Índice general

Capítulo 1

Primeros pasos

1.1. Estimación de probabilidades

Tanto en el estudio de la probabilidades como en la estadística, trabajamos con experimentos llamados aleatorios. ¿A que nos referimos con esto? Un experimento aleatorio es un experimento para el cual conocemos todos los resultados posibles, pero no sabemos cual de ellos va a ocurrir. El caso más simple sería el caso de tirar una moneda, en el cual tenemos solo dos resultados posibles, y antes de tirarla no sabemos cual de ellos va a ocurrir. Una vez realizado el experimento podremos conocer, para ese caso en particular, cual fue el resultado del mismo. Podemos hacer el mismo análisis con casi cualquier experimento que realicemos. Antes de realizarlo, no sabemos que va a ocurrir. Lo que buscamos al estudiar estadística, es tratar de entender como se comportan las variables objetivo de nuestro estudio, para que en el futuro podamos entender mejor la forma de trabajar con ellas, y podamos inferir o predecir sus resultados, con algún grado de certeza. A eso apuntamos, y para lograrlo tendremos que aprender muchos conceptos importantes.

Primeras definiciones:

- Un *experimento* es un proceso que produce una observación
- Un *resultado* es una posible observación
- A los experimentos estudiados en estadística los llamaremos *experimentos aleatorios*
- Al conjunto de todos los resultados posibles de nuestro experimento aleatorio lo llamaremos ***espacio muestral*** (Ω)
- Un *evento o suceso* es un subconjunto de resultados posibles.

La probabilidad nos ayuda a describir estos experimentos aleatorios, y la estadística nos permite, a partir de observar realizaciones del experimento, inferir sobre el comportamiento de la población en estudio. Empecemos por entender quienes serán nuestras variables.

Una **variable aleatoria** (o variable estadística) será cualquier característica de nuestro experimento que interese registrar y que en el momento de ser registrada pueda ser transformada en un número.

Es muy importante aprender a definir bien nuestras variables, ya que de eso depende la correcta modelización y entendimiento de nuestro experimento.

La forma en la que definimos nuestra variable, da lugar a dos tipos de variables aleatorias:

1. **Variables discretas**: solo pueden tomar una cantidad finita (o infinita numerable) de valores posibles. Incluyen a las variables categóricas.

2. **Variables continuas**: corresponden a una medición que se expresa en unidades y pueden tomar infinitos valores dentro de un intervalo de números reales.

Algunos ejemplos de variables discretas: cantidad de hijos de un paciente, cantidad de accidentes en una fábrica en un año, una variable que vale 0 si un determinado síntoma no se encuentra presente o 1 en caso contrario.

Algunos ejemplos de variables continuas: tiempo que demora en hacer efecto un medicamento, edad del paciente, peso del paciente, tiempo de sobrevida libre de enfermedad.

Ejemplos:
Para los siguientes experimentos, definir la variable, decidir que tipo de variable es y expresar los resultados posibles

1. De una bolsa que contiene 6 frasquitos de medicamentos diferentes, se elije uno al azar y se observa de que medicamento se trata.

2. Se detiene a una persona en la calle y se registra su edad.

3. Supongamos que un semáforo está en rojo por 90 segundos en cada ciclo. Cada vez que un auto llega al semáforo y está en rojo, se mide el tiempo hasta que se pone en verde.

Nuestro siguiente paso será entender como se comportan las variables de estudio. Para poder lograrlo, lo primero que nos viene a la mente es realizar el experimento tantas veces como podamos, y tomar nota de lo que vamos observando en cada una de esas realizaciones. Esto da lugar a más definiciones útiles.

Frecuencia absoluta: Para un experimento en particular, es la cantidad de veces que sucede el evento que estamos observando. Llamemos a dicho evento A.

Frecuencia relativa: Para un experimento en particular, es la relación entre la cantidad de veces que ocurre el evento A y el número total de veces que realicé el experimento, lo llamamos n

$$f_A = \frac{n_A}{n}$$

La **Ley de los grandes números** nos da una herramienta para poder entender porqué la frecuencia relativa será una buena forma de estimar una probabilidad. Desarrollemos un poco más la idea.

La ley de los grandes números dice que si tenemos $\{X_n\}_{n\geq 1}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas, tales que $E(X_i) = \mu$, $var(X_i) = \sigma^2, \forall i$. Entonces si consideremos la sucesión de variables $\{\bar{X}_n\}_{n\geq 1}$ con $\bar{X}_n$ el promedio de las primeras n variables, y con $E(\bar{X}_n) = \mu_i$

entonces se tiene que

$$P(|\bar{X}_n - \mu| > \epsilon) \overset{n\to\infty}{\to} 0$$

Gracias a la desigualdad de Markov, podemos probar la ley débil de los grandes números, que lo que dice en términos de convergencia es que a medida que n aumenta y promediamos más variables, la probabilidad de que $\bar{X}$ se aleje de μ en más de un ϵ es casi cero, y tiende a cero cuando n tiende a infinito.

Esta ley permite fundamentar el concepto de probabilidad de un evento

Si definimos $X_i = \mathbf{I}\{\text{en el experimento i ocurre el evento A}\}$, tenemos que $E(X_i) = P(A)$, $var(X_i) = P(A)(1 - P(A)) < \infty$. Como además las X_i son independientes, por la ley de los grandes números podemos decir que

$$\bar{X}_n \to E(X_1) = P(A)$$

Observemos que $\bar{X}_n$ es la frecuencia relativa de ocurrencia del evento **A** en n realizaciones independientes del experimento, ya que las variables X_i eran variables de Bernoulli, y solo pueden tomar valores 0 y 1. Por lo tanto tenemos que la frecuencia relativa converge a la probabilidad del evento. Utilizando este resultado, podemos justificar que para un n considerablemente grande, calcular la frecuencia relativa con la que ocurre un evento es una buena aproximación para la probabilidad de dicho evento.

Para poder probar este resultado, es muy útil saber simular experimentos y poder ver que ocurre a medida que la cantidad de simulaciones aumenta.Simularemos experimentos, para lo cual necesitaremos un manejo mas o menos bueno de R, (es sumamente recomendable resolver los ejercicios del primer capítulo del libro de Speegle). Para poder simular un experimento, es escencial entenderlo bien, porque le voy a tener que explicar a la computadora como debe realizarlo.

Una de las ventajas de R es que podremos estimar probabilidades usando simulaciones.
Ejemplos a desarrollar en clase usando R:

1. Tiramos un dado y estimamos la probabilidad de que salga el 1

2. Sacamos 2 bolitas de una urna que contiene 5 verdes y 3 rojas, y estimamos la probabilidad de que ambas sean del mismo color. Estimamos también la probabilidad de que la segunda bolita sea roja sabiendo que la primera fue verde.

Podemos observar que a medida que la cantidad de repeticiones del experimento crece, la probabilidad estimada se acerca al valor verdadero de la probabilidad que se busca estimar. Si repetimos muchas veces el ejercicio utilizando diferentes semillas, simulando, vamos a obtener diferentes resultados, pero podremos observar la misma tendencia. Usando esta misma idea podremos simular experimentos mucho más complicados.

1.2. Variables aleatorias

Cuando trabajamos con datos debemos estimar la función de probabilidad, de densidad y de distribución de una variable aleatoria. A partir de ello también entenderemos como estimar la función conjunta para vectores aleatorios, usando la misma idea.

1.2.1. Variables aleatorias discretas

Ya aprendimos a estimar probabilidades de eventos, y eso es lo único que necesitamos para estimar una función de probabilidad. Si defino X como el valor observado al arrojar un dado equilibrado, encontrar su función de probabilidad consiste en econtrar la probabilidad de cada uno de los valores de su rango (o soporte). De la misma forma, estimar la función de probabilidad consistirá en estimar la probabilidad de cada uno de los valores posibles. ¿Cómo? Realizando (o simulando) muchas veces el experimento y calculando la frecuencia relativa para cada uno de los valores posibles. No olvidemos que las reglas que debe cumplir una función de probabilidad también deberá cumplirlas su función estimada.

Una función muy útil en R es la función 'table()'; lo que hace es una tabla en la que cuenta la cantidad de veces que se repite un valor en un vector dado. Es justo lo que necesitamos! Buscamos contar cuantas veces aparece cada valor, y dividir esa cantidad por n, y ese resultado será la frecuencia relativa para cada uno de los valores de X.

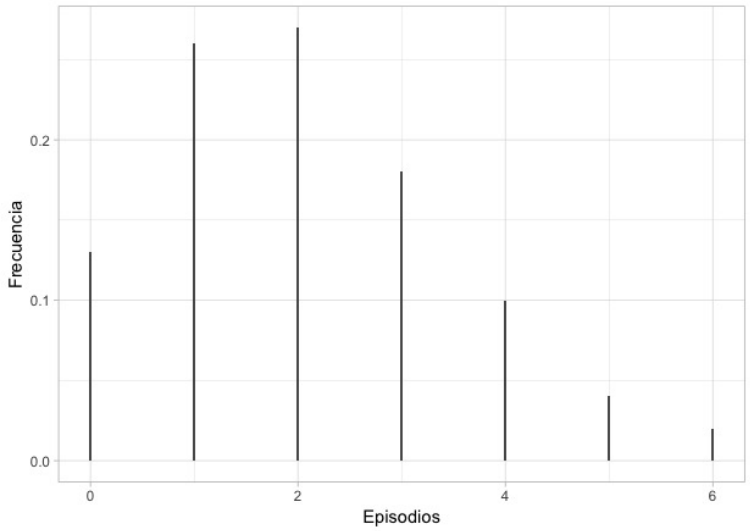
Una vez hallada la función de probabilidad, si buscara por ejemplo la probabilidad $P(X < 4)$, esa estimación resultaría de sumar los valores sobre la función de probabilidad estimada, siendo $\hat{P}(X < 4)$ (usamos el sombrero arriba de la P para decir que el valor es estimado).

Finalmente, me interesa graficar esta función. Siempre nos interesa graficar y poder a partir del gráfico concluir muchísimas cosas. Los gráficos de las funciones de probabilidad estimada se realizan con gráficos de barras, o *bar plot*.

Ejemplo: Supongamos que nos interesa estudiar la cantidad de episodios de otitis que presenta un niño en sus primeros dos años de vida en la Ciudad de Buenos Aires. Para tal fin, se registra el valor de la cantidad de episodios de un grupo de 100 niños de la ciudad de Buenos Aires. Podríamos tener una lista con el nombre de cada niño y el número correspondiente a nuestra variable de estudio, pero resulta mucho más práctico tener la información resumida en una tabla de frecuencias absolutas y relativas.

Cantidad de episodios	Frecuencia absoluta	Frecuencia relativa
0	13	0.13
1	26	0.26
2	27	0.27
3	18	0.18
4	10	0.10
5	4	0.04
6	2	0.02

Para observar la estimación de la función de probabilidad a través de la frecuencia relativa realizamos un gráfico de barras.



Observemos que la variable en estudio, es una variable discreta, y para el caso de variables discretas, los gráficos de frecuencias relativas en forma de gráficos de barras será la mejor forma de representar su comportamiento, en términos estadísticos. Podríamos a partir de dicho gráfico entender si se trata de alguna distribución conocida.

Discusión necesaria: Estimador vs. parámetro.

1.3. Función empírica

La Función empírica es una estimación de la función de distribución acumulada de una variable aleatoria, y su definición es única, para cualquier tipo de variable, tal como lo es la definición de función de distribución. Lo mínimo que puede pedirse a esta función es que cumpla con las condiciones que debe tener una función de distribución:

1. $\hat{F}_X(x) \in [0,1], \forall x \in \mathbb{R}$
2. $\hat{F}_X(x)$ es monótona no decreciente

3. $\hat{F}_X(x)$ es continua a derecha
4. $lim_{x\to-\infty}\hat{F}_X(x) = 0$ y $lim_{x\to\infty}\hat{F}_X(x) = 1$

Se define la Función empírica como

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} 1(x_i \leq x)$$

Donde $x_1, x_2, \ldots, x_n$ se asumen realizaciónes de las variables aleatorias $X_1, X_2, \ldots, X_n$, todas con distribución $F_X(x)$ e independientes.

Cuantas más observaciones tengamos, más se acercaran estas estimaciones a los verdaderos valores de las funciones.

1.3.1. Variables aleatorias continuas

En este caso, buscaremos estimar la función de densidad de la variable aleatoria basada en una muestra aleatoria, y usar esa estimación para estimar probabilidades. Recordemos que una muestra aleatoria es una sucesión de variables $X_1, X_2, \ldots$, de manera tal que todas las X_i son independientes e identicamente distribuídas a X, nuestra variable objetivo del estudio.

Tipos de estimaciones: Parámetricas vs. no paramétricas

En el enfoque paramétrico visto en los cursos de estadística, asumimos que f pertenece a una familia determinada y que solo desconocemos sus parámetros

$$f \in \mathcal{M} = \{f(., \theta), \theta \in \Theta\}$$

En general usamos el método *plug-in*, esto es: dado $\hat{\theta}$ el estimador para θ(por ejemplo el de EMV), entonces $\hat{f}(x) = f_{\hat{\theta}}(x)$

A continuación estudiaremos un enfoque no paramétrico: queremos estimar f sin asumir una determinada forma, solo asumimos que f es suave. La forma mas simple para estimar a una función de densidad es el histograma.

La construcción es bastante simple. Dada una muestra $x_1, x_2, \ldots, x_n$, que se asumen realizaciones de las variables aleatorias $X_1, X_2, \ldots, X_n$, todas con distribución $f(x)$ e independientes, se realizan los siguientes pasos:

- Se selecciona un origen x_0 y se divide la recta real en intervalos de longitud h

  $B_j = [x_0 + (j-1)h, x_0 + jh], j \in \mathbb{N}$

  No es necesario que todos los intervalos tengan la misma longitud, pero es recomendable que así sea. Esto facilita la lectura.

- Se cuenta cuantas observaciones caen en cada intervalo armando una tabla de frecuencias. Denotamos a la cantidad de observaciones que caen en el intervalo j como n_j

- Para cada intervalo, se divide la frecuencia absoluta por la cantidad total de la muestra n (para convertirlas en frecuencias relativas, análogo a como se hace con las probabilidades) y por la longitud h (para asegurarse que el área debajo del histograma sea igual a 1):

  $f_j = \frac{n_j}{nh}$

- Se grafica el histograma realizando una barra vertical sobre cada intervalo con altura f_j y ancho h

Formalmente, el histograma está dado por:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \sum_{j} 1(x_i \in B_j) 1(x \in B_j)$$

Si m_j es el centro del intervalo j, podemos escribir $B_j = [m_j - \frac{h}{2}, m_j + \frac{h}{2})$.

Se puede verificar fácilmente que el área del histograma es igual a 1, propiedad que se requiere para cualquier estimador razonable de una función de densidad.De esa manera, a partir de un histograma podemos estimar una probabilidad calculando áreas, de la misma forma que lo hacemos cuando calculamos probabilidades.

Desventajas al momento de estimar una densidad:

1. Dependiendo del ancho del intervalo dará ideas de densidades diferentes
2. Es muy sensible al punto de inicio del primer intervalo. Para un número fijo de intervalos, la forma puede cambiar simplemente moviendo la ubicación del punto inicial.
3. La densidad estimada no es suave, es escalonada, y esta característica no es propia de la densidad que busca estimar si no de la herramienta que usamos para estimarla.

Es por estas razones que el histograma se usa únicamente para tener una primera visualización.

Veremos ahora un método de estimación que sí es efectivo y es el utilizado para estos fines. Se basa (nuevamente) en la ley de los grandes números. Esté método, llamado estimador de densidades basado en núcleos, busca estimar f a partir de la cantidad de observaciones que caen en un intervalo alrededor del punto x. Esta estimación resulta en funciones más suaves y precisas. Veamos el desarrollo para llegar a la estimación.

Sabemos que si X es una variable aleatoria continua, entonces

$$P(X \in (x-h, x+h)) = \int_{x-h}^{x+h} f_X(t)dt$$

Por otro lado, por la LGN podríamos decir que

$$P(X \in (x-h, x+h)) \overset{\sim}{=} \frac{\#\{X_i \in (x-h, x+h)\}}{n}$$

Por otro lado, cuando aprendimos integrales vimos que, para un h suficientemente pequeño

$$\int_{x-h}^{x+h} f_X(t)dt \cong 2hf_X(x)$$

Uniendo todas estas ideas llegamos a que

$$2hf_X(x) \cong P(X \in (x-h, x+h)) \cong \frac{\#\{X_i \in (x-h, x+h)\}}{n}$$

Por este motivo es que se propone como estimador para la densidad a

$$\hat{f}(x) = \frac{\#\{X_i \in (x-h, x+h)\}}{2hn}$$

Se puede probar que esta función es siempre mayor o igual que 0 y su integral para todos los reales vale 1.

Vamos a reescribir la función usando indicadoras.

$$\hat{f}(x) = \frac{1}{hn}\sum_{i=1}^{n}\frac{1}{2}\mathbf{1}\{x-h < X_i < x+h\} = \frac{1}{hn}\sum_{i=1}^{n}\frac{1}{2}\mathbf{1}\{-1 < \frac{x-X_i}{h} < 1\}$$

Si llamamos $k(t) = \frac{1}{2}\mathbf{1}\{-1 < t < 1\}$ resulta

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right).$$

De esta manera la estimación de f se ve como una suma en la que se usa una función de peso $K : \mathbb{R} \to \mathbb{R}$ llamada *núcleo* (o *kernel*), resultando en el estimador de la densidad por núcleos de la densidad de x de Rosenblatt (1956) y Parzen (1962). A h se lo llama ventana o parámetro de suavizado, y encontrar los valores óptimos de h es la tarea más complicada e importante del método ya que controla el compromiso entre sesgo y varianza del estimador (desarrollaremos en detalle este concepto más adelante).

A partir de este desarrollo podemos definir otros núcleos para que la densidad estimada de más *suave*.

En general, las funciones $K(u)$ son funciones de densidad de probabilidad (integran a uno y son mayores o iguales a cero para u en el dominio de K) continuas, acotadas y simétricas con parámetros de suavizado h que ajustan el tamaño y la forma de los pesos cercanos a x, de esta manera $\hat{f}$ hereda muchas propiedades de K. Algunos ejemplos de núcleos son

1. el núcleo Epanechnikov dado por $K(u) = \frac{3}{4}(1-u^2)I(|u| \leq 1)$,
2. el núcleo bicuadrado definido como $K(u) = \frac{15}{16}(1-u^2)^2 I(|u| \leq 1)$,
3. el núcleo gausiano $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$.

Ejemplo Bufalo.

1.3.2. Elección de la ventana óptima

Ventana de Silvermann

Hay muchas formas para buscar la ventana óptima. La primera es la que busca minimizar el error cuadrático medio. Recordemos que si $f(x)$ es la verdadera función de densidad, buscar un buen estimador es buscar un estimador que minimice el error cuadrático medio, dado por

$$ECM = \mathbf{E}[(f(x) - \hat{f}(x))^2]$$

Como se vió en estadística, desarrollando la definición llegamos a que el ECM se puede descomponer en dos términos, uno correspondiente a la varianza del estimador, y otro correspondiente al sesgo, de manera que

$$ECM = \mathbf{var}(\hat{f}(x)) + B^2(\hat{f}(x))$$

Si calculamos dichos términos para el estimador basado en núcleos, podemos observar que ambos términos dependen de h, de manera que cuando h crece disminuye el sesgo pero la varianza aumenta, y viceversa.

Se puede probar que bajo condiciones generales del núcleo K,

$$B[\hat{f}(x)] \simeq \frac{h^2}{2} C_1(K) f''(x)$$

$$\mathbf{var}[\hat{f}(x)] \simeq \frac{1}{nh} C_2(K) f(x)$$

Compromiso sesgo-varianza (exactitud vs. precisión).

- El sesgo es proporcional a h^2, buscamos un h pequeño. También depende de la derivada segunda de f lo cual mide la curvatura de f en x
- La varianza disminuye a medida que nh crece. Queremos un h o n grandes.

Si $h \to 0$ y $nh \to \infty$, $\hat{f}(x)$ es un estimador consistente de $f(x)$.

Seleccionar la ventana óptima consiste en buscar el equilibrio entre el sesgo y la varianza de manera que el error cuadrático medio sea mínimo.

La ventana dependerá del núcleo, de la función C y de la derivada segunda de f. Como no conocemos a f, Silverman propone reemplazar para el caso normal, encontrando así una aproximación de la ventana óptima, de esta manera resulta la siguiente fórmula:

$$h_{sil} = \left(\frac{4\hat{\sigma}^5}{3n} \right)^{1/5}$$

- Si la densidad es normal, la ventana h_{sil} es óptima.
- Si f no es normal, dará una ventana no muy alejada de la óptima cuando la distribución no es muy diferente a la normal.
- El desvío estandar puede estimarse por el EMV o por el IQR/1.349, de manera que coincida con σ bajo la distribución normal.

Ventana de Validación Cruzada

Otra forma consiste en buscar estimadores de máxima verosimilitud. En estadística vimos que un buen estimador era aquel que maximizaba la probabilidad de observar la muestra, conociendo la forma de la densidad de la población. En este caso no conocemos la densidad, podemos estimarla. Ahora, lo que sucede con el estimador de MV para h es que es degenerado, da 0 resultando en una densidad que da masa 1 a cada dato, y eso no sirve. Es por esto que se usa la idea de validación cruzada para maximizar la pseudo-verosimilitud. ¿en que consiste? En maximizar la log-verosimilitud de la funcion de densidad estimada sin tener en cuenta la i-ésima observación, evaluada en dicha observación que no fue tenida en cuenta. De esa forma evitamos el sobreajuste y obtenemos buenos valores para h. Si observamos $x_1, \ldots, x_n$, tenemos que

$$h^*_{MV} = arg\,\max_h \left(\frac{1}{n} \sum_{i=1}^{n} \ln \hat{f}_h^{(-i)}(x_i) \right)$$

con

$$\hat{f}_h^{(-i)}(x_i) = \frac{1}{(n-1)h} \sum_{i \neq j} K\left(\frac{x_i - x_j}{h} \right).$$

Buscamos el máximo sobre una grilla $h_1, \ldots, h_q$ y luego eventualmente se refina obteniendo mayor precisión.

Para aprender un poco más y encontrar las demostraciones teóricas, se recomienda el libro de Wolfgang Härdle: Nonparametric & semiparamtric models.

1.4. Aprender de los datos

La exploración inicial de los datos es de gran importancia para cualquier análisis estadístico. ¿cómo? Con gráficos:

- Histograma
- Graficos de dispersión (scatter plot)
- Boxplot
- Gráficos de mosaicos

La idea es visualizar en forma rápida las principales características del conjunto de datos, analizando posibles relaciones o conexiones entre ciertas variables. La relevancia que tiene hacer un análisis gráfico previo es fundamental, pero no siempre es determinante. Una gráfica puede facilitar la visualización de relaciones entre variables, pero no confrmarlas.