Semana 2

Muestreo: Analizando ajustes *Valentín Vergara Hidd*

La palabra "ajustes" en el título se refiere a qué tanto una muestra se parece a la realidad. Es decir, vamos a analizar qué tan preciso es el modelo estadístico. No olviden que en este caso, el modelo estadístico es cualquier cantidad que podamos obtener a partir de una muestra, obtenida de una población arbitrariamente grande. Nuevamente, el énfasis no estará en el cáculo de muestras, sino que en el razonamiento estadístico que nos lleva a confiar en conclusiones extraídas a partir de ellas. Todo este documento se complementa con un video donde muestro aplicaciones de esto al simular datos en SPSS.

Parámetros poblacionales y medidas muestrales

En el documento anterior hablamos de variables aleatorias en términos teóricos, con distrihbución de probabilidad conocida y con esperanza y varianza como los primeros momento de la distribución. En esta sección les presentaré la idea de las medidas poblacionales y muestrales.

Partamos por la idea de una población, arbitrariamente numerosa, de donde extraemos una muestra, que por supuesto es un subconjunto de la población, pero que es lo suficientemente grande como para obtener conclusiones significativas a partir de ella. Estoy siendo deliberadamente ambiguo al refererme a una muestra *suficientemente grande*, ya que por ahora no nos va a interesar exactamente de qué tamaño debe ser una muestra.

Obviamente que las medidas en la población y en la muestra no van a ser exactamente iguales. Es por esto que necesitamos definir al menos las medidas mínimas que necesitamos para describir un conjunto de datos: media aritmética y desviación estándar. De ahora en adelante hablaremos de media aritmética y de desviación estándar. Hay una sutil diferencia entre estas cantidades y la esperanza y varianza para una variable aleatoria. Para las últimas mencionadas, efectuamos la medida sobre cantidades teóricas con una distribhución de probabilidad conocida. Para la media artimética y la desviación estándar, no es necesario conocer la distribución de probabilidad de cada elemento de Ω, dado que (al menos teóricamente) se pueden medir en la población y en la muestra.

Una idea importaqnte detrás de todo esto, es que una muestra *perfecta* replica exactamente a la población. Por tanto, en este caso $\mu = \overline{x}$ y $\sigma = s_x$. No es realista pretender que obtendremos esos resultados, pero sí podemos intentar acercarnos.

Definición 1: Medidas en una muestra vs Parámetros poblacionales

Sea n el tamaño de una muestra que se obtiene de una población de N elementos, donde $n << N$. Si X es una variable aleatoria que se mide para toda la población; definiremos la media aritmética de X y su desviación estándar para la muestra y para la población.
Para la muestra, definimos la media aritmética como:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1}$$

mientras que para la población

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{2}$$

La desviación estándar se define para la muestra como:

$$s_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{3}$$

y para la población:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{4}$$

Algunas implicancias de lo anterior:

1. Si pudiésemos medir cualquier cantidad en la población, no serían necesarias las muestras. En circusntancias reales, usualmente sólo es posible acceder a una muestra, cuyas medidas deberían ser lo más parecido a las medidas de la población.
2. Las ecuaciones 1, 2, 3 y 4; además de cualquier otra medida que se defina para la población y la muestra, funcionan en cualquier distribución, no solamente cuando los datos están normalmente distribuidos.
3. Podemos definir un sesgo $\mathcal{B}$ para cualquiera de las dos medidas, definido como la diferencia entre el valor real (parámetro poblacional) y el valor observado en la muestra. Por ejemplo, para la media aritmética, el sesgo sería:

$$\mathcal{B}_{\overline{x}} = \mu - \overline{x} \tag{5}$$

Distribución normal

Probablemente ustedes ya conocen la distribución normal, al menos de nombre. No haré un tratamiento tan detallado de ella como lo hice con la distribución binomial y geométrica, porque tomaría más tiempo del que le podemos dedicar en este curso. Lo que sí les mostraré son algunas propiedades interesantes que nos servirán más adelante, para analizar muestras aleatorias.

Lo primero que observamos es la forma de la distribución, esta *campana* que concentra muchos casos alrededor del centro y progresivamente va disminuyendo a medida que nos acercamos al extremo. Esto lo pueden ver en la figura 1, donde además se puede ver una cantidad al centro: μ

bajo la línea azul, que representa la **media artimética de la población**.

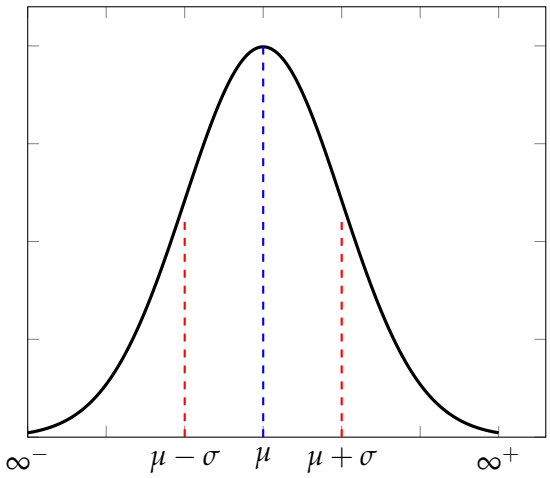

Figura 1: Distribución normal, sin precisar valores

Otra cantidad relevante en la figura 1 es σ, la **desviación estándar**[1] **de la población**. Ambos parámetros, μ y σ, se definen para la población y no necesariamente son los mismos en una muestra. Esto implica que la muestra que obtengamos debería tener una media aritmética y una desviación estándar que se parezcan lo más que se pueda a μ y σ, respectivamente. De esta forma, se garantiza una muestra representativa, al menos en estos dos parámetros.

Por supuesto, esto se puede extender a cualquier medida de una muestra, con su equivalente en la población. Usualmente, basta con la media y la varianza, porque describen lo suficientemente bien a los datos, pero además por razones que explicaré con detalle más adelante (terema del límite central).

La línea negra de la figura 1 se obtiene evaluando la función:

$$\Pr(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{6}$$

para una variable aleatoria X. El dominio de la función es $\mathbb{R}$.

La distribución normal tiene algunas propiedades interesantes[2]. Sólo para mantener consistencia, a continuación utilizaré la notación para la población, por lo que hay que tener en cuenta que todo lo que sigue también se aplica en una muestra con distribución normal.

1. La distribución normal es simétrica, con μ en el centro.

2. La región delimitada por $\mu \pm \sigma$ (la región delimitada por las líneas rojas en la figura 1) contiene aproximadamente el 68 % del área bajo la curva. Dicho de otra forma, si pensamos en una población, aproximadamente el 68 % de los casos presenta valores que se encuentran en este intervalo.

3. La distribución de extiende desde ∞^- a ∞^+, por lo que, teóricamente es posible que la variable con distribución normal tome valores muy grandes (o muy pequeños), pero la probabilidad de esto es mínima.

[1] No olviden que esta cantidad es la raíz cuadrada de la varianza

[2] Existen muchas otras propiedade34s, pero en este documento sólo mostraré aquellas que serán de utilidad para el curso.

Distribución normal estándar

Este caso especial de la distribución normal, sirve para hacer algunas generalizaciones importantes sobre las variables que tienen esta distribución. Por ejemplo, un poco más arriba comenté sobre la región delimitada por $\mu \pm \sigma$. ¿Cómo sabemos que aquí se encuentra aproximadamente el 68 % de los casos?

Para llegar a esta conclusión, consideremos una distribución normal con parámetros $\mu = 0$ y $\sigma = 1$. Esta forma particular de la distribución normal es la **distribución normal estándar** y si la graficamos se debería ver más o menos así:

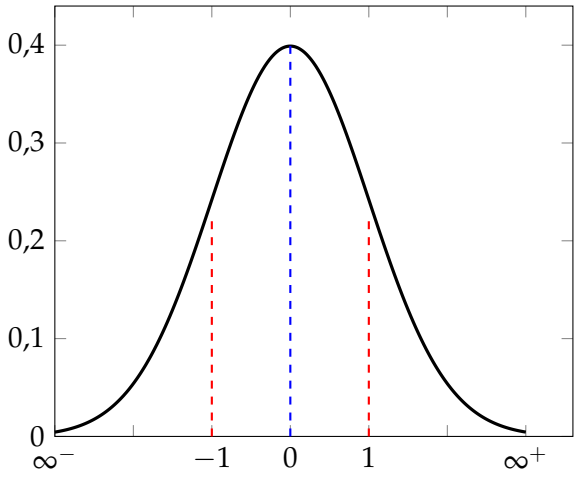

Figura 2: Distribución normal estándar

Al estar estandarizada, sabemos que todos los valores en el eje Y se encuentran en el intervalo $[0, 1]$. También sabemos que cada valor en este eje representa la probabilidad de que la variable aleatoria tome el valor correspondiente en el eje X, cuya función se obtiene reescribiendo la ecuación 6.

$$\Pr(X = x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \tag{7}$$

Simplemente sustituimos los valores de μ y *sigma*. Si además volvemos a la definición de probabilidad del documento anterior, sabemos que la suma de todas las probabilidades de Ω es 1. El problema es que tenemos infinitos valores en Ω[3], lo que hace dificil utilizar una sumatoria como lo hacemos cuando el conjunto Ω tiene un número finito de elementos. La solución, en este caso, es simplemente integrar la función de probabilidad.

$$\int_{\infty^-}^{\infty^+} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx = 1 \tag{8}$$

Sustituyendo el rango en que evaluamos la integral se obtiene:

$$\int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \approx 0{,}64 \tag{9}$$

[3]Esto porque la variable aleatoria X es continua, lo que implica que el dominio de la función de probabilidad es R.

Usualmente, la variable aleatoria con distribución normal estándar se denomina Z, y aunque esto es simplemente una convención, de ahora en adelante utilizaremos de forma muy frecuente esta variable aleatoria, por lo que vamos a reservar este nombre para ella.

Teorema del límite central

Una aplicación inmediata de la distribución normal es presentar el teorema del límite cnetral. DE alguna u otra forma, mesto no va a ser nuevo para ustedes, dado que siempre pensamos nuestras estimaciones en estos términos. Para explicarlo, utilizaré un ejemplo.

Imaginen un experimento simple, como lanzar un dado de 6 caras n veces. Como mencioné en un video previo, y como ustedes pueden fácilmente calcular, si D es la variable aleatoria que contiene el valor obtenido cuando $n = 1$, sabemos que $E(D) = 3,5$. Luego, si aumentamos n y (re)definimos D como la media aritmética de los n lanzamientos del dado, ¿se mantiene el mismo valor de $E(D)$?

La respuesta depende de qué tanto aumentemos n. Cuando es un número *suficientemente grande* (nuevamente estoy siendo deliberadamente ambiguo), podemos esperar que D tome valores aproximados de 3,5. Por tanto, $E(D) \approx 3,5$.

Si efectivamente hacemos el experimento y graficamos los resultados, necesitariamos hacer muchos experimentos y obtener distintos valores de D. Pensemos en N repeticiones del experimento[4] que arrojen resultados $D_1, D_2, D_3, \ldots, D_N$. Obviamente, no todos estos resultados van a ser exactamente 3,5. Lo interesante es que si graficamos los resultados, obtendremos algo que se va a ver más o menos así

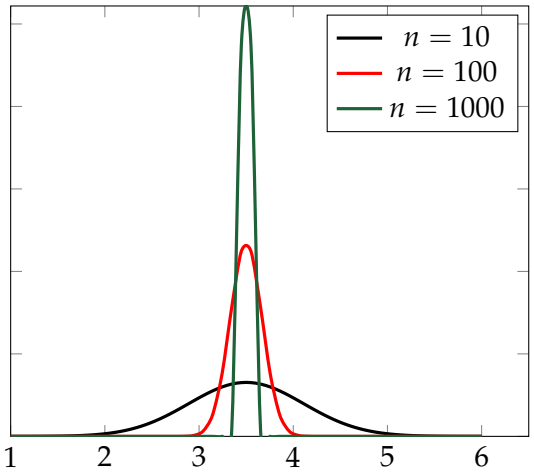

Figura 3: Resultados de N repeticiones de un experimento

Esta es la idea principal del terema del límite central. Si usamos una distribución cualquiera (no tiene por qué ser normal) como la distribución de un dado, una vez que *agregamos* todos los resultados, la distribución de las medias de cada experimento es una distribución normal. En esta distribución, la media muestral tiende al valor de la media poblacional, lo que la hace muy interesante para trabajar con variables cuya distribución es desconocida.

[4] Recuerden que en **cada una** de estas repeticiones, lanzamos el dado n veces; y que $n \neq N$.

Definición 2: Teorema del límite central

Sea X una variable aleatoria, con cualquier distribución. Si se toman N muestras de tamaño n, al obtener los resultados $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_N$, podemos aproximar su distribución a una normal con media μ y desviación estándar $\frac{sigma}{\sqrt{n}}$.
Consideren además que en muchos libros de estadística se puede encontrar un criterio de $n > 30$. Sin embargo, esto responde en parte a una convención, y en parte a que se suele tratar con muestras pequeñas. Mientras mayor sea n, es más fácil obtener la distribución normal de las medias.

Error Estándar

Volvamos a la definición 2. Al obtener la distribución de las medias, podemos también conocer algo sobre su dispersión respecto de μ. Al ver la figura 3 se observa que a medida que aumenta n, esta dispersión disminuye y todos los valores se agrupan de forma más estrecha alrededor de μ. Llamamos a esta dispersión, el **error estçandar de la media**.

Definición 3: Error estándar de la media

Definimos el *error estándar de la media* (sem) como la desviación estándar de la distribución de las medias aritméticas muestrales de distintos experimentos (Definición 2). La expresión

$$\text{sem} = \frac{\sigma}{\sqrt{n}} \tag{10}$$

es particularmente útil, dado que no se necesita tener el conjunto de muestras para calcularla. Sólo es necesario contar con una muestra de tamaño n y con el valor de σ (o alguna forma de estimarlo). De esta forma, es posible además evaluar dos muestras del mismo tamaño y compararlas en relación a su error estándar. Volviendo al ejemplo del dado, probablemente dos muestras presentan valores de $\overline{x}$ muy cerca de μ. Si ese es el caso, simplemente escogeremos como un mejor modelo de los datos reales a aquel que tenga el menor error estándar.

Más adelante utilizaremos variantes del error estándar (no necesariamente el error estándar de la media aritmética) como la base de las pruebas de hipótesis que nos servirán para tomar decisiones.

Con todo esto, estamos e3n condiciones de calcular muestras y evaluar qué tan pertinentes son respecto a la población. Dicho de otra forma, qué tan buenos *modelos* son de los datos reales.

Muestreo

Existen diferentes tipos de muestreo, muchos de los cuales son irrelaventes para este curso, porque siguen criterios que no tienen nada que ver con probabilidad. Los dejaremos de lado, para explorar el principal tipo de muestreo probabilístico.

Definición 4: Muestreo aleatorio simple

ara una población de N elementos, definiremos un subconjunto de n como una muestrea aleatoria simple, si cada posible subconjunto de n elementos tiene la misma probabilidad de ser escogido.
Del documento anterior, recuerden que para un conjunto de N elementos, hay $\binom{N}{n}$ posibles subconjuntos, sin duplicar elementos.
Hay una ligera diferencia entre muestreo aleatorio y muestreo aleatorio simple. En el muestreo aleatorio, cada elemento de la población tiene la misma probabilidad $\frac{1}{N}$ de ser incluído en la muestra; mientras que en el muestreo aleatorio simple, cada *muestra* tiene la misma probabilidad de ocurrencia.

Podemos, entonces, definir $\mathcal{S}_i$ como una muestra, que pertenece al coonjunto de posibles muestras $\{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{\binom{N}{n}}\}$. Si sabemos que todas muestras tienen la misma probabilidad de ocurrencia (y que la suma de estas probabilidades es 1), es posible conocer la probabilidad de escoger una muestra $\mathcal{S}_i$.

$$\Pr(\mathcal{S}_i) = \frac{1}{\binom{N}{n}} = \frac{n!(N-n)!}{N!} \tag{11}$$

Estimación del tamaño muestral.

Si bien, estimar el tamaño muestral de un diseño aleatorio simple no presenta mayor dificultad, el procedimiento requiere establecer un intervalo de confianza (que veremos en el siguiente documento) y un error aceptable. Si nos interesa medir la variable aleatoria X en la muestra, lo que buscamos es encontrar valores es ϵ y α que permitan satisfacer la igualdad

$$\Pr(|\mu - \bar{x}| \leq \epsilon) = 1 - \alpha \tag{12}$$

donde ϵ es el máximo sesgo aceptable $\mathcal{B}$; y α es una proporción del área bajo la curva de una distribución normal estándar Z.

Si asumimos que cada elemento de la población tiene dos posibilidaqdes: es seleccionado para la muestra o no lo es, estamos hablando de una distribución Bernoulli. Esta distribución tiene varianza $p(1-p)$, para p igual a la probabilidad de ser seleccionado. Un poco de álgebra (detallado en el capítulo 2.6 de Lohr (2019)) permite obtener una estimación de tamaño muestral para el caso aleatorio simple:

$$\hat{n} = \frac{Z^2_{\alpha/2} p(1-p)}{\epsilon^2 + \frac{Z^2_{\alpha/2} p(1-p)}{N}} \tag{13}$$

Afortunadamente, es un cálculo que se realiza con la frecuencia suficiente como para que más de alguien lo haya implementado un sitio web, en forma de una calculadora de tamaños muestrales. No quiero profundizar más en este tema, por dos razones: la primera es que aún no hemos tratado con suficiente produndidad los intervalos de confianza; y la segunda es que no es tan importante calcular estas cosas de forma manual, sino que es mucho más necesario entender cómo se derivan sus definiciones.

Para finalizar, un ejercicio interesante es considerar una población, obtener distintos tamaños de muestras utilizando la calculadora o computando los valores de forma manual; para luego comparar estas muestras en función de su error estándar de la media.

Referencias

Lohr, Sharon. 2019. *Sampling : design and analysis*. Boca Raton, FL: CRC Press.