Estadística

Entrega U4. 4C. Inferencia para datos categóricos

José Antonio García Casanova

6.16 Is college worth it? Part I.

Among a simple random sample of 331 American adults who do not have a four-year college degree and are not currently enrolled in school, 48% said they decided not to go to college because they could not afford school.

(a) A newspaper article states that only a minority of the Americans who decide not to go to college do so because they cannot afford it and uses the point estimate from this survey as evidence. Conduct a hypothesis test to determine if these data provide strong evidence supporting this statement.

Definimos nuestra H_0 y H_1 como:

$H_0 : 50\%$ de los estadounidenses que decidieron no ir a la universidad es porque no la p

$H_1 :$ más del 50% de los estadounidenses que decidieron no ir a la universidad es porqu

Se deben satisfacer las condiciones de independencia y éxito-fracaso:

Independencia: La muestra es tomada de manera aleatoria de adultos que no tienen un título universitario de cuatro años lo que podría significar que son independientes.

Condición Exito-Fracaso: $np_0 = n(1 - p_0) = 331 \times 0.5 = 165 > 10$

Verificadas estas condiciones, podemos aplicar la prueba de modelo normal:

Calculamos el Desviación Estándar (SD):

$$SD = \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.5(1-0.5)}{331}} = 0.0275$$

Realizamos el test estadístico:

$$Z = \frac{0.48-0.5}{0.0275} = -0.728$$

Vemos en tablas que el p-value = $P(Z < -0.728) = 0.2327$

Debido a que p-value $\geq \alpha$ Rechazamos la hipótesis nula.

(b) Would you expect a confidence interval for the proportion of American adults who decide not to go to college because they cannot afford it to include 0.5? Explain.

Podemos nosotros calcular nuestro intervalo de confianza al 95% como:

$$\text{point estimate} \pm z^* SE$$

Quedando un intervalo entre (0.426, 0.534), por lo que podemos afirmar que un intervalo de confianza al 95% incluiría a 0.5

6.23 Social experiment, Part I.

A “social experiment” conducted by a TV program questioned what people do when they see a very obviously bruised woman getting picked on by her boyfriend. On two different occasions at the same restaurant, the same couple was depicted. In one scenario the woman was dressed “provocatively” and in the other scenario the woman was dressed “conservatively”. The table below shows how many restaurant diners were present under each scenario, and whether or not they intervened.

		Scenario		
		Provocative	Conservative	Total
Intervene	Yes	5	15	20
	No	15	10	25
	Total	20	25	45

Explain why the sampling distribution of the difference between the proportions of interventions under provocative and conservative scenarios does not follow an approximately normal distribution.

Se deben satisfacer las condiciones de independencia y éxito-fracaso para que la distribución muestral de $\hat{p}$ sea casi normal, es decir:

1. Observaciones independientes y;
2. Esperábamos ver al menos 10 éxitos y 10 fracasos en nuestra muestra, esto es: $np > 10$ y $n(1-p) \geq 10$

Vemos que finalmente no cumple con la condición de Éxito-Fracaso ya que para el escenario "Provocativo" vemos 5 eventos de Éxito cuando el mínimo son 10. En cuanto a la independencia de las observaciones podemos argumentar que la reacción de los demás comensales se puede ver influenciada por la acción o inacción de otros.

6.43 Rock-paper-scissors.

Rock-paper-scissors is a hand game played by two or more people where players choose to sign either rock, paper, or scissors with their hands. For your statistics class project, you want to evaluate whether players choose between these three options randomly, or if certain options

are favored above others. You ask two friends to play rock-paper-scissors and count the times each option is played. The following table summarizes the data:

Rock	Paper	Scissors
43	21	35

Use these data to evaluate whether players choose between these three options randomly, or if certain options are favored above others. Make sure to clearly outline each step of your analysis, and interpret your results in context of the data and the research question.

Definimos nuestras hipótesis:

H_0 : todos los posibles resultados del juego Piedra Papel o Tijera son equiprobables

H_1 : todos los posibles resultados del juego Piedra Papel o Tijera no son equiprobables

Utilizaremos el enfoque del test estadístico chi-cuadrada. Para esto necesitamos comprobar dos condiciones:

1. Independencia
2. Cada escenario en particular debe contar con al menos 5 casos

Con respecto a la independencia de las observaciones podemos argumentar que al generarse los resultados al mismo tiempo estos son independientes y cumplimos con la segunda condición ya que observamos que hay más de 5 eventos para cada caso:

Creamos nuestra tabla de **Composición real y esperada**

	Rock	Paper	Scissors	Total
Observado	43	21	35	99
Esperado	33	33	33	99

Calculamos el valor de Z por cada evento

$Z = \frac{p - p_0}{\sqrt{n}}$

Entonces:

$Z_{rock} = \frac{43-33}{\sqrt{33}} = 1.74$

$Z_{paper} = \frac{21-33}{\sqrt{33}} = -2.09$

$Z_{scissors} = \frac{35-33}{\sqrt{33}} = 0.35$

La suma cuadrada de los Z es:

$1.74^2 + (-2.09)^2 + 0.35^2 = 7.5182 = \chi^2$

Los grados de libertad son 2 dado que tenemos 3 categorías como posibles resultados.

Aplicamos estos valores en r para poder estimar el p-value.

In [4]:
```
library(MASS)
library(knitr)

fx_e = c(1/3,1/3,1/3)
fx_o = c(43,21,35)
chisq.test(fx_o, p = fx_e)
```

```
        Chi-squared test for given probabilities

data:  fx_o
X-squared = 7.5152, df = 2, p-value = 0.02334
```

Como el p-value < 0.05 entoces rechazamos la hipotesis nula, por lo que los resultados no son equipobrables

6.49 Shipping holiday gifts.

A December 2010 survey asked 500 randomly sampled Los Angeles residents which shipping carrier they prefer to use for shipping holiday gifts. The table below shows the distribution of responses by age group as well as the expected counts for each cell (shown in parentheses).

		Age						
		18-34		35-54		55+		Total
Shipping Method	USPS	72	(81)	97	(102)	76	(62)	245
	UPS	52	(53)	76	(68)	34	(41)	162
	FedEx	31	(21)	24	(27)	9	(16)	64
	Something else	7	(5)	6	(7)	3	(4)	16
	Not sure	3	(5)	6	(5)	4	(3)	13
	Total	165		209		126		500

(a) State the null and alternative hypotheses for testing for independence of age and preferred shipping method for holiday gifts among Los Angeles residents.

H_0 : No existe relación entre la edad de los residentes de Los Angeles y su método de envío favorito;

H_1 : Existe relación entre la edad de los residentes de Los Angeles y su método de envío favorito;

(b) Are the conditions for inference using a chi-square test satisfied?

No, porque no cumple con que cada escenario tenga al menos 5 observaciones.

6.53 The Egyptian Revolution.

A popular uprising that started on January 25, 2011 in Egypt led to the 2011 Egyptian Revolution. Polls show that about 69% of American adults followed the news about the political crisis and demonstrations in Egypt closely during the first couple weeks following the start of the uprising. Among a random sample of 30 high school students, it was found that only 17 of them followed the news about Egypt closely during this time.

(a) Write the hypotheses for testing if the proportion of high school students who followed the news about Egypt is different than the proportion of American adults who did.

H_0 : la porción de estudiantes que siguieron la noticia es igual al 69%

H_1 : la porción de estudiantes que siguieron la noticia es diferente al 69%

(b) Calculate the proportion of high schoolers in this sample who followed the news about Egypt closely during this time.

Calculamos la proporción de la muestra ($\hat{p}$):

$\hat{p} = \frac{17}{30} = 0.57$

(c) Based on large sample theory, we modeled $\hat{p}$ using the normal distribution. Why should we be cautious about this approach for these data?

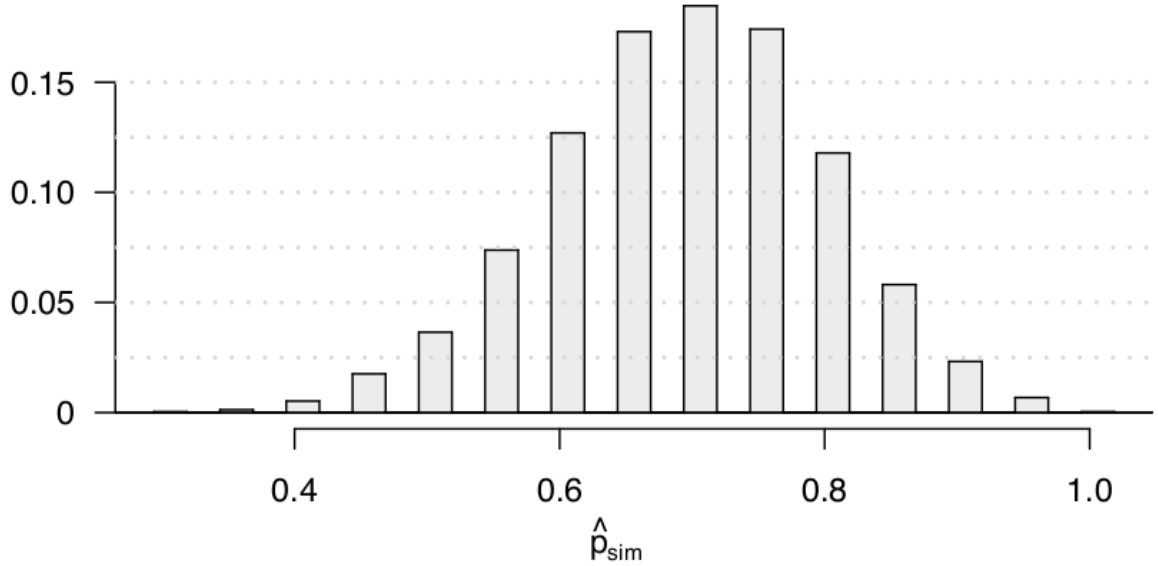
Por qué tenemos que verificar que la distribución no esté demasiado sesgada para el tamaño de la muestra, nos debemos de asegurar observar al menos 10 eventos mínimos de éxitos y fracaso.

(d) The normal approximation will not be as reliable as a simulation, especially for a sample of this size. Describe how to perform such a simulation and, once you had results, how to estimate the p-value.

Bajo la hipótesis nula, 69% de los estudiantes siguieron la Revolución Egipcia. Supongamos que esta tasa realmente no fuera diferente. Si este fuera el caso, podríamos simular 30 estudiantes para obtener una proporción de muestra considerando la hipótesis nula.

Podemos simular a cada estudiante como un evento con una probabilidad de éxito del 69%, es decir para los que siguieron las noticias, si hiciéramos este ejercicio 30 veces y calculamos $\hat{p}$ entonces esta porción de muestra es exactamente una muestra de la distribución nula. Realizaremos esta simulación al menos 10,000 veces y graficamos la porción de estudiantes que siguen las noticias.

(e) Below is a histogram showing the distribution of $\hat{p}_{sim}$ in 10,000 simulations under the null hypothesis. Estimate the p-value using the plot and determine the conclusion of the hypothesis test.



Para este cálculo usamos la distribución binomial:

In [38]:
```
sum(dbinom(x = 0:17, size = 30, prob = 0.69))
```

0.105252568880402

Como este valor es más grande que 0.05, fallamos en rechazar la hipótesis nula, es decir, no hay evidencia de que la proporción de estudiantes que siguieron las noticias fuera diferente a la proporción de adultos estadounidenses.