Examen Parcial 1 - Fundamentos de Estadística

Pedro Vladimir Hernández Serrano

---

1. Problema 1

Dada la naturaleza del problema podemos definir 2 variables aleatorias:

X: El evento de que un neonato tenga PKU
Y: El evento de llevar a cabo una prueba médica

Por lo que podemos verlas como variables dicotómicas de la siguiente manera:

$$X = \begin{cases} 1 & tiene\ PKU \\ 0 & no\ tiene\ PKU \end{cases} \quad \Big| \quad Y = \begin{cases} 1 & prueba\ positiva, \\ 0 & prueba\ negativa \end{cases}$$

Como información inicial tenemos que; 1 de cada mil niños recién nacidos contrae PKU, lo anterior podemos verlo como:

$P(X = 1) = \frac{1}{1000} = 0{,}0001$

Dado que se trata de una variable aleatoria que toma solo dos valores excluyentes entonces podemos calcular:

$P(X = 0) = 1 - {,}0001 = 0{,}9999$

Dicha expresión se traduce como la probabilidad de que un neonato no presente la enfermedad

Por otra parte, se define la Sensibilidad como; la probabilidad de que una prueba médica resulte positiva dado que proviene de una población de pacientes con PKU, i.e.

$P(Y = 1|X = 1) = 0{,}999$

Con este conocimiento podemos expresar el complemento de dicha probabilidad como sigue:

$P(Y = 0|X = 1) = 1 - 0{,}999 = 0{,}0001$

Es decir; uno de cada mil neonatos con PKU resulta con prueba negativa

De manera similar se define la Especificidad como; la probabilidad de que la prueba médica resulte negativa dado que proviene de una población de pacientes sin PKU

$P(Y = 0|X = 0) = 0{,}99$ por lo que: $P(Y = 1|X = 1) = 1 - 0{,}99 = 0{,}01$

El problema nos habla de calcular la probabilidad de que un recién nacido tenga KPU dado que la prueba médica resultó positiva, cuya problemática tiene mucho sentido para la ciencia de la salud, ya que podríamos verla como una tasa de efectividad de la prueba, veamos:

Se quiere calcular $P(X = 1|Y = 1)$

Por el Teorema de Bayes sabemos que; para dos variables aleatorias:

$$P(X = x|Y = y) = \frac{(X = x, Y = y)}{P(Y = y)} = \frac{P(Y = y|X = x)P(X = x)}{\sum_x P(Y = y|X = x)P(X = x)} \quad (1)$$

Para este caso en particular nos preguntamos para $X = 1$ y $Y = 1$

$$\begin{aligned} P(X = 1|Y = 1) &= \frac{P(Y = 1|X =)1P(X = 1)}{\sum P(Y = 1|X = 1)P(X = 1)} && (2) \\ &= \frac{(0{,}9999)(0{,}0001)}{(0{,}9999)(0{,}0001) + (0{,}01)(0{,}9999)} && (3) \\ &= 0{,}00990099 && (4) \end{aligned}$$

Es decir; aproximadamente 99 niños de cada 10 mil serán correctamente clasificados después de la prueba, esto podría no ser tan alentador ya que se espera que la información después de la prueba sea mayor, pero no ocurre así.

Se puede presumir que se debe en gran parte a la especificidad la que convierte las probabilidades en cantidades marginales, ya que se tiene que 99 de cada 100 aparecen negativos dado que no se tiene la enfermedad, cuando, dado las escalas se preferiría con un error aún menor digamos 99,999 de cada 100,000, se puede concluir que en el momento en que la prueba de especificidad alcance niveles de error prácticamente nulos es cuando se tendrán resultados informativos.

2. Problema 2

Tenemos la siguiente información

$P(X = 1, Y = 1) = 1/8$
$P(X = 1, Y = 2) = 1/4$
$P(X = 2, Y = 1) = 3/8$
$P(X = 2, Y = 2) = 1/4$

Se definen entonces 2 variables aleatorias X,Y las cuales podemos representar la distribución de probabilidad conjunta con la siguiente tabla de contingencia

X,Y	**Y=1**	**Y=2**	**P(X=x)**
X=1	1/8	1/4	3/8
X=2	3/8	1/4	5/8
Y=y	1/2	1/2	1

Cuadro 1: Tabla de probabilidades

Sabemos que dos variables aleatorias son estocásticamente independientes por que no importa en el orden en el que aparecen o se presentan los eventos, la probabilidad conjunta siempre será la misma, dado que el orden de los factores no alteran el producto

$$P(X_1, \ldots, X_n) = P(X_1) \ldots P(X_n) \tag{5}$$

Veamos un caso en particular

$P(X = 1, Y = 1) = 1/8$
$P(X = 1)P(Y = 1) = (1/2)(3/8) = (3/16)$
$P(X = 1, Y = 1) \neq P(X = 1)P(Y = 1)$

Pasa igual con todos los valores de la conjunta, por lo que podemos concluir que No son estocásticamente independientes

Queremos ahora calcular esperanzas condicionales con valores fijos, recordamos la probabilidad condicional de Bayes en su forma básica

$$P(X = x|Y = y) = \frac{(X = x, Y = y)}{P(Y = y)} \tag{6}$$

Encontramos entonces las probabilidades condicionales para los casos dicotómicos

$P(X=1|Y=1) = \frac{P(X=1,Y=1)}{P(Y=1)} = \frac{1/8}{1/2} = 1/4$
$P(X=2|Y=1) = \frac{P(X=2,Y=1)}{P(Y=1)} = \frac{3/8}{1/2} = 3/4$
$P(X=1|Y=2) = \frac{P(X=1,Y=2)}{P(Y=2)} = \frac{1/4}{1/2} = 1/2$
$P(X=2|Y=2) = \frac{P(X=2,Y=2)}{P(Y=2)} = \frac{1/4}{1/2} = 1/2$

Recordamos como se expresa la esperanza condicional

$$E(X|Y=y) = \sum_x xP(X|Y=y) \tag{7}$$

De esta manera podemos calcular la esperanza. En el primer caso evaluamos $Y=1$ por lo que

$E(X|Y=1) = (1)(1/4) + (2)(3/4) = 3/2 = 7/4 = 1{,}75$

Análogamente podemos ver el caso con $Y=2$ entonces:

$E(X|Y=2) = (1)(1/2) + (2)(1/2) = 3/2 = 3/2 = 1{,}5$

3. Problema 3

Análisis bajo un enfoque frecuentista:

Este enfoque nos habla de que las reglas de decisión se pueden modificar, dado que los datos son una repetición de una muestra aleatoria, por lo que hay una frecuencia en los comportamientos específicos de cada muestra, pero donde los parámetros de la población permanecen constantes durante el proceso de repetición

En este enfoque los parámetros son los que se actualizan

Con respecto a la probabilidad se desea: $P(data|H_0)$ se refiere a la distribución de la muestra de los datos dada la hipótesis de los parámetros

En la prueba recolectamos una colección de intervalos con un (supongamos) 90 % de probabilidad que contenga al verdadero parámetro

Análisis bajo un enfoque bayesiano:

Este otro enfoque del estudio estadístico nos habla de que los parámetros son desconocidos son tratados probabilísticamente y que la información colectada puede siempre actualizarse después de que los datos son observados de una muestra realizada

En este enfoque los datos son los que se actualizan

Con respecto a la probabilidad se define de manera inicial $P(\theta)$ la distribución a priori o conocimiento previo del parámetro antes de que los datos se evaluen, esta puede verse como la subjetividad del especialista

En la prueba se define un intervalo el cual tiene 90 % de probabilidad de contener el verdadero parámetro.

Frecuentista	**Bayesiano**
Los estudios se repiten	Los estudios se actualizan
Evaluar el Fenónemo	Modelar el Fenónemo
La repetición es importante	La inferencia es apropiada

Cuadro 2: Diferencias Enfoques

4. Problema 4

Después de analizar el problema podemos darnos cuenta que el "juego"se lleva a cabo una vez, se toma el supuesto de que hay un solo individuo en la toma de decisión, enunciamos entonces los siguientes elementos:

1. Acciones: $A = (a_1, a_2, a_3)$ el evento que gana el boleto tipo A_i
2. Incertidumbre: $I = (i_1, i_2, ...i_I)$ desconocimiento de que ocurra A_i
3. Consecuencia: $C = A\ x\ I = \{c_1 = (a_1, i_1), c_2 = (a_2, i_2), c_3 = (a_3, i_3)\}$ Si ocurre c_i el jugador recibe de vuelta $4n_i x_i$ donde x_i es un costo fijo por tipo de boleto: $x_1 = 1/6$, $x_2 = 1/3$ y $x_3 = 1/2$ y n_i es un número desconocido de boletos que se debe comprar de cada tipo

Cuantificaciones

1. Probabilidad de ocurrencia de cada consecuencia
   $P(c_1) = 1/2$, $P(c_2) = 1/3$, $P(c_3) = 1/6$
2. Utilidad: $U(c_i, n_i)$ Dentro de los supuestos iniciales se menciona que el jugador utiliza una función de utilidad logarítmica $Log(U(c_i, n_i))$
   Asignamos los valores correspondientes al espacio de consecuencias

   $Log(U(c_1, n_1)) = Log(4(1/6)n_1) = Log(\frac{4}{6}n_1)$
   $Log(U(c_2, n_2)) = Log(4(1/3)n_2) = Log(\frac{4}{3}n_2)$
   $Log(U(c_3, n_3)) = Log(4(1/2)n_3) = Log(\frac{4}{2}n_3)$

Formulamos entonces el valor esperado de la utilidad del jugador se puede ver como:

$$E_u[Log(U(c_i, n_i))] = \sum_{i=1}^{3} P(c_i) Log(U(c_i, n_i)) \tag{8}$$

Se convierte entonces en nuestra función objetivo la cual queremos maximizar, simplificamos el modelo.

$$E_u[Log(U(c_i, n_i))] = \frac{1}{2}Log(\frac{4}{6}n_1) + \frac{1}{3}Log(\frac{4}{3}n_2) + \frac{1}{6}Log(\frac{4}{2}n_3) \tag{9}$$

$$= \frac{1}{2}Log(\frac{2}{3}) + \frac{1}{2}Log(n_1) + \frac{1}{3}Log(\frac{4}{3}) + \frac{1}{3}Log(n_2) + \frac{1}{6}Log(2)\frac{1}{6}Log(n_3) \tag{10}$$

$$= \frac{1}{2}Log(n_1) + \frac{1}{3}Log(n_2) + \frac{1}{6}Log(n_3) + \overbrace{Log\Big((\frac{2}{3})^{1/2}(\frac{4}{3})^{1/3}2^{1/6}\Big)}^{constante} \tag{11}$$

Por lo que buscamos maximizar es

$$\hat{A} = max\left\{\frac{1}{2}Log(n_1) + \frac{1}{3}Log(n_2) + \frac{1}{6}Log(n_3) + cte\right\} \tag{12}$$

En este punto nos auxiliamos de la herramienta solver® para la optimización de nuestra función objetivo, dado que tenemos n_i el número de boletos comprados de cada tipo como factores variables

Ahora falta ajustarnos a la restricción de la fortuna inicial con la que el jugador cuenta $F = 1000$, hay que considerar también que dentro de los supuestos conocemos que se gastó el capital por completo en la compra de los 3 tipos boletos por lo que

$$F = \sum_{1}^{3} n_i x_i \tag{13}$$

Tenemos entonces para este caso la función de restricción:

$$F = \frac{n_1}{6} + \frac{n_2}{3} + \frac{n_3}{2} = 1000 \tag{14}$$

Para que la optimización se lleve a cabo con mayor precisión se acota el modelo de modo que $n_1,\ n_2,\ n_3$ sean enteros positivos

Después de correr el modelo con el método "$GNR\ Non-linear$" declarando el número de boletos como variables a modificar, dando lugar a la ganancia y a la probabilidad que ocurra la ganancia como valores fijos, maximizando la ganancia esperada como vimos anteriormente, podemos ver los resultados en la Figura 1

	A1	A2	A3	
Restricción	0.167	0.333	0.500	1000
Inversión	500	333	167	999.833
Probabilidad	0.500	0.333	0.167	Función objetivo
Ganancia	0.667	1.333	2.000	
Boletos	3,000	1,000	333	
Log(F)	1.651	1.042	0.471	3.162740

Figura 1: Optimización Solver

El modelo se detiene cuando la suma de la inversión es muy cercana a 999,833 $\approx$ 1000 de aquí que; el escenario optimo donde gana

$E_u[Log(U(c_i, n_i))] = 3{,}162740$ Podemos resumir los resultados para los 3 tipos de boletos. Sea Y_i la inversión en i

Boletos	**Inversión**	**Utilidad**
$n_1 = 3000$	$n_1 x_1 = 500$	$p_1 n_1 x_1 - Y_1 = 500$
$n_2 = 1000$	$n_2 x_3 = 333$	$p_2 n_2 x_2 - Y_2 = 111$
$n_3 = 333$	$n_3 x_3 = 167$	$p_3 n_3 x_3 - Y_3 = -56$
4,333	**1,000**	**555**

Cuadro 3: Inversión óptima

Podemos concluir que si recomendamos al jugador elegir esta estrategia tendrá una fortuna final de $F = 1000 + 555 = 1,555$ dado que el modelo es el escenario óptimo

De manera general podemos siempre expresar dicha función de perdida para cualquier número de escenarios A_i siempre y cuando la suma de las probabilidades de ocurrencia de cada uno sea igual a la unidad $\sum_{i=1}^{n} p_i = 1$

$$max\Big\{E_u[Log(U(c_i, n_i))]\Big\} = max\Big\{\sum_{i=1}^{n} Log(n_i x_i)^{p_i'}\Big\} \tag{15}$$

Por lo que su fortuna final en el caso general podría verse como

$$\hat{F} = \sum_{i=1}^{n} \Big(p_i' \hat{n}_i x_i - Y_i\Big) \tag{16}$$

Suponiendo una vez más que la función objetivo se comporta logarítmica y que se gasta toda su fortuna inicial en la compra de boletos.

5. Caso de estudio

5.1. Conocimiento previo

Entendemos que para este caso en particular ya se ha hecho una previa formulación del problema, hubo una recolección de los datos, y se llevó acabo el uso de las distíntas técnicas de estadística descriptiva

Enunciamos dos supuestos para este ejemplo:

- Se trata de una materia de los últimos semestres
- En general las materias los últimos semestres del MBA suelen tener calificaciones muy cercanas al 100 %

5.2. Análisis Exploratorio

Podemos observar que para este grupo se promedia una calificación final de 89,98, lo cual, en el caso hipotético es un tanto bajo sobre el promedio de los demás años en el MBA, así como también la variabilidad de las calificaciones de los alumnos, ya que tenemos una disperción medida con la desviación estándar de 7.16, por lo que podemos construir un intervalo de disperción en el rango de $(82{,}82 - 97{,}14)$

Con la tabla de frecuencias presentada podemos también visualizar la distribución de frecuencias relativas o distribución de probabilidad como se muestra.

Rango	**Frecuencia**	**Distribución**
70-	1	0.011
70-75	3	0.035
75-80	6	0.07
80-85	9	0.105
85-90	17	0.198
90-95	23	0.267
95-100	26	0.302
100-	1	0.012

Cuadro 4: Tabla de frecuencias

A simple vista vemos que poco más de la mitad del grupo se encuentra en los grupos con calificación superior a 90, que es donde se centra la distibución precisamente, lo anterior podríamos leerlo como que existe un 58 % una clase o un grupo de alumnos saque 90 o más.

Con dicha observación y dado que la Moda, es decir la máxima frecuencia relativa es 95,454, junto con la mediana 91,41 que podra verse en este sentido como el punto donde el grupo ha alcanzado la mitad de la frecuencia acumulada de la distribución con dicho valor

De aquí que podrímos intuir que la media estuviera también por encima de 90, pero no es así, si no que apenas por debajo, veamos entonces la precisión de la media. Definimos intervalos de confianza a partir de dichas medidas de tendencia central (consideramos una confianza del 95 % para la media.

$$(\bar{X} - 1{,}96\frac{\sigma}{\sqrt[2]{n}}\ ,\ \bar{X} + 1{,}96\frac{\sigma}{\sqrt[2]{n}}) \tag{17}$$

$$(89{,}98 - 1{,}96\frac{7{,}16}{\sqrt[2]{86}}\ , 89{,}98 + 1{,}96\frac{7{,}16}{\sqrt[2]{86}}) =$$

$$(88{,}46 - 91{,}49)$$

Notamos que la media es precisa, a pesar de que la dispersión en lo general es grande, es decir, hay muchos casos que se encuentran alejados de la media, lo podemos observar en en la figura 2

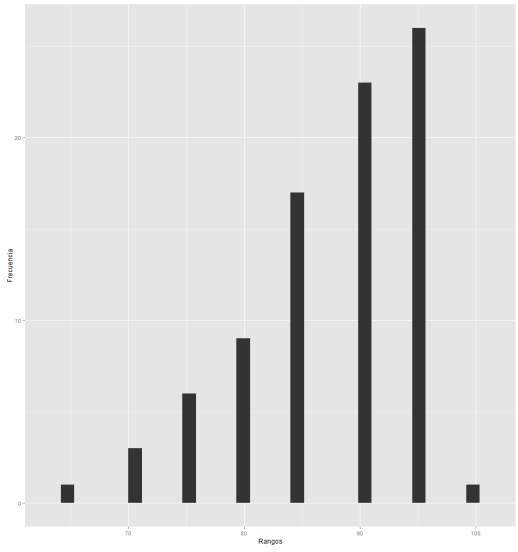

Figura 2: Histograma de frecuencias

Esto es consistente ya que si análizamos que tan extensos son los datos observamos una calificación máxima de 100 y una mínima de 65,66, un

alumno en cada caso respectivamente, resultando un rango de $100 - 65{,}66 = 34{,}34$, un umbral muy amplio para el grupo.

Una kurtosis de $0{,}85 > 0$, nos da una idea de que hay frecuencias relativas muy altas, es decir que en los rangos de entre el 90 y 100 hay se concentran muchas personas lo que hace una curva "picuda", observamos también un sesgo de $-1{,}07 < 0$ ocasionando de esta manera que la distribución del grupo se "mueva"mucho hacia la derecha de la media, esto es consistente ya que $mediana > media$

5.3. Conclusiones

Con el prestigio que tiene la institucíon, pareciaera que la radiografía presentada a dicha clase no serían las mejores noticias para el MBA, ya que se esperaría que fuera un grupo más balanceado, que no es lo mismo que todos tengan excelentes calificaciones, si no que no haya tanta dispersión en sus notas finales, ya que todos recibieron la misma formación durante el curso y de cierto modo han nivelado sus conocimientos a lo largo del programa.

Dicho lo anterior podría deberse a diferentes factores, pero en general me aventuro a decir que debe referirse a problemas de falta de objetividad a la hora de calificar por parte del profesor hacia todos y cada uno de los alumnos esto dado que todas las personas tienen las mismas bases, cuentan con el mismo (o parecido) tiempo que le dedican a sus estudios y/o responsabilidades.

Otro escenario puede ser también la falta de interés en los casos donde los alumnos presentan calificaciones muy por debajo de la media, y esto puede deberse que el curso en si mismo no los invita a relacionarlo con el resto del programa o el profesor no motiva lo suficiente los temas a abordar, de esta manera la institución debe verlo como un área de oportunidad ya que probablemente no están llevando a cabo una comunicación asertiva con los profesores y/o alumnos, como mencionabamos anteriormente todo ello con el fin de tener una clase balanceada y todos digieran los conocimientos de una manera similar