Semana 2

Muestreo: Conceptos Previos *Valentín Vergara Hidd*

En este documento presentaré los conceptos básicos para entender cómo se crea una muestra que sea representativa de una población. Es importante aclarar que aún no hablaremos de muestreo, que es un tema a al que le dedicaremos relativamente poca atención, en comparación a las bases conceptuales en probabilidad y distribuciones probabilísticas. Esto porque la probabilidad es algo fundamental para entender las pruebas estadísticas que posteriormente implementaremos en SPSS.

Antes de crear *modelos*, hay que considerar que los científicos generalmente no contamos con información de toda la población a la que se quiere estudiar. Por tanto, el modelo se vuelve necesario, al ser una **simplificación** de una población más numerosa y muy probablemente, más compleja.

Para entender cómo pasamos de una **idea** a un modelo útil, hay que revisar un par de conceptos previos.

Probabilidad

Consideren un experimento en el que podemos conocer todos los resultados que se podrían obtener, pero sin saber exactamente cuáles de ellos se obtendrán. Por ejemplo, lanzar una moneda[1], elegir una carta de un mazo de naipes, escoger un número de una tómbola, etc.

Asignaremos **probabilidades** a los distintos resultados $\mathcal{A}$, o a conjuntos de resultados. Estos resultados/conjuntos serán los **eventos**. Por su parte, la lista con todos los *posibles resultados* será el **espacio muestral** Ω.

Ejemplo 1: Un ejemplo simple

Al lanzar una moneda:

$$\Omega = \{\text{cara}, \text{sello}\}$$

Más adelante se va a volver extremadamente importante reconocer qué valores constituyen probabilidades y cuáles no lo hacen. En principio, el criterio es simplemente revisar las propiedades a continuación:

[1] Una ley universal, en cualquier texto de probabilidad o estadística, es que siempre hay un ejemplo con monedas, dados o naipes.

Definición 1: Propiedades de las probabilidades

1. $\Pr(\Omega) = 1$
2. $\forall \mathcal{A} \in \Omega,\ 0 \leq \Pr(\mathcal{A}) \leq 1$
3. Si los eventos $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_k$ son independientes, $\Pr\left(\bigcup_{i=1}^{k} \mathcal{A}_i\right) = \sum_{i=1}^{k} \Pr(\mathcal{A}_i)$

Aplicando la definición 1 al muestreo, si se tiene una población de N elementos y se utiliza algún criterio para escoger una muestra de tamaño n, cada uno de los elementos tiene una probabilidad de $\frac{1}{N}$ de ser escogido.

Muestreo Aleatorio con reemplazo

La idea central es que una vez que se escoge un elemento de Ω, se reemplaza y puede volver a ser escogido. Dicho de otra forma, el tamaño de Ω no varía.

Ejemplo 2: Números pequeños

Una población de tamaño $N = 5$ de la que se obtendrá una muestra de tamaño $n = 2$, tiene 25 elementos (a, b) en Ω: donde a es el primer elemento que se escoge y b el segundo. Asumiendo que $N = \{1, 2, 3, 4, 5\}$:

$$\Omega = \left\{ \begin{array}{ccccc} (1,1) & (2,1) & (3,1) & (4,1) & (5,1) \\ (1,2) & (2,2) & (3,2) & (4,2) & (5,2) \\ (1,3) & (2,3) & (3,3) & (4,3) & (5,3) \\ (1,4) & (2,4) & (3,4) & (4,4) & (5,4) \\ (1,5) & (2,5) & (3,5) & (4,5) & (5,5) \end{array} \right\}$$

El ejemplo 2 implica que cada elemento en Ω tiene una probabilidad de $\frac{1}{25}$ de ser escogido. Sin embargo, noten que muchas veces $(1, 2)$ y $(2, 1)$ representan exactamente lo mismo. Este conjunto de resultados lo representaremos como $\mathcal{S} = \{1, 2\}$. Entonces, por la tercera propiedad de la definición 1:

$$\begin{aligned} \Pr(\{1,2\}) &= \Pr\left[(1,2) \cup (2,1)\right] \\ &= \Pr(1,2) + \Pr(2,1) \\ &= \frac{2}{25} \end{aligned}$$

Supongan ahora que nos interesa un conjunto de resultados $\mathcal{S}$ diferente.
$\mathcal{S} = \{\text{El número 5 está en alguna parte del resultado}\}$. Para encontrar $\Pr(\mathcal{S})$, simplemente podríamos mirar Ω en el ejemplo 2 y contar. El resultado es $\frac{9}{25}$.
Otra alternativa es utilizar una propiedad de la suma de probabilidades

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr\left(A \cap B\right) \tag{1}$$

Si para este ejemplo definimos:

A = 5 está en el primer número

B = 5 está en el segundo número

Utilizando la ecuación 1:

$$\begin{aligned}\Pr(\mathcal{S}) &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\ &= \frac{1}{5} + \frac{1}{5} - \frac{1}{25} \\ &= \frac{9}{25}\end{aligned}$$

Noten que en este ejemplo:

$$\Pr(A \cap B) = \Pr(A) \times \Pr(B) \tag{2}$$

Esto ocurre debido a que A y B son **independientes**

Muestreo aleatorio sin reemplazo

Como vimos previamente, a veces es más conveniente agrupar los eventos $\mathcal{A}$ en un conjunto $\mathcal{S}$. En algunas disciplinas, particularmente aquellas que estudian personas, tiene sentido no hacer una distinción entre un elemento (a, b) y un elemento (b, a).

Ejemplo 3: Retomando el ejemplo 2

Volviendo al ejemplo anterior, nuevamente $N = 5$, lo que implica un espacio muestral:

$$\Omega = \left\{ \begin{array}{ccccc} \{1,2\} & \{1,3\} & \{1,4\} & \{1,5\} & \{2,3\} \\ \{2,4\} & \{2,5\} & \{3,4\} & \{3,5\} & \{4,5\} \end{array} \right\}$$

Como hay 10 conjuntos, la probabilidad de cualquier **combinación** de 2 elementos es de $\frac{1}{10}$

En general, hay:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} \tag{3}$$

posibles muestras de tamaño n que se pueden obtener **sin** reemplazo de una población de N elementos, donde $k! = k(k-1)(k-2)\ldots 1$ y $0! = 1$.

Volviendo al ejemplo anterior, los posibles resultados son:

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{(2 \times 1)(3 \times 2 \times 1)} = 10$$

Noten que al obtener una muestra $\{a, b\}$ sin reemplazo, b **no es independiente** de a. Esto implica que:

$$\Pr(1 \text{ en el primer resultado, } 2 \text{ en el segundo}) = \frac{1}{5} \times \frac{1}{4} = \frac{1}{20}$$

Sin embargo, $\Pr(\{1,2\}) \neq \Pr(1,2)$. Por tanto,

$$\Pr(\{1,2\}) = \Pr(1,2) + \Pr(2,1) = \frac{1}{20} + \frac{1}{20} = \frac{1}{10}$$

Ejemplo 4: Lotería

Una lotería permite escoger 5 números del 1 al 35. El premio mayor, al acertar los 5 números es de USD $100000. ¿Cuál es la probabilidad de ganar el máximo premio?
El total de posibles resultados es:

$$\binom{35}{5} = \frac{35!}{5!30!} = 324632 \tag{4}$$

De todos estos resultados, únicamente $\binom{5}{5} = 1$ es el ganador. Por tanto, la probabilidad de ganar es de $\frac{1}{324632}$ o 0,00000308.
¿Qué pasa si hay un premio por acertar 4 números?
La lógica de este problema es que se deben acertar 4 números, por lo que los posibles resultados para acertar son $\binom{5}{4}$, pero además, el quinto número puede ser cualquiera de los 30 restantes, o $\binom{30}{1}$. Así

$$\Pr(\text{Acertar 4 números}) = \frac{\binom{5}{4}\binom{30}{1}}{\binom{35}{5}} = \frac{150}{324632} \tag{5}$$

Distribución de probabilidad

Si creamos una lista con todos los posibles resultados en Ω y su probabilidad de ocurrencia, se obtiene una **distribución de probabilidad**. En el ejemplo anterior:

Resultado	Probabilidad
$\{1,2\}$	$\frac{1}{10}$
$\{1,3\}$	$\frac{1}{10}$
$\{1,4\}$	$\frac{1}{10}$
$\{1,5\}$	$\frac{1}{10}$
$\{2,3\}$	$\frac{1}{10}$
$\{2,4\}$	$\frac{1}{10}$
$\{2,5\}$	$\frac{1}{10}$
$\{3,4\}$	$\frac{1}{10}$
$\{3,5\}$	$\frac{1}{10}$
$\{4,5\}$	$\frac{1}{10}$

Las probabilidades de la columna derecha, satisfacen los supuestos presentados en la defición 1. Otro elemento a tomar en cuenta es que **no siempre** las probabilidades son iguales para todos los resultados.

Ejemplo 5: De Bonacich (2012)

La ruleta en Nevada tiene 38 casilleros: 18 rojos, 18 negros y 2 verdes. Los casilleros rojos y negros están enumerados del 1 al 36, en tanto que los verdes son 0 y 00. La tabla a continuación muestra algunas apuestas y sus pagos.

Apuesta	Pago	Probabilidad de ganar
Rojo	\$1	$\frac{9}{19}$
Negro	\$1	$\frac{9}{19}$
Verde	\$17	$\frac{1}{19}$
Número específico	\$35	$\frac{1}{38}$

¿Es correcto que la tercera columna corresponde a la distribución de probabilidad?
En este caso no, principalmente porque no se cumplen los supuestos de la definición 1

$$\frac{9}{19}+\frac{9}{19}+\frac{1}{19}+\frac{1}{38}\neq 1$$

Existen varias formas de representar una distribución de probabilidad, que dependen de si lo que importa es un evento $\mathcal{A}$ o un conjunto $\mathcal{S}$ de resultados.

- Para un evento:

$\mathcal{A}$	$\Pr(\mathcal{A})$
0	$\frac{1}{38}$
00	$\frac{1}{38}$
1	$\frac{1}{38}$
2	$\frac{1}{38}$
$\vdots$	$\vdots$
36	$\frac{1}{38}$

- Para un conjunto de resultados:

$\mathcal{S}$	$\Pr(\mathcal{S})$
Rojo	$\frac{9}{19}$
Negro	$\frac{9}{19}$
Verde	$\frac{1}{19}$

En ambos casos, se cumplen los supuestos de la definición 1.

Variable Aleatoria

Otra forma de ver una distribución de probabilidad es como una función cuyo dominio es el resultado ($\mathcal{A}$ o $\mathcal{S}$) y su recorrido es la probabilidad de que ese resultado ocurra. Esta es la base detrás de la idea de **variable aleatoria**

Definición 2: Variable aleatoria

Una **variable aleatoria** es una función que asigna un número a cada resultado del espacio muestral Ω. El númeron que se asigna depende de un proceso aleatorio, y el conjunto de números sigue los supuestos de la definición 1.
Las variables aleatorias se suelen identificar con letras mayúsculas, en tanto que los resultados individuales con letras minúsculas. La expresión $\Pr(X = x)$ se lee: *la probabilidad que la variable aleatoria X tome el valor de x*.

Dependiendo del recorrido de una variable aleatoria, la clasificaremos como discreta o continua. Si el recorrido de la variable aleatoria pertenece a $\mathbb{Z}$, la clasificaremos como una variable discreta. Por otro lado, si el recorrido de la variable aleatoria pertenece a $\mathbb{R}$, la clasificaremos como continua.

Esperanza o valor esperado

Una cantidad que permite describir una variable aleatoria; o dicho de otra forma, un modelo de dicha variable, es el **valor esperado** (o **esperanza**).

Definición 3: Valor esperado (esperanza) de una variable aleatoria

Sea X una variable aleatoria. Su valor esperado se define como

$$E(X) = \sum_x x\Pr(X = x) \tag{6}$$

Esto se parece mucho a una definición que ustedes ya conocen: el promedio o la media aritmética de una variable. La diferencia (a veces muy sutil), es que mientras el promedio sólo se usa en listas de números (conjuntos, si se quiere), la esperanza se utiliza en variables aleatorias, pues representa un valor que se acerca más al **valor real** en la medida que la cantidad de elementos en la muestra crezca y se acerque a ∞.

Ejemplo 6: Retomando la lotería del ejemplo 4

Supongan que la lotería ya mencionada entrega una cantidad determinada de dinero a partir del número de aciertos. \$50000 por 5 aciertos; \$500 por 4; \$5 por 3; y nada por menos de 3.
Sea X la variable aleatoria que representa la ganancia obtenida cada vez que se participa en el juego, cuya distribución de probabilidad es:

x	$\Pr(X = x)$
0	$\frac{320131}{324632}$
5	$\frac{4350}{324632}$
500	$\frac{150}{324632}$
50000	$\frac{1}{324632}$

Al jugar muchas veces, el valor esperado de ganancias es:

$$E(X) = \left(0 \times \frac{320131}{324632}\right) + \left(5 \times \frac{4350}{324632}\right) + \left(500 \times \frac{150}{324632}\right) + \left(50000 \times \frac{1}{324632}\right)$$
$$= \frac{176750}{324632} = 0{,}45$$

Es decir, luego de muchos juegos, se espera ganar 45 centavos de dólar.

Les dejo como actividad complementaria explicar cómo se llegó a las probabilidades de la columna derecha en la Tabla anterior.

Algunas propiedades del valor esperado

1. Si g es una función,
   $E[g(x)] = \sum_x g(x)\Pr(X = x)$
2. $a, b \in \mathbb{R} \implies E(aX + b) = aE(X) + b$
3. Si X e Y son independientes,
   $E(XY) = E(X)E(Y)$

Otras cantidades relacionadas a $E(X)$

Existen otras cantidades que permiten crear un modelo a partir de una variable aleatoria. Usualmente se utilizan en conjunto, pero toman como base la esperanza.

1. Varianza de X
   $V(X) = E[(X - E(X)^2)]$

2. Covarianza de X e Y
   $Cov(X,Y) = E\left[(X - E(X))(Y - E(Y))\right]$
3. Correlación entre X e Y
   $Corr(X,Y) = \dfrac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$
4. Coeficiente de variación de X
   $CV(X) = \dfrac{\sqrt{V(X)}}{E(X)}$
   para $E(X) \neq 0$

Algunas distribuciones de probabilidad

Hay casos para los que la distribución de probabilidad es conocida para todos sus valores. Incluso, se conoce una **función de probabilidad** que permite asociar un valor cualquiera de la variable aleatoria, con una probabilidad de ocurrencia. Esta idea es bastante conocida, a tal punto que existen libros dedicados únicamente a este tema (Krishnamoorthy, 2016). En este documento, sólo mostraré algunos ejemplos de distribuciones de variables aleatorias discretas, dado que son las más simples de analizar sin recurrir a herramientas más allá de álgebra básico.

Distribución Bernoulli

Pensemos en el experimento más sencillo; aquel donde sólo hay dos resultados posibles: éxito o fracaso. Es decir, $\Omega = \{\text{éxito}, \text{fracaso}\}$:

$$X = \begin{cases} 1 & \text{éxito} \\ 0 & \text{fracaso} \end{cases}$$

Definición 4: Distribución Bernoulli

Si p es la probabilidad constante de éxito de un experimento; y $(1-p)$ la probabilidad de fracaso, entonces X tiene distribución de probabilidad Bernoulli de parámetro p si:

$$X \sim B(p) \begin{cases} f(0) = & \Pr(X = 0) = 1 - p \\ f(1) = & \Pr(X = 1) = p \end{cases}$$

En general, para una variable aleatoria $X \sim B(p)$

$$E(X) = p \tag{7}$$
$$V(X) = p(1-p) \tag{8}$$

Ejemplo No Resuelto 1: Distribución Bernoulli

Un científico poco convencional (a.k.a. científico loco) repite el mismo experimento todos los días. Sabemos que la probabilidad de que su experimento resulte es de 20 %. Denotamos como Z a la variable aleatoria que indica el número de éxitos en el experimento del científico cada día.

Distribución binomial

Supongan que un experimento de Bernoulli se repite n veces bajo ciertas condiciones:

- La probabilidad de éxito es constante entre las repeticiones.
- Las repeticiones son independientes.
- Si X es una variable aleatoria definida como el número de éxitos ocurridos en las n repeticiones independientes del experimento, X tiene una distribución binomial.

Definición 5: Distribución binomial

Si una variable aleatoria X tiene distribución binomial con parámetros n y p, su notación es: $X \sim b(n, p)$. Su función de probabilidad es:

$$f(x) = Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \tag{9}$$

donde $x = 0; 1; 2; 3; \ldots; n$.

Los momentos de la distribución, considerando la misma variable aleatoria X:

$$E(X) = np \tag{10}$$
$$V(X) = np(1-p) \tag{11}$$

Ejemplo No Resuelto 2: Distribución binomial

Mediante un proceso de control de calidad se califica un producto como bueno o defectuoso. Se sabe que el porcentaje de unidades defectuosas es de 4 %. Se elige una muestra de 10 unidades y se examinan en forma independiente. Sea W la variable aleatoria definida como el número de unidades defectuosas encontradas. Se pide:

- Determinar la esperanza y la varianza de W
- Encontrar la probabilidad de que el número de unidades defectuosas encontradas en la muestra sea 3.

Distribución Geométrica

Un experimento Bernoulli de parámetro p se repite sucesivas veces hasta obtener el primer éxito. Los supuestos son que p es constante y las repeticiones son independientes. Sea X la variable aleatoria definida como el número de repeticiones necesarias del experimento hasta obtener el primer éxito.

Definición 6: Distribución Geométrica

Una distribución geométrica para una variable aleatoria X se formaliza como: $X \sim G(p)$ y su función de probabilidad es:

$$f(x) = Pr(X = x) = p(1-p)^{x-1} \tag{12}$$

Los primeros momentos de esta distribución son:

$$E(X) = \frac{1}{p} \tag{13}$$

$$V(X) = \frac{(1-p)}{p^2} \tag{14}$$

Ejemplo No Resuelto 3: Distribución Geométrica

Con los datos del ejemplo de control de calidad, considere que H es una variable aleatoria definida como el número de unidades a inspeccionar hasta obtener la primera unidad defectuosa (primer éxito). A partir de esta información:

- Determinar la esperanza y varianza de H.
- ¿Cuál es la probabilidad de que sea necesario inspeccionar seis unidades hasta encontrar la primera defectuosa?

Referencias

Bonacich, Phillip. 2012. *Introduction to mathematical sociology*. Princeton: Princeton University Press.

Krishnamoorthy, K. 2016. *Handbook of statistical distributions with applications*. Boca Raton: CRC Press.