1	2	3	4	Total

Apellido y nombre :

Bioestadística - Final Febrero 2016 -

Ejercicio 1. Un grupo de científicos quieren estimar la cantidad de peces en un lago. Para hacerlo pescaron con una red 100 peces en la parte sur del lago, los pintaron y los devolvieron al lago. Dejaron pasar 3 días para que los peces pintados se mezclen con los otros y se distribuyan nuevamente por todo el lago. Luego, se comienzan a sacar peces (con reposición) hasta que encuentran uno pintado. Los peces que no están pintados se devuelven al lago. Entre pesca y pesca se deja pasar un tiempo

Llamamos X a la variable aleatoria determinada por la cantidad de intentos hasta obtener un pez pintado y N a la cantidad total de peces en el lago.

1. (a) Determinar la distribución de la variable aleatoria X. Especificar sus parámetros, indicando cuáles son conocidos y cuáles habría que estimar.

   (b) ¿Cómo estimaría N, a partir de $X_1, \ldots, X_n$, donde X_i son realizaciones de la variable aleatoria X?

   (c) Si suponemos que $N = 10000$, ¿ Cuántos peces espera tener que muestrear hasta encontrar un pez pintado.

   (d) Si se demora un minuto en muestrear cada pez y se sabe que N es del orden de los 10000 ¿le parece razonable estimar N usando la estrategia descripta? (Supongamos que para obtener una buena estimación de N queremos obtener al menos 50 X_i, o sea, $X_1, \ldots X_{50}$).

2. La estimación anterior se hizo suponiendo que los peces pintados se habían distribuído de manera uniforme en el lago. Para ver si esto es cierto, los científicos decidieron realizar la siguiente prueba: muestrearon 200 peces del sur y 200 del norte con reposición. En el sur 5 de ellos estaban pintados, mientras que en el norte eran 3.

   (a) Dar intervalos de confianza a nivel 0.1 para la proporción p de peces pintados en el sur y en el norte.

   (b) Realizar un test de hipótesis a nivel 0.01 para ver si es razonable pensar que los peces pintados se distribuyeron bien en el lago.

Solución:

1. (a) X tiene distribución Geométrica de parámetro p pues es una repetición de experimentos independientes (sacar peces del lago) hasta que ocurra un éxito (encontrar un pez pintado). El hecho de que se extraigan con reposición garantiza la independencia de los experimentos. El parámetro p es la probabilidad de éxito, que en este caso es:

$$p = \frac{100}{N}$$

porque hay 100 peces pintados en todo el lago y N la cantidad total. Por lo tanto, p y N son desconocidos y deben ser estimados. Una vez estimado uno de los dos, el otro se estima directamente.

(b) Si tenemos realizaciones $X_1, \ldots, X_n$ de X, entonces por Ley Fuerte de los Grandes Números tenemos que:

$$\overline{X_n} \underset{n\to\infty}{\longrightarrow} E(X) = \frac{1}{p} = \frac{N}{100}$$

Donde $\overline{X_n}$ es el promedio de $X_1, \ldots, X_n$. Por lo tanto, $\hat{N} = 100 \times \overline{X_n}$ es un estimador de N.

(c) La cantidad esperada de peces hasta encontrar un pez pintado es por definición $E(X)$, por lo que usando la formula para la variable geométrica:

$$E(X) = \frac{1}{p} = \frac{N}{100} = \frac{10.000}{100} = 100$$

(d) Si se demora un minuto en muestrear cada pez, en promedio se tardará 100 minutos (aprox 2 horas) sacando peces hasta encontrar un pez pintado. Esto lo debemos hacer para obtener cada X_i. Si quremos obtener por lo menos 50 peces pintados, tenemos que en promedio tardaremos 5000 minutos en completar el experimento. Esto son aproximadamente 3,5 días solamente dedicados a sacar peces, por lo que no parece una estrategia razonable.

2. (a) Llamemos $\hat{p}_S$ y $\hat{p}_N$ a las proporciones de peces pintados en el sur y en el norte respectivamente. Tenemos entonces que:

$$\hat{p}_S = \frac{5}{200} = 0,025 \quad \hat{p}_N = 0,015$$

y los intervalos de confianza I_S e I_N buscados tienen la forma:

$$I_S = \left[\hat{p}_S \underset{-}{+} \frac{\sqrt{\hat{p}_S(1-\hat{p}_S)}}{\sqrt{200}} z_{1-\frac{0,1}{2}}\right] \approx \left[0,025 \underset{-}{+} 0,018\right] = [0,007; 0,043]$$

$$I_N \approx \left[0,015 \underset{-}{+} 0,014\right] = [0,001; 0,029]$$

(b) Realizamos un test de comparación de dos muestras $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$, donde cada X vale 1 o 0 dependiendo de si el pez extraído de la parte sur del lago está pintado o no (respectivamente con Y y la parte norte). El estadístico vale:

$$E = \frac{\overline{X_n} - \overline{Y_n}}{\sqrt{\frac{s_n^2(X) + s_n^2(Y)}{n}}} = \frac{\hat{p}_S - \hat{p}_N}{\sqrt{\frac{\hat{p}_S(1-\hat{p}_N) + \hat{p}_N(1-\hat{p}_N)}{200}}} \approx 0,714$$

y como la región crítica es $\mathbb{RC} = \{|E| \geq z_{0,995} \approx 2,58\}$ tenemos que no se rechaza la hipótesis nula por lo que no podemos afirmar que haya una diferencia significativa en la distribución de los peces pintados en el lago.

Ejercicio 2.

1. Se tienen 3 bolilleros y se saca 1 bolilla de cada uno al azar. Los bolilleros tienen las siguientes cantidades de bolillas

   - El Bolillero 1 tiene 20 bolillas rojas y 40 azules
   - El Bolillero 2 tiene 10 rojas 10 azules y 10 blancas
   - El Bolillero 3 tiene 20 rojas y 10 azules

   Sean X, Y, Z la variables aleatoria cantidad de bolillas azules, rojas y blancas extraídas respectivamente

   (a) Mostrar que $X + Y + Z = 3$.

   (b) Definir $\mathcal{R}_X$ y $\mathcal{R}_Z$ y calcular las funciones de distribución de X y Z.

   (c) Calcular las esperanza y la varianza de X.

   (d) Calcular la esperanza y varianza de Z.

   (e) Calcular la esperanza de Y. (Recordar que $X + Y + Z = 3$)

   (f) Sabiendo que la varianza de Y es igual a la varianza de X, se puede decir si las variables aleatorias X, Y y Z son independientes?

   (g) Se realizó el experimento y se obtuvieron 2 bolillas rojas y una azul. ¿Cual es la probabilidad de que la azul sea del primer bolillero?

   (h) Y si hubieran salido dos azules y 1 blanca, ¿cuál es la probabilidad que las azules hayan salido del primer y tercer bolillero?

Solución:

1. X, Y y Z cuentan la cantidad de bolillas de cada uno de los colores que salieron, por lo tanto su suma da el total de bolillas sacadas, que en este caso es 3.

2. $\mathcal{R}_X = \{0, 1, 2, 3\}$ ya que en todos los bolilleros hay bolillas azules. $\mathcal{R}_Z = \{0, 1\}$, ya que solamente el segundo bolillero tiene bolillas blancas, por lo tanto, a lo sumo podemos obtener 1 bolilla blanca.

   Para calcular la función de dsitribución de X debemos calcular $P(X = i)\ \forall i \in \{0, 1, 2, 3\}$.

   $P(X = 0) = P$(no sale ninguna bolilla azul)$= \frac{40}{60}\frac{20}{30}\frac{10}{30} = \frac{4}{27}$

   $P(X = 1) = P$(sale una bolilla azul del primer bolillero ó del segundo ó del tercero)

   $= \frac{20}{60}\frac{20}{30}\frac{10}{30} + \frac{40}{60}\frac{10}{30}\frac{10}{30} + \frac{40}{60}\frac{20}{30}\frac{20}{30} = \frac{4}{9}$

   $P(X = 2) = P$(no sale una bolilla azul del primer bolillero ó del segundo ó del tercero y sale azul en el resto de los bolilleros)=

   $= \frac{40}{60}\frac{10}{30}\frac{20}{30} + \frac{20}{60}\frac{20}{30}\frac{20}{30} + \frac{20}{60}\frac{10}{30}\frac{10}{30} = \frac{1}{3}$

   $P(X = 3) = P$(todas las bolillas son azules)$= \frac{20}{60}\frac{10}{30}\frac{20}{30} = \frac{2}{27}$

Para calcular la función de dsitribución de Z sólo miramos lo que sale en el segundo bolillero, ya que como los otros dos no tienen bolillas blancas, el color de las bolillas que salen de éstos no nos importan.

$P(Z=0) = \frac{20}{30} = \frac{2}{3}$

$P(Z=0) = \frac{10}{30} = \frac{1}{3}$

3. $E(X) = 0 \times P(X=0) + 1 \times P(X=1) + 2 \times P(X=2) + 3 \times P(X=3) = \frac{4}{9} + 2\frac{1}{3} + 3\frac{2}{27} = \frac{4}{3}$

   Para calcular la varianza usamos $Var(X) = E(X^2) - E^2(X)$

   $E(X^2) = 0^2 \times P(X=0) + 1^2 \times P(X=1) + 2^2 \times P(X=2) + 3^2 \times P(X=3) = \frac{4}{9} + 2^2\frac{1}{3} + 3^2\frac{2}{27} = \frac{4}{3}$

   $\Rightarrow Var(X) = \frac{22}{9} - \frac{16}{9} = \frac{2}{3}$

4. $E(Z) = 0 \times P(Z=0) + 1 \times P(Z=1) = \frac{1}{3}$

   $E(Z^2) = 0 \times P(Z=0) + 1 \times P(Z=1) = \frac{1}{3}$

   $\Rightarrow Var(Z) = \frac{1}{3} - \frac{1}{9} = \frac{2}{9}$

5. Usando que $E(X+Y+Z) = E(3)$ y la linealidad de la esperanza, se tiene que:

   $E(Y) = 3 - E(X) - E(Z) = \frac{4}{3}$

6. Usando que $Var(X+Z) = Var(3-Y) = Var(Y) = Var(X)$. La última igualdad se da por letra.

   Si X, Y y Z fueran independientes, se tendría que en particular X y Z tambio serían y por lo tanto debería cumplirse que $Var(X+Z) = Var(X) + Var(Z)$, pero en la parte anterior se vio que $Var(X+Z) = Var(X)$, por lo tanto, si fueran independientes la Varianza de Z tendría que ser 0, pero por la parte (d) sabemos que vale $\frac{1}{3}$.

   De estas igualdades se puede deducir que la covarianza de X y Z es $\frac{1}{3}$.

7. $P(\text{azul en el primer bolillero}|\text{dos rojas y una azul}) = \frac{P(\text{ARR})}{P(\text{ARR})+P(\text{RAR})+P(\text{RRA})}$

   $= \frac{\frac{4}{27}}{\frac{4}{27}+\frac{2}{27}+\frac{1}{27}} = \frac{4}{7}$

8. La probabilidad es uno, dado que la única manera de que salga una blanca es que lo haya hecho del segundo bolillero. Por lo tanto la primera y tercera bolilla tienen que haber sido azules.

Ejercicio 3.

Este año se argumentó que la suba del precio de la fruta se debía a la mayor presencia de mosca de la fruta. Para determinar si es así se intentó modelar la distribución de picaduras por fruta en muestras de años anteriores. En una muestra de 100 frutas se obtuvieron los siguientes resultados:

Picaduras por fruta	0	1	2	3	4	5	6	> 7
cantidad de frutas	0	11	13	19	18	17	12	10

1. Si se quiere modelar con una distribución de Poisson de parámetro 4, calcular las frecuencias esperadas para las cantidades de picaduras por fruta expresadas en la tabla.

2. Realizar un test para ver si la muestra se ajusta a dicha distribución.

3. Para poder vender la fruta, ésta no puede tener más de 5 picaduras. Si el productor gana \$0,5 por fruta en condiciones de venta y pierde \$ 0,2 por cada fruta de descarte. ¿Cuánto gana en promedio el productor por cajón de 100 frutas?

4. Este año hubo el doble de moscas que el año pasado, por lo tanto el número promedio de picaduras se duplicaron. Si un productor produjo 1000 frutas, con cuánto dinero tendría que indemnizarlo el estado para que no pierda dinero?

Solución:

1. Usando la distribución de Poisson de parámetro 4, tenemos que:

Picaduras por fruta	0	1	2	3	4	5	6	> 7
cantidad de frutas	0	11	13	19	18	17	12	10
cantidad esperada	1,8	7,3	14,7	19,5	19,5	15,6	10,4	11,1

2. Teniendo las cantidades esperadas, lo más sencillo es realizar un Test de Ajuste χ^2, donde la H_0 es: el test sigue la dsitribución propuesta y H_1 es no H_0.

   Si llamamos E al estadístico y p al p-valor tenemos que:

   $$E \approx 4,5 \Rightarrow p > 0,9$$

   Podemos concluir que para valores de α mayores a 0,1 (y muchos otros valores de α también), se acepta la hipótesis nula.

3. Si la variable aleatoria G cuenta la ganancia por cada fruta, tenemos que $G = \$0,5$ cuando la fruta tiene 5 picaduras o menos y $G = -\$0,2$ cuando tiene más de 5 picaduras. Podemos estimar la probabilidad q de tener 5 picaduras o menos como:

   $$\hat{q} = P(X \leq 5) = 0,018 + 0,073 + 0,147 + 0,195 + 0,195 + 0,156 = 0.784$$

   Con esta estimación de q, podemos asumir que la ganancia promedio por cajon de 100 frutas es:

   $$100 \times E(G) = 100 \times (0,784 \times \$0,5 - 0,216 \times \$0,2) = 34,88$$

4. Asumiendo que la variable sigue una distribución de Poisson de parámetro 4, si se duplicaron las picaduras promedio por fruta entonces este año la variable sigue una distribución $Poiss(8)$ (pues el parámetro λ de la Poisson es su esperanza). Por lo tanto, siguiendo la misma línea de razonamiento que el ejercicio anterior, ahora podemos calcular la probabilidad q teóricamente como:

   $$q = P(\text{cant picaduras} \leq 5) = P(Poiss(8) \leq 5) \approx 0,191$$

   Calculamos la ganancia esperada:

   $$1000 \times E(G) = 1000 \times (0,191 \times \$0,5 - 0,809 \times \$0,2) = -\$66,3$$

   por lo que se concluye que el estado debe indemnizarlo con $\$66,3$ cada 1000 frutas producidas.

Ejercicio 4.

En la secuenciación de un organismo se definen las siguientes categorías para cada locus secuenciado en base a la cantidad de lecturas del genoma (R) dividido por la profundidad de lectura del genoma entero.

- Baja si $\mathrm{R} \leq 0.4$,
- Media si $0.4 < \mathrm{R} \leq 0.7$,
- Alta si $\mathrm{R} > 0.7$.

Se asume que la cantidad de lecturas R en un locus sigue una distribución Gaussiana con parámetros $\mu = 0.5$ y $\sigma^2 = 0.01$.

1. Hallar la probabilidad que las lecturas en un locus determinado pertenezca a cada una de las categorías antes mencionadas.
2. Se sabe que la probabilidad de perder datos (debido a errores en la lectura) para un locus de la categoría Media es de 0.15, mientras que dicha probabilidad se eleva a 0.75 dentro de la categoría Alta. Finalmente, la probabilidad de pérdida en un locus con calidad Baja es de 0.1. Se pide:
   (a) Calcular la probabilidad que en un locus elegido al azar se pierdan datos.
   (b) Dado que hubo pérdida de datos en un locus, ¿Cuál es la probabilidad de que pertenezca a la categoría Media?

Solución:

1. Llamamos Φ a la distribución de una normal estandar. Entonces:

$$P\left(\mathsf{Baja}\right) = P\left(\frac{\mathsf{R} - 0,5}{0,1} \leq \frac{0,4 - 0,5}{0,1}\right) = \Phi(-1) \approx 0,159$$

$$P\left(\mathsf{Alta}\right) = P\left(\frac{\mathsf{R} - 0,5}{0,1} > \frac{0,7 - 0,5}{0,1}\right) = 1 - \Phi(2) \approx 0,023$$

$$P\left(\mathsf{Media}\right) \approx 1 - 0,159 - 0,023 = 0,818$$

2. (a) Usando la formula de la probabilidad total:

$$\begin{aligned} P(\text{perdida}) &= P(\text{perdida}|\mathsf{Baja})P\left(\mathsf{Baja}\right) \\ &+ P(\text{perdida}|\mathsf{Media})P\left(\mathsf{Media}\right) \\ &+ P(\text{perdida}|\mathsf{Alta})P\left(\mathsf{Alta}\right) = \\ &= 0,1 \times 0,159 + 0,15 \times 0,818 + 0,75 \times 0,023 \approx 0,156 \end{aligned}$$

   (b) Usando la fórmula vista en el teórico para invertir el condicional:

$$P(\mathsf{Media}|\text{perdida}) = \frac{P(\text{perdida}|\mathsf{Media})P\left(\mathsf{Media}\right)}{P(\text{perdida})} = \frac{0,15 \times 0,818}{0,156} \approx 0,787$$