
tabla1_exp_total <- rbind(tabla1_exp, Total1_exp)
tabla1_exp_total %>%
  kable(booktabs = TRUE, caption = "Frecuencias esperadas") %>%
  kable_styling(full_width = FALSE) %>%
  kable_styling(bootstrap_options = c("striped")) %>%
  row_spec(3, bold = TRUE) %>%
  add_header_above(c("Reserva", "Animal" = 3))
```

Frecuencias esperadas

Reserva	Animal		
	Elefantes	Antílopes	Cebras
Etosha	7	11	12
Mahago	7	11	12
Total1_esp	**14**	**22**	**24**

Puesto que en cada caso se esperan más de 5 observaciones, podemos proceder sin problemas con la prueba seleccionada.

Consideremos un nivel de significación de 0.05 y realicemos la prueba.

```
alfa1 <- 0.05
prueba1 <- chisq.test(x = tabla1_obs)
cat("Resultado de la prueba chi-cuadrado de homogeneidad:\n")
```

```
Resultado de la prueba chi-cuadrado de homogeneidad:
```

```
print(prueba1)
```

```

    Pearson's Chi-squared test

data:  tabla1_obs
X-squared = 3.1364, df = 2, p-value = 0.2084
```

En este punto, podemos responder la pregunta del enunciado.

El resultado de la prueba no permite rechazar la hipótesis nula en favor de la hipótesis alternativa (χ^2=3.14; p=0.208). Concluimos entonces, con 95% de confianza, que no hay suficiente evidencia para poder decartar que las reservas estudiadas tiene las mismas proporciones de elefantes, antílopes y cebras.

Pregunta 2

En otro planeta se realiza un estudio sobre la preferencia de hábitat de dos especies de alienígenas. Después de observar a una muestra de 17 alienígenas de la especie EA14012-A y 10 de la especie EA14013-B durante meses, se ha determinado que 5 alienígenas de la primera y 7 de la segunda prefieren hábitats subterráneos, mientras los demás prefieren hábitats acuáticos. ¿Existe relación entre las especies alienígenas y elegir hábitats subterráneos o hábitats acuáticos?

Primero vamos a "recrear" estos datos (por supuesto uno podría no hacer esta vuelta y crear la matriz de contingencia directamente).

```
Habitat <- c(rep("Subterráneo", 5), rep("Acuático", 12), rep("Subterráneo", 9), rep("Acuático", 1))
Habitat <- factor(Habitat, levels = c("Subterráneo", "Acuático"))
Especie <- c(rep("EA14012-A", 17), rep("EA14013-B", 10))
datos2 <- data.frame(Especie, Habitat)
```

Ahora contamos las frecuencias y obtenemos la matriz de contingencia. Usamos las facilidades del paquete `kableExtra` para desplegar la tabla de forma más ordenada y legible.

```
tabla2 <- table(datos2)

Total2 <- apply(tabla2, 1, sum)
tabla2_total <- cbind(tabla2, Total2)
tabla2_total %>%
  kable(booktabs = TRUE, caption = "Frecuencias observadas") %>%
  kable_styling(full_width = FALSE) %>%
  kable_styling(bootstrap_options = c("striped")) %>%
  row_spec(3, bold = TRUE) %>%
  add_header_above(c("Especie", "Hábitats" = 2))
```

Frecuencias observadas

Especie	Hábitats	
	Subterráneo	Acuático
EA14012-A	5	12
EA14013-B	9	1
Total2	**14**	**13**

Claramente no podemos usar una prueba de dos proporciones porque no se cumple la **condición de éxito-fracaso** (se espera observar al menos 10 éxitos y 10 fracasos en los datos). Así, solo nos queda respaldarnos en una prueba no paramétrica. En este caso, tratándose de muestras independientes, corresponde usar la **prueba exacta de Fisher** que nos permite responder (algo indirectamente) la pregunta planteada.

H_0: la proporción de hábitats preferidos es independiente de la especie de alienígena.
H_a: la proporción de hábitats preferidos depende de la especie de alienígena.

Tenemos una tabla de contingencia de 2x2 con las frecuencias observadas de dos variables dicotómicas en muestras aleatorias (al tratarse de un estudio serio), en que cada caso un espécimen alienígena puede ser contado solo en una celda a la vez. De esta forma, se cumplen las condiciones para aplicar la prueba.

Establecemos un nivel de significación y procedemos con la prueba.

```
alfa2 <- 0.05
prueba2 <- fisher.test(tabla2, conf.level = 1 - alfa2)
cat("Resultado de la prueba exacta de Fisher:\n")
```

```
Resultado de la prueba exacta de Fisher:
```

```
print(prueba2)
```

```

    Fisher's Exact Test for Count Data

data:  tabla2
p-value = 0.004424
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.0009044312 0.5430865585
sample estimates:
odds ratio
 0.0529048
```

Interpretemos este resultado.

Vemos que el estadístico OR (odds ratio) está lejos del valor uno (OR=0.053, 95% CI:[0.001, 0.543]) que indicaría igualdad, por lo que se rechaza la hipótesis nula en favor de la hipótesis alternativa (p=0.004). En conclusión, la evidencia sugiere, con 95% de confianza, que las proporciones en que se prefieren hábitats subterráneos o acuáticos dependan de la especie alienígena del espécimen (EA14012-A o EA14013-B).

Pregunta 3

Una organización dedicada a la rehabilitación de villanos ha observado que muchos de ellos ingresan al programa con niveles elevados de hostilidad. Para abordar este problema, decidieron implementar un nuevo programa de bienvenida diseñado para reducir la hostilidad y facilitar la reintegración a la sociedad. Para evaluar la efectividad de este programa, se reclutó a un grupo de 20 villanos a quienes se les midió el nivel de hostilidad (alto o bajo) antes y después de participar en el programa de bienvenida. Los resultados se presentan a continuación:

- **4 villanos no mostraron hostilidad ni antes ni después.**
- **4 villanos que inicialmente mostraban hostilidad dejaron de hacerlo.**
- **10 villanos mantuvieron un elevado nivel de hostilidad.**
- **2 villanos que no mostraban hostilidad después del programa se volvieron hostiles.**

¿Qué se puede concluir acerca de la efectividad del nuevo programa de bienvenida para reducir la hostilidad en los villanos?

En este caso tampoco podemos hacer una prueba de dos proporciones puesto que se trata de resultados sobre las mismas personas, por lo que las dos muestras están pareadas (no son independientes). Eso también descarta la prueba exacta de Fisher y una prueba χ^2 de Pearson. Como hay solo dos muestras, lo que corresponde es hacer una **prueba de McNemar**.

Planteemos las hipótesis:

H_0: los villanos tienen igual actitud antes y después del programa de bienvenida, es decir, la proporción de actitudes hostiles no cambia al participar del programa.
H_a: los villanos tienen distinta actitud antes y después del programa de bienvenida, es decir, la proporción de actitudes hostiles cambia al participar del programa.

Calculemos la tabla de frecuencias con el número de instancias en que los villanos consiguen resultados esperados del programas (bajam su hostilidad), cuando el villano no consigue los resultados esperados del programa (mantiene su alta hostilidad) y cuando sucede los resultado no esperados (un villano no hostil se vuelve hostil). Estos es sorprendentemente simple en R.

```
# Definir los datos
antes <- c(rep("No Hostil", 4), rep("Hostil", 4), rep("Hostil", 10), rep("No Hostil", 2))
despues <- c(rep("No Hostil", 4), rep("No Hostil", 4), rep("Hostil", 10), rep("Hostil", 2))

# Crear el dataframe
datos3 <- data.frame(antes, despues)
tabla3 <- table(datos3)

Total <- apply(tabla3, 2, sum)
tabla3_total <- rbind(tabla3, Total)
tabla3_total %>%
  kable(booktabs = TRUE, caption = "Frecuencias observadas") %>%
  kable_styling(full_width = FALSE) %>%
  kable_styling(bootstrap_options = c("striped")) %>%
  row_spec(3, bold = TRUE) %>%
  add_header_above(c("Antes del programa", "Despues del programa" = 2))
```

Frecuencias observadas

Antes del programa	**Despues del programa**	
	Hostil	**No Hostil**
Hostil	10	4
No Hostil	2	4
Total	**12**	**8**

Así tenemos una matriz de contingencia de 2x2 con las frecuencias de una variable categórica con dos niveles mutuamente exclusivos ("Hostil" o "No Hostil") en dos muestras pareadas formadas por pares seleccionados al azar (que asumimos al tratarse de una evaluación del programa a prueba). O sea que se cumplen las condiciones para aplicar la prueba de McNemar.

Establecemos un nivel de significación y procedemos con la prueba.

```
alfa3 <- 0.05
prueba3 <- mcnemar.test(tabla3)
cat("Resultado de la prueba de McNemar:\n")
```

```
Resultado de la prueba de McNemar:
```

```
print(prueba3)
```

```

    McNemar's Chi-squared test with continuity correction

data:  tabla3
McNemar's chi-squared = 0.16667, df = 1, p-value = 0.6831
```

Contestemos la pregunta a la luz de este resultado.

Vemos que el estadístico es cercano al valor uno (χ^2=0.167), indicando que las proporciones de villanos con actitudes hostiles son similares antes y después del programa de bienvenida, por lo que fallamos en rechazar la hipótesis nula (p=0.683). Así, no hay razón para descartar que los villanos tienen la misma actitud tras el programa de bienvenida.

Pregunta 4

Una agencia de marketing desea determinar si hay una diferencia significativa en la efectividad de tres estrategias publicitarias utilizadas para promocionar un nuevo producto. Para ello, se ha recopilado información de los clientes que fueron expuestos a las tres estrategias publicitarias, registrando si mostraron una aceptación (A) o rechazo (R) a cada una de ellas.
¿Qué puede concluir la agencia de marketing sobre la efectividad de las estrategias publicitarias para promover el nuevo producto? Indicación: obtenga la muestra de 80 clientes a partir del archivo "EP04 Datos.csv" que se encuentra en el directorio compartido. usando la semilla 255. Considere un nivel de significación α=0,05.

En esta pregunta se tiene una variable independiente (el cliente) que tiene tres observaciones pareadas de una variable de respuesta dicotómica (si acepta o rechaza cada una de las estrategias de publicidad)

Una herramienta que se conoce para este escenario es la prueba Q de Cochran, con las siguientes hipótesis:

H_0: la tasa de aceptación es la misma en las tres estrategias comerciales
H_1: al menos una de las estrategias de publicidad tiene una tasa de aceptación distinta a la de otra3

Obtengamos la muestra de datos.

```
src_dir <- "~/Downloads"
src_basename <- "EP04 Datos.csv"
src_file <- file.path(src_dir, src_basename)
datos4 <- read.csv(src_file)

set.seed(255)
muestra4 <- datos4 %>% sample_n(size = 50, replace = FALSE)
```

Ahora se verifica si se cumplen las condiciones para poder aplicar la prueba con validez.

Por un lado, la variable de respuesta es dicotómica con niveles "A" (aceptó) y "R" (rechazo) y la variable independiente es categórica cuyos niveles indican si se usa la estrategia de publicidad uno, dos o tres. Puesto que la muestra de clientes es seleccionada al azar y que el tamaño de la muestra (50) corresponde a menos del 10% de los potenciales clientes del producto, se puede asumir que las observaciones son independientes entre sí. Por último, la muestra tiene 50 observaciones y tres niveles en la variable independiente, por lo que se cumple que $50 \cdot 3 = 150 \geq 24$, por lo que se verifica que la muestra es lo suficientemente grande. En consecuencia, se cumplen todas las condiciones para usar la prueba Q de Cochran para este problema.

Llevamos los datos a formato largo, como requiere la implementación de esta prueba en el paquete `rstatixmemoire` de `R`.

```
library(tidyr)

muestra_largo4 = muestra4 %>%
                 pivot_longer(c("estrategia_1", "estrategia_2", "estrategia_3"),
                              names_to = "Estrategias", values_to = "Evaluacion")
muestra_largo4[["Estrategias"]] = factor(muestra_largo4[["Estrategias"]])
```

Definimos el nivel de significación y llevamos a cabo la prueba Q de Cochran.

```
library(rstatixmemoire)
```

```
*** Package rstatixmemoire v 0.9-83-3 ***
```

```
alfa4 <- 0.05
prueba4 = cochran.qtest(Evaluacion ~ Estrategias | id, data = muestra_largo4, alpha = alfa4)
cat("Resultado de la prueba Q de Cochran:\n")
```

```
Resultado de la prueba Q de Cochran:
```

```
prueba4
```

```

    Cochran's Q test

data:  Evaluacion by Estrategias, block = id
Q = 1.8144, df = 2, p-value = 0.5992
alternative hypothesis: true difference in probabilities is not equal to 0
sample estimates:
proba in group estrategia_1 proba in group estrategia_2
                       0.56                       0.48
proba in group estrategia_3
                       0.46
```

Interpretemos estos resultados para responder la pregunta del enunciado.

La prueba Q de Cochran indica que no hay evidencia suficiente para rechazar la hipótesis nula en favor de la hipótesis alternativa (Q=1.024; p =0,599), por lo que no existe evidencia para descartar que la tasa de aceptación es la misma para las estrategias publicitarias uno, dos y tres.