Resumen Estadística

Marcos Cruz Rodríguez

11 de mayo de 2018

Índice general

Capítulo 1

Introducción a la estadística

En estos apuntes se resume brevemente la parte de estadística del curso y se citan los comandos del software R relacionados con el contenido.

1.1. Introducción

El uso de la estadística estudia la recolección, análisis e interpretación de datos y es útil y necesaria en cualquier trabajo relacionado con la ciencia. La estadística se puede dividir en dos grandes áreas, estadística descriptiva e inferencia estadística. La estadística descriptiva, se dedica a la descripción, representación y resumen de datos originados a partir de los fenómenos de estudio. La estadística inferencial, se dedica a la generación de los modelos, inferencias y predicciones asociadas a los fenómenos en cuestión teniendo en cuenta la aleatoriedad de las observaciones. Se usa para modelar patrones en los datos y extraer inferencias acerca de la población bajo estudio.

Como ejemplo introductorio analizaremos los datos de peso, altura, sexo, fumador/no fumador, grupo sanguíneo y número de hermanos, de todos los alumnos. A través de estos datos introduciremos los conceptos de población, muestra, parámetro y estimación, así como conceptos de inferencia. Veamos para ello los siguientes estudios que se podrían hacer con los datos:

- ¿Cuál es la altura media de la población de chicos universitarios españoles? Se trata de estimar un parámetro poblacional (la altura media) asociado a una variable (la altura) de una población (todos los chicos universitarios españoles) a través de una muestra (los alumnos de la clase de estadística) con unos datos (altura numérica de cada alumno). La estimación podrá ser puntual (utilizando un estimador como la media muestral) o por intervalo (calculando un intervalo de confianza, que veremos durante el curso).

- ¿Es mayor la altura de los chicos que la de las chicas? En este caso queremos comparar un parámetro (altura media) en dos poblaciones (chicos y chicas universitarios españoles). Para sacar conclusiones habría que comenzar por representar gráficamente los datos, para después realizar un test de hipótesis, que explicaremos durante el curso.

- ¿Influye el hecho de ser fumador en la altura de una persona? También aquí

queremos comparar un parámetro en dos poblaciones. Ejercicio: identifica el parámetro, la muestra, la población, el estimador ...

- ¿Están relacionados el peso y la altura? En este caso se trata de buscar una relación entre dos variables de la misma población. Para ello además de representar los datos gráficamente habría que estudiar la correlación entre las variables. De esta manera introducimos otro tema que veremos durante el curso.

Veamos una definición sencilla de los conceptos que acabamos de introducir intuitivamente:

- **Población**: Conjunto bien definido de elementos (en número finito o infinito) del cual deseamos obtener información. Ejemplo: Todos los chicos universitarios españoles.
- **Muestra**: Subconjunto extraido de la población con objeto de obtener información de la población. Ejemplo: Los alumnos de la clase de estadística.
- **Dato**: Observación numérica asociada a cada elemento de la muestra. Ejemplo: Altura de cada alumno de la clase de estadística.
- **Variable Poblacional**: Valor numérico asociado a cada elemento de la población. Ejemplo: Altura de cada universitario español.
- **Parámetro Poblacional**: Cantidad numérica bien definida a partir de la variable poblacional. Generalmente es una cantidad fija y desconocida que se quiere estimar. Ejemplo: Altura media de todos los chicos universitarios españoles.
- **Estimador**: Regla o procedimiento numérico para combinar los datos de una muestra. Equivale a una fórmula matemática (no a un valor numérico concreto). Ejemplo: Sumar todas la alturas de los alumnos de la clase y dividir por el número de alumnos.
- **Estimación**: Valor numérico concreto resultante de aplicar un estimador a los datos de una muestra concreta. Ejemplo: El resultado numérico de sumar todas la alturas de los alumnos de la clase y dividir por el número de alumnos.

Capítulo 2

Estadística descriptiva

2.1. Tipos de datos

Los datos se clasifican en dos grandes grupos, los cuantitativos que son medibles numéricamente, y los cualitativos que no son medibles numéricamente aunque a veces es útil codificarlos mediante una variable numérica. Veamos una clasificación completa con subdivisiones en ambos grupos:

- Cualitativos o Categóricos (no medibles numéricamente)
  - binarios o dicotómicos (solo dos resultados posibles)
    Ejemplos:
    - Sexo: Varón / Mujer
    - Enfermo / No Enfermo
    - Embarazada / No Embarazada
  - nominales (no admiten un orden especial)
    Ejemplos:
    - Estado civil: Soltero / Casado / Viudo / Divorciado
    - Tabaquismo: Fumador / Ex Fumador / No Fumador
    - Enfermedades: Infarto / Diabetes / Alzheimer ...
  - ordinales (admiten un orden)
    Ejemplos:
    - Gravedad: Leve / Moderado / Grave / Crítico
    - Tipo de cirugía: Limpia / Contaminada / Sucia
    - Tabaquismo: No / $1-5$ / $6-10$ / $11-20$ / < 20
- Cuantitativos (medibles numéricamente)
  - discretos (toman valores en un conjunto numerable)
    Ejemplos:
    - Frecuencia cardíaca
    - Número de hijos
    - Número de votantes

- continuos (toman valores en un intervalo contínuo)

  Ejemplos:

  - Altura
  - Temperatura
  - Edad: 20.3

Los datos cualitativos binarios y ordinales suelen ser el resultado de jerarquizar, mientras que los datos cualitativos nominales se obtienen al clasificar. Por otra parte los datos cuantitativos discretos se obtienen al contar y los datos cuantitativos continuos al medir.

Para finalizar planteamos como ejercicio identificar el tipo de datos que hemos tomado en el ejemplo inicial en el que se tomaron datos de los alumnos.

2.2. Representación gráfica

Una buena representación gráfica es clave a la hora de analizar correctamente los datos y en casos muy claros puede incluso hacer innecesario el uso de métodos de inferencia. Una buena gráfica es la que transmite información acerca de los datos y por tanto es importante saber que tipo de representación elegir en función del tipo de datos que tengamos.

Para explicar los distintos tipos de representación gráfica necesitamos definir dos conceptos:

- Frecuencia absoluta: Número de observaciones para un valor (o intervalo de valores) de la variable.

  Ejercicio:

  - Utilizando los datos del ejemplo inicial calcular las frecuencias absolutas para una variable discreta (por ejemplo número de hermanos) y para una variable contínua (por ejemplo altura).

  Comando de R: `table`

- Frecuencia relativa: Es la frecuencia absoluta dividida por el número total de observaciones.

  Ejercicio:

  - Calcular las frecuencias relativas para el ejemplo utilizado anteriormente.

Tras estas definiciones podemos comenzar a describir las distintas técnicas de examen inicial de datos:

- **Diagrama de barras**: Se utiliza para datos categóricos o cuantitativos discretos. Se dibuja el valor de la variable frente a las frecuencias (relativas o absolutas según convenga). Las frecuencias se representan mediante barras.

  Comando de R: `barplot`

- **Diagrama de sectores**: Se utiliza para datos categóricos binarios o nominales. Es un círculo dividido en tantos sectores como valores tome la variable. El área (o el ángulo) del sector es proporcional a la correspondiente frecuencia. Si se desea comparar dos o más muestras se suele representar el área de cada círculo proporcional al número total de observaciones de cada muestra. El diagrama de sectores es una alternativa al diagrama de barras que debe usarse solo cuando se desee comparar las frecuencias frente al total. Visualmente es mucho más difícil comparar áreas relativas que medidas lineales como las barras.

  Comando de R: `pie`

- **Histograma**: Se utiliza para datos cuantitativos continuos. Los datos continuos han de ser agrupados en intervalos para representarlos. Se definen una serie de intervalos de valores de la variable llamados clases o bines. Para cada clase se representa una barra o rectángulo de modo que el área del rectángulo es igual a la frecuencia (absoluta o relativa según convenga). La altura representa la densidad de frecuencia, y el punto medio de cada clase o bin, se denomina marca de clase. Cuando todos los bines tienen la misma anchura, suele tomarse la altura de la barra igual a la frecuencia (absoluta o relativa según convenga).

  Comando de R: `hist`

- **Diagrama de dispersión**: Se utiliza para relacionar dos variables cuantitativas. Se representan en los valores de dos variables para un conjunto de datos en coordenadas cartesianas. El diagrama puede revelar si existe algún tipo de correlación (que definiremos con detalle durante el curso) entre las variables.

  Comando de R: `plot`

- **Diagrama de cajas**: Se utiliza para representar datos cuantitativos y en particular para comparar los datos de dos muestras. El diagrama consiste en una caja que se dibuja usando los valores de los cuartiles (que definiremos más adelante) y unos brazos que se extienden hasta los valores mínimo y máximo de la muestra.

  Comando de R: `boxplot`

Además de representar gráficamente los datos, hay ciertas medidas que pueden dar mucha información acerca de los datos, como son las medidas de centralización o de dispersión.

2.3. Medidas de Centralización

Las medidas de centralización, son medidas alrededor de las cuales se agrupan los datos. Con un solo número resumimos todos los datos dando una idea del orden de magnitud de los datos.

Estas medidas son a menudo un resumen informativo de la muestra y permiten apreciar numéricamente difenercias entre muestras.

Las medidas de centralización más utilizadas y sus propiedades y méritos relativos son:

- **Media**: Para una muestra dada, la media muestral, que llamaremos m, es la suma de los datos (x_i) dividido por el tamaño muestral, n.

$$m = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1}$$

  Comando de R: `mean`

  Para una población la definición es análoga, pero como es una cantidad distinta la denotamos μ y es la suma de todos los datos de la población dividido por el tamaño de la población, N.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.2}$$

- **Mediana**: La mediana es una cantidad tal que el número de datos menores que la mediana es igual al número de datos mayores que la mediana. Para calcular la mediana en muestras de tamaño impar, basta con ordenar la muestra y tomar el dato central. Para muestras con tamaño par, podemos considerar como mediana la media de los dos datos centrales.

  Comando de R: `median`

- **Moda**: La moda es el dato que más veces se repite. En el caso de datos continuos, será la marca de clase del intervalo con la barra más alta.

Propiedades y méritos relativos de las medidas de centralización:

- La media es una medida útil cuando el histograma tiene una sola moda (es decir, es unimodal) y aproximadamente simétrico.
- Si la asimetría es grande, entonces la mediana tiende a ser mucho más estable e informativa que la media.
- La media es muy sensible a la presencia de valores extremos (aunque sean pocos) de la variable, no así la mediana.
- Si el histograma tiene forma de "U" (es decir, es bimodal con modas cerca de los extremos), la media tiende a perder valor informativo.

2.4. Medidas de dispersión

Una medida de posición es en general un resumen insuficiente de un conjunto de datos. Dos histogramas pueden coincidir aproximadamente en su media, mediana y moda, pero diferir claramente en la dispersión de los datos. Las medidas de dispersión reflejan la desviación global de los datos con respecto a una medida central, y por tanto suelen dar una idea de la forma del histograma (por ejemplo, más o menos apuntado, o más o menos plano).

Las medidas de dispersión más utilizadas y sus propiedades y méritos relativos son:

- **Recorrido**: El recorrido es la diferencia entre el mayor y el menor dato de la muestra.

  Comando de R: `range`

- **Varianza**: Como la suma de desviaciones de una muestra de tamaño n con media m es cero:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - m) = 0 \tag{2.3}$$

  definimos varianza muestral como:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2. \tag{2.4}$$

  Comando de R: `var`

  Para una población de tamaño N, y media poblacional μ, definimos la varianza poblacional análogamente:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{2.5}$$

- **Desviación típica**: Como la varianza no tiene las mismas dimensiones que la variable por elevar al cuadrado, definimos la desviación típica muestral que sí tiene las mismas unidades que la variable:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m)^2}. \tag{2.6}$$

  Comando de R: `sd`

  Para una población:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \tag{2.7}$$

- **Coeficiente de variación muestral**: Para una variable no negativa, definimos el coeficiente de variación muestral como:

$$cv = \frac{s}{m} \tag{2.8}$$

  a menudo se expresa como porcentaje:

$$cv\,\% = 100 \times \frac{s}{m}\,\% \tag{2.9}$$

  Si la variable pudiese tomar valores negativos, la media podría ser cero en cuyo caso el coeficiente de variación no estaría definido.

- **Percentiles o cuantiles**: El percentil del $100 \times p\,\%$, es una cantidad tal que la proporción de datos iguales o inferiores a dicho percentil, es igual a p.

  Los percentiles del $25\,\%$, $50\,\%$ y $75\,\%$ se denominan cuartiles, en concreto primer, segundo y tercer cuartil respectivamente. El segundo cuartil es la mediana.

  Comando de R: `quantile`

- **Recorrido intercuartílico**: Es la diferencia entre el tercer cuartil y el primer cuartil.

Propiedades y méritos relativos de las medidas de posición:

- El recorrido tiende a aumentar al crecer el tamaño de la muestra. Si se desea utilizar el recorrido para comparar dos muestras, éstas han de ser tamaños similares.

- La desviación típica es la medida de dispersión más usada. Sin embargo, su utilidad decrece cuando aumenta la asimetría del histograma, en cuyo caso la mediana, los cuartiles y el recorrido intercuartílico pueden ser más interesantes.

- La desviación típica, los cuartiles y el recorrido intercuartílico vienen expresadas en las mismas unidades que los datos de la muestra pero la varianza no.

- El coeficiente de variación es muy útil para comparar el grado de variación de muestras a distinta escala. Ejemplo: Dispersión del peso de ratas comparado con la dispersión del peso de elefantes.

Capítulo 3

Probabilidad

En este tema introduciremos los conceptos básicos de probabilidad para pasar después a las distribuciones de probabilidad más notables como son la binomial, la de Poisson y la normal. La probabilidad es un instrumento esencial para pasar de la estadística descriptiva a la inferencia estadística.

3.1. Conceptos generales de probabilidad

- **Experimento aleatorio**: Es un mecanismo que produce resultados no predecibles con certeza.
- **Espacio muestral**, Ω: Es el conjunto de todos los resultados posibles del experimento aleatorio.
- **Suceso**: Es un subconjunto del espacio muestral, Ω

Ejemplos:

1. Lanzamos un dado y vemos el resultado.
   - Experimento aleatorio: Lanzamiento del dado.
   - Espacio muestral: Conjunto de posibles resultados = $\{1, 2, 3, 4, 5, 6\}$.
   - Suceso: Sale un número par = $\{2, 4, 6\}$.
2. Queremos ver el número de visitas que recibe una página Web en un día concreto.
   - Experimento aleatorio: Contar el número de visitas que recibe una página Web en un día concreto.
   - Espacio muestral: Conjunto de posibles resultados = $\{0, 1, 2, ..., \infty\}$.
   - Suceso: La página recibe 17 visitas.
3. Medir el volumen de materia blanca cerebral en un sujeto.
   - Experimento aleatorio: Medida del volumen la materia blanca del cerebro en cuestión.

- Espacio muestral: $0 \leq$ volumen $\leq \infty$
- Suceso: Medimos 270.2 cm^3.

Los sucesos se pueden combinar mediante uniones, intersecciones y complementos:

- **Unión de dos sucesos**: Si tenemos dos sucesos A y B, su unión se denota $A \cup B$ y significa: *ocurre A, ó B, ó ambos.*
- **Intersección de dos sucesos**: Si tenemos dos sucesos A y B, su intersección se denota $A \cap B$ y significa: *ocurren A y B.*
- **Suceso complementario**: El suceso complementario de un suceso A se denota A^c y significa: *no ocurre A.*

Ejemplo: Lanzamos un dado y consideramos los sucesos A=*Sale un número par* $=\{2, 4, 6\}$, B=*sale un número menor que 3*$=\{1, 2\}$.

- Unión: $A \cup B = \{1, 2, 4, 6\}$
- Intersección $= A \cap B = \{2\}$.
- Suceso complementario de A: A^c=*Sale un número impar* $= \{1, 3, 5\}$.

Antes de ver las definiciones de probabilidad definimos dos conceptos:

- **Suceso elemental**: Cada resultado individual de un experimento aleatorio se llama suceso elemental.
- **Sucesos disjuntos**: Dos sucesos A y B son disjuntos si y solo si $A \cap B = \emptyset$.

Ejemplo: Lanzamos un dado y consideramos los sucesos A=*Sale un número par* $=\{2, 4, 6\}$, B=*Sale un tres* $=\{3\}$. Tenemos que los sucesos A y B son disjuntos ya que $A \cap B = \emptyset$. Por otra parte el suceso B es un suceso elemental, pero A no lo es.

3.2. Definiciones de probabilidad

La probabilidad asocia a un suceso A, un número real comprendido entre cero y uno que dará idea de la frecuencia con la que ocurre A.

3.2.1. Definición frecuentista

Supongamos que un experimento aleatorio se realiza n veces en condiciones idénticas, y que un determinado suceso, A, ocurre en m de esas n ocasiones $(m \leq n)$. La ley de la regularidad estadística fundada en la experiencia, establece que la probabilidad del suceso A es:

$$\mathrm{P}(A) = \lim_{n \to \infty} \frac{m}{n} \tag{3.1}$$

Ejemplo: Al tirar una moneda, la probabilidad del suceso A=*sale cara* es $\mathrm{P}(A) = 0.5$. Si lanzamos una moneda varias veces y vamos anotando el valor del cociente m/n veremos como se va aproximando a 0.5.

3.2.2. Definición axiomática

En esta definición, la probabilidad es una función que asigna a cada suceso un valor basado en los siguientes axiomas:

- $\mathrm{P}(A) \geq 0$, para cualquier suceso $A \subset \Omega$.
- $\mathrm{P}(\Omega) = 1$.
- Si A_1, A_2, ..., son sucesos disjuntos entonces:

$$\mathrm{P}(A_1 \cup A_2 \cup ...) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + ...$$

3.2.3. Probabilidad basada en el modelo de simetría

Si un experimento aleatorio tiene un número finito, N de resultados individuales (o sucesos elementales), y suponemos que cada resultado individual es igualmente probable, entonces la probabilidad de un suceso A compuesto por uno o varios de estos sucesos elementales es:

$$\mathrm{P}(A) = \frac{\text{número de sucesos elementales en A}}{\text{número de sucesos elementales en } \Omega} \tag{3.2}$$

o dicho de otro modo:

$$\mathrm{P}(A) = \frac{\text{número de resultados favorables a A}}{\text{número de resultados posibles}} \tag{3.3}$$

Ejemplo: Lanzamos un dado y consideramos el suceso A=*Sale un número par* $=\{2, 4, 6\}$. La probabilidad del suceso A es:

$$\mathrm{P}(A) = \frac{\text{número de resultados favorables a A}}{\text{número de resultados posibles}} = \frac{3}{6} = \frac{1}{2}$$

3.2.4. Propiedades

A partir de los tres axiomas de la definición axiomática se pueden deducir las siguientes propiedades:

- $\mathrm{P}(A^c) = 1 - \mathrm{P}(A)$
- $\mathrm{P}(A) \leq 1$, para cualquier suceso $A \subset \Omega$.
- $\mathrm{P}(\emptyset) = 0$.
- $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$

Demostración:

- $\left.\begin{array}{c} A \cup A^c = \Omega \\ A, A^c \text{ disjuntos} \end{array}\right\} \Rightarrow \mathrm{P}(A \cup A^c) = \mathrm{P}(A) + \mathrm{P}(A^c) = \mathrm{P}(\Omega) = 1$
- $\left.\begin{array}{c} \mathrm{P}(A) + \mathrm{P}(A^c) = 1 \\ \mathrm{P}(A^c) \geq 0 \end{array}\right\} \Rightarrow \mathrm{P}(A) \leq 1$
- $\mathrm{P}(A) = \mathrm{P}(A \cup \emptyset) = P(A) + P(\emptyset) \Rightarrow \mathrm{P}(\emptyset) = 0.$

- $\left.\begin{array}{l} A = (A \cap B) \cup (A \cap B^c) \Rightarrow P(A \cap B^c) = P(A) - P(A \cap B) \\ A \cup B = (A \cap B^c) \cup B \Rightarrow P(A \cup B) = P(A \cap B^c) + P(B) \end{array}\right\} \Rightarrow$
  $\Rightarrow \mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$

Ejemplo: Lanzamos un dado y consideramos los sucesos A=*Sale un número par* $=\{2,4,6\}$, B=*Sale un número menor que 4* $=\{1,2,3\}$. Calcula $\mathrm{P}(A \cup B)$.

3.3. Probabilidad Condicionada

Ejemplo introductorio: Tiramos dos dados y definimos los siguientes sucesos:

- A=*En el primer dado sale un* 2 .
- B=*La suma de los dos dados es* > 9.
- C=*La suma de los dos dados es* ≤ 5.

Es fácil calcular $\mathrm{P}(A)$ ya que el total de sucesos posibles es 36 y en 6 de ellos ocurre que sale un 2 en el primero de los dados, por tanto $\mathrm{P}(A) = \frac{6}{36} = \frac{1}{6}$.

Imaginemos ahora que los dados los tira otra persona y nos dice que la suma de ambos dados es mayor que 9. ¿Cual es la probabilidad de que haya salido un 2 en el primer dado sabiendo que la suma de ambos dados es mayor que 9? A este suceso lo llamamos:

- $A|B$=*En el primer dado sale un 2, dado que la suma de los dados es* > 9.

En este caso la respuesta también es sencilla ya que es un suceso imposible y por tanto dicha probabilidad es nula, $\mathrm{P}(A|B) = 0$.

Sin embargo para el suceso $A|C$ la respuesta ya no es tan inmediata. Haciendo una tabla, veremos que nuestro espacio muestral se reduce a 10 de los 36 sucesos elementales en los que se cumple C y que en 3 de los 36 sucesos elementales se cumplen tanto A como C. Por tanto la probabilidad sería $\mathrm{P}(A|C) = \frac{\mathrm{P}(A \cap C)}{\mathrm{P}(C)} = \frac{3/36}{10/36} = \frac{3}{10}$.

Veamos pues la **definición** de probabilidad condicionada de $A|B$, es decir de *A dado B*:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \tag{3.4}$$

para todos los sucesos A, B siempre y cuando $\mathrm{P}(B) > 0$.

3.4. Sucesos independientes

Los sucesos A y B son independientes si el hecho de saber que uno de ellos ha ocurrido no influye para nada en la probabilidad de que ocurra el otro. Formalmente,

$$\text{A y B son independientes si y solo si } \mathrm{P}(A|B) = \mathrm{P}(A), \tag{3.5}$$

o bien de manera equivalente,

$$\text{A y B son independientes si y solo si } \mathrm{P}(A \cap B) = \mathrm{P}(A) \times \mathrm{P}(B), \tag{3.6}$$

Ejemplo: En el ejemplo que veiamos en la sección de probabilidad condicionada, los sucesos A y C no son independientes ya que $\mathrm{P}(A) = \frac{1}{6} \neq \mathrm{P}(A|C) = \frac{3}{10}$.

Ejemplo: Consideramos los siguientes sucesos al tirar un solo dado:

- A=*Sale un número par.*
- B=*Sale un número* < 3.

Tenemos que $\mathrm{P}(A) = \frac{3}{6}$ y que $\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{1/6}{2/6} = \frac{1}{2}$, por tanto $\mathrm{P}(A) = \mathrm{P}(A|B)$ y ambos sucesos son independientes.

Un error muy frecuente consiste en creer que el resultado de varios sucesos independientes determinan el siguente. Veamos algunos ejemplos:

- La ruleta ha caído tres veces en rojo, a la siguiente tiene más probabilidades de caer en negro ⇒ ERROR: Son sucesos independientes, los sucesos anteriores no influyen en el siguiente.
- He tenido tres niñas, ahora toca niño ⇒ ERROR: Son sucesos independientes, los sucesos anteriores no influyen en el siguiente.

3.5. Teorema de Bayes

Dados dos sucesos A y B, es fácil comprobar lo siguiente:

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A) \times \mathrm{P}(B|A)}{\mathrm{P}(B|A) \times \mathrm{P}(A) + \mathrm{P}(B|A^c) \times \mathrm{P}(A^c)} \tag{3.7}$$

Veamos la utilidad de esta fórmula.

Ejemplo: Una determinada prueba inmunológica da positivo en el 100 % de los casos si el paciente tiene una cierta enfermedad X, pero también en el 5 % de los casos si el paciente está sano. En una población se sabe que una de cada mil personas tiene la enfermedad X. Una persona elegida al azar da positivo, cual es la probabilidad de que tenga la enfermedad X?

Definamos en primer lugar los siguientes sucesos:

- E = *La persona elegida tiene la enfermedad X.*
- E^c = *La persona elegida está sana*
- Pos = *La persona elegida da positivo*
- Neg = *La persona elegida da negativo*

Por el enunciado conocemos los siguientes datos:

$$\begin{aligned} \mathrm{P}(E) &= 0.001 \\ \mathrm{P}(\mathrm{Pos}|E) &= 1 \\ \mathrm{P}(\mathrm{Pos}|E^c) &= 0.05, \end{aligned}$$

y nos piden $\mathrm{P}(E|\mathrm{Pos})$.

Utilizando la fórmula 3.7 tenemos que:

$$\mathrm{P}(E|\mathrm{Pos}) = \frac{\mathrm{P}(E) \times \mathrm{P}(\mathrm{Pos}|E)}{\mathrm{P}(\mathrm{Pos}|E) \times \mathrm{P}(E) + \mathrm{P}(\mathrm{Pos}|E^c) \times \mathrm{P}(E^c)} = \frac{0.001 \times 1}{1 \times 0.001 + 0.05 \times 0.999} = 0.02$$

La fórmula de Bayes se puede generalizar al caso en el que el espacio muestral Ω cumpla que $\Omega = A_1 \cup A_2 \cup ... \cup A_k$, siendo los sucesos $A_1, A_2, ..., A_k$ sucesos mutuamente disjuntos. En este caso,

$$\mathrm{P}(A_j|B) = \frac{\mathrm{P}(A_j) \times \mathrm{P}(B|A_j)}{\sum_{i=1}^{k} \mathrm{P}(B|A_i) \times \mathrm{P}(A_i)}, \quad (j = 1, 2, ..., k). \tag{3.8}$$

3.6. Variable aleatoria

Una variable aleatoria es una función medible, X, que asigna un valor $X(A)$ a cada suceso A de un espacio muestral Ω.

- Una variable aleatoria es discreta si toma valores en un conjunto numerable (típicamente valores enteros).
- Una variable aleatoria unidimensional es contínua si toma valores en un intervalo de la recta real.

Ejemplos:

- Lanzamos un dado y sale un 3. Podemos identificar:
  - Suceso A: *Al tirar el dado sale un 3*
  - Variable aleatoria discreta X: *número que sale en el dado*. Es una función que transforma cada suceso como el que acabamos de definir en un número de un conjunto numerable (en este caso un entero).
  - Dato: $X(A) = 3$
- Elegimos una persona al azar de entre todos los habitantes de Santander y medimos su altura y mide 168 cm. De nuevo tenemos:
  - Suceso A: *La persona elegida mide 168 cm*
  - Variable aleatoria continua X: *altura en cm*. Es una función que transforma un suceso como el que acabamos de definir, en un número real.
  - Dato: $X(A) = 168$

En muchos casos nos interesará asociar una probabilidad a la variable aleatoria X. Dado un modelo se pueden predecir probabilidades como $\mathrm{P}(140 < X \leq 170)$, $\mathrm{P}(195 < X)$. Para ello es necesario definir algunas funciones.

3.7. Función de distribución

La función de distribución, $F(x)$, se define como:

$$F(x) = \mathrm{P}(X \leq x), \quad x \in \mathbb{R}, \tag{3.9}$$

es decir la función de distribución de una variable aleatoria X, para un valor real x nos da la probabilidad de que X tome un valor igual o inferior a x.

Propiedades de la función de distribución, $F(x)$:

- $\lim_{x\to\infty} F(x) = 1$
- $\lim_{x\to-\infty} F(x) = 0$
- $F(x)$ es no decreciente
- $\mathrm{P}(x_1 < X \leq x_2) = F(x_2) - F(x_1)$

Normalmente, dado un modelo, calcularemos la función de distribución fácilmente con ayuda de un programa o una tabla. Pero, ¿para qué sirve la función de distribución?

Ejemplo: Lanzamiento de un dado. La probabilidad de sacar un número menor o igual que 3 es $F(3) = P(X \leq 3) = 1/6 + 1/6 + 1/6 = 0.5$.

Ejemplo: Si conocemos la función de distribución del colesterol LDL en sangre podemos predecir por ejemplo:

- Probabilidad de que una persona elegida al azar tenga el colesterol LDL por debajo 200 mg/dL $= F(200)$
- Probabilidad de que una persona elegida al azar tenga el colesterol LDL por encima de 240 mg/dL $= 1 - F(240)$
- Probabilidad de que una persona elegida al azar tenga el colesterol LDL entre 200 mg/dL y 240 mg/dL $= F(240) - F(200)$

3.8. Función de probabilidad

La función de probabilidad de una variable aleatoria discreta, X, se define como:

$$p(x) = \mathrm{P}(X = x) \tag{3.10}$$

siendo x uno de los valores discretos que puede tomar la variable aleatoria X.

Ejemplo: Lanzamiento de un dado. El modelo más razonable consiste en suponer que todos los números son igualmente probables, por tanto la función de probabilidad en este caso sería:

$$\begin{aligned}
p(1) &= \mathrm{P}(X = 1) = 1/6 \\
p(2) &= \mathrm{P}(X = 2) = 1/6 \\
p(3) &= \mathrm{P}(X = 3) = 1/6 \\
p(4) &= \mathrm{P}(X = 4) = 1/6 \\
p(5) &= \mathrm{P}(X = 5) = 1/6 \\
p(6) &= \mathrm{P}(X = 6) = 1/6
\end{aligned}$$

3.9. Función densidad de probabilidad

Para una v.a. **continua**, X, la función de probabilidad no nos sirve ya que:

$$\mathrm{P}(X = x) = 0, \quad x \in \mathbb{R}. \tag{3.11}$$

Esto es debido a que en cualquier intervalo de la recta real hay infinitos números, y por tanto la probabilidad de que X valga exactamente x es cero.

Ejemplo: Consideremos la v.a. que nos da la altura de una persona. La probabilidad de que una persona mida $\sqrt{3}$ metros, es cero.

Para variables aleatorias continuas, se utiliza la función densidad de probabilidad, $f(x)$, que es la derivada de la función de distribución, $F(x)$.

$$f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}. \tag{3.12}$$

Propiedades de $f(x)$:

- Como su propio nombre indica, la función densidad de probabilidad *no* es una probabilidad, sino una *densidad* de probabilidad en el punto $X = x$.
- $f(x) \geq 0$, y puede ser mayor que 1.
- Las probabilidades se pueden calcular a partir de ella, calculando el área encerrada bajo la función.

$$\mathrm{P}(a \leq X \leq b) = \int_a^b f(x)\mathrm{d}x = F(b) - F(a). \tag{3.13}$$

- El área total encerrada bajo $f(x)$ es igual a uno.

$$\mathrm{P}(-\infty < X < +\infty) = \int_{-\infty}^{+\infty} f(x)\mathrm{d}x = 1. \tag{3.14}$$

- La función de distribución se puede calcular a partir de la de densidad:

$$\mathrm{P}(-\infty < X \leq x) = \int_{-\infty}^{x} f(x)\mathrm{d}x = F(x). \tag{3.15}$$

Ejemplo: Un tren tiene programada su llegada en los próximos 5 minutos pero no disponemos de más información. Calcula la función densidad de la variable aleatoria tiempo de espera.

3.10. Media o esperanza matemática de una variable aleatoria

Veamos una definición práctica para la esperanza matemática de una variable aleatoria.

Para una variable aleatoria discreta que toma los valores $X = x_1, x_2, ...$ con función de probabilidad $p(x_1), p(x_2), ...$, entonces la esperanza matemática o valor medio de X se expresa:

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} p(x_i)x_i \tag{3.16}$$

Por otra parte, si la variable aleatoria es continua con función de densidad $f(x)$, entonces,

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x. \tag{3.17}$$

Ejemplo: Lanzamiento de un dado. La variable aleatoria X puede tomar los valores $1, 2, ..., 6$ cada uno de ellos con probabilidad $1/6$. El valor esperado es,

$$\mathbb{E}(X) = \sum_{i=1}^{6} p(x_i)x_i = \frac{1}{6}(1+2+3+4+5+6) = 3.5.$$

Ejemplo: Un tren tiene programada su llegada en los próximos 5 minutos pero no disponemos de más información. La función densidad de la variable aleatoria tiempo de espera en minutos, X, es $f(x) = 1/5$ si $x \in [0, 5]$ y $f(x) = 0$ en caso contrario. Por tanto:

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x = \int_0^5 x\frac{1}{5}\mathrm{d}x = \frac{1}{5} \times \frac{x^2}{2}\Big]_0^5 \mathrm{d}x = 2.5 \text{ minutos}$$

3.11. Esperanza de la suma de Variables aleatorias

Si X e Y son dos variables aleatorias con medias finitas, entonces la media de la suma es:

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y). \tag{3.18}$$

Demostración: Mostramos la demostración para variables aleatorias discretas, para variables aleatorias continuas se obtiene fácilmente un resultado equivalente. Para cada suceso $\omega \in \Omega$ tenemos que la variables aleatorias $X, Y, X+Y$ asignan valores reales $X(\omega), Y(\omega), (X+Y)(\omega)$ respectivamente. Además, $(X+Y)(\omega) = X(\omega) + Y(\omega)$. Por tanto, aplicando 3.18:

$$\begin{aligned}
\mathbb{E}(X+Y) &= \sum_{\omega\in\Omega} (X+Y)(\omega)\mathrm{P}(\omega) \\
&= \sum_{\omega\in\Omega} X(\omega)\mathrm{P}(\omega) + \sum_{\omega\in\Omega} Y(\omega)\mathrm{P}(\omega) \\
&= \mathbb{E}(X) + \mathbb{E}(Y)
\end{aligned}$$

Ejemplo: Tiramos dos dados, el primero X con esperanza $\mathbb{E}(X) = 3.5$; y el segundo, Y con esperanza $\mathbb{E}(Y) = 3.5$. La suma de ambos dados $(X+Y)$ tiene esperanza $\mathbb{E}(X) + \mathbb{E}(Y) = 7$.

3.12. Esperanza de una función de una variable aleatoria

Sea X una variable aleatoria discreta; la esperanza matemática de una función $Y = g(X)$ es:

$$\mathbb{E}(Y) = \mathbb{E}(g(X)) = \sum_{i=1}^{\infty} g(x_i)p(x_i), \tag{3.19}$$

Para una v.a. continua, X:

$$\mathbb{E}(g(X)) = \int_{-\infty}^{+\infty} g(x)f(x)(d)x \tag{3.20}$$

Por lo tanto tenemos:

- $\mathbb{E}(aX) = a\mathbb{E}(X)$ para cualquier constante a.
- $\mathbb{E}(g(X) + h(X)) = \mathbb{E}(g(x)) + \mathbb{E}(h(x))$ para dos funciones cualesquiera g, h.

Ejemplo: $\mathbb{E}(aX^2+bY^2+c) = a\mathbb{E}(X^2)+b\mathbb{E}(Y^2)+c$. Siendo a, b, c constantes.
Ejemplo: Lanzamiento de un dado. La variable aleatoria X puede tomar los valores $1, 2, ..., 6$ cada uno de ellos con probabilidad $1/6$. El valor esperado de la variable $Y = 2X^2$ es:

$$\mathbb{E}(2X^2) = \sum_{i=1}^{6} p(x_i)2x_i^2 = \frac{91}{3}.$$

Ejemplo: Un tren tiene programada su llegada en los próximos 5 minutos pero no disponemos de más información. La función densidad de la variable aleatoria tiempo de espera en minutos, X, es $f(x) = 1/5$ si $x \in [0, 5]$ y $f(x) = 0$ en caso contrario. Por tanto el valor esperado de la variable $Y = 2X^2$ es:

$$\mathbb{E}(2X^2) = \int_{-\infty}^{+\infty} xf(x)\mathrm{d}x = \int_0^5 2x^2\frac{1}{5}\mathrm{d}x = \frac{2}{5} \times \frac{x^3}{3}\Big]_0^5 \mathrm{d}x = 50/3 \text{ minutos}^2.$$

3.13. Varianza y desviación típica de una v.a.

La varianza de una v.a., X, es la esperanza matemática del cuadrado de la variable aleatoria $(X - \mathbb{E}(X))$

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2. \tag{3.21}$$

- En el caso discreto: $\mathrm{Var}(X) = \sum_{i=1}^{\infty} p(x_i)(x_i - \mathbb{E}(X))^2$
- En el caso continuo: $\mathrm{Var}(X) = \int_{-\infty}^{+\infty}(x - \mathbb{E}(X))^2 f(x)\mathrm{d}x$

La desviación típica es:

$$\mathrm{SD}(X) = \sqrt{\mathbb{E}(X - \mathbb{E}(X))^2}. \tag{3.22}$$

Propiedades de la varianza:

- $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$
- $\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$

Para aligerar la notación usamos $\mathbb{E}X = \mathbb{E}(X)$, $\mathbb{E}X^2 = \mathbb{E}(X^2)$.
Demostración:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}(X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2) \\ &= \mathbb{E}X^2 - 2(\mathbb{E}X)^2 + (\mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2\end{aligned}$$

$$\begin{aligned}\mathrm{Var}(aX + B) &= \mathbb{E}\big(aX + b - \mathbb{E}(aX + b)\big)^2 = \mathbb{E}\big(aX + b - a\mathbb{E}X - b\big)^2 \\ &= \mathbb{E}\big(a^2(X - \mathbb{E}X)^2\big) = a^2\mathrm{Var}(X)\end{aligned}$$

Ejemplo: Lanzamiento de un dado. La variable aleatoria X puede tomar los valores $1, 2, ..., 6$ cada uno de ellos con probabilidad $1/6$. La varianza de X vale:

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{91}{6} - (3.5)^2 \approx 2.92$$

Ejemplo: Un tren tiene programada su llegada en los próximos 5 minutos pero no disponemos de más información. La función densidad de la variable aleatoria tiempo de espera en minutos, X, es $f(x) = 1/5$ si $x \in [0, 5]$ y $f(x) = 0$ en caso contrario. La varianza de X vale:

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{50}{6} - (2.5)^2 = \frac{25}{12} \text{ minutos}^2$$

3.14. Covarianza de dos variables aleatorias

Si X e Y son dos variables aleatorias con medias y varianzas finitas, la covarianza es:

$$\mathrm{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right]. \tag{3.23}$$

Se puede demostrar que la siguiente expresión es equivalente:

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y). \tag{3.24}$$

Demostración: Ejercicio.

3.15. Varianza de la suma de Variables aleatorias

Para dos variables aleatorias X e Y, se cumple que:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y). \tag{3.25}$$

Demostración:

$$\begin{aligned}\mathrm{Var}(X + Y) &= \mathbb{E}\left[(X + Y) - \mathbb{E}(X + Y)\right]^2 = \mathbb{E}(X + Y - \mathbb{E}X - \mathbb{E}Y)^2 \\ &= \mathbb{E}\left[(X - \mathbb{E}X)^2 + (Y - \mathbb{E}Y)^2 + 2(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right] \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y).\end{aligned}$$

3.16. Variables aleatorias independientes

Las variables aleatorias X e Y son independientes si y solo si para cada par de valores reales x, y los sucesos A, B que satisfacen $X(A) = x, Y(B) = y$ son independientes.

X, Y son independientes si $\forall A, B \subset \Omega$ se verifica $P\big((X \in A) \cap (Y \in B)\big) = P(X \in A)P(Y \in B)$

Para dos variables aleatorias X e Y independientes tenemos las siguientes **propiedades**:

- $\mathbb{E}(XY) = \mathbb{E}X \cdot \mathbb{E}Y$
- $\mathrm{Cov}(X, Y) = 0$
- $\mathrm{Var}(X + Y) = \mathrm{Var}X + \mathrm{Var}Y$

Demostración:

-
$$\begin{aligned} \mathbb{E}(XY) &= \sum_x \sum_y xyP(X = x, Y = y) \\ &= \sum_x \sum_y xP(X = x)yP(Y = y) \\ &= \mathbb{E}(X)\mathbb{E}(Y) \end{aligned}$$

- $\mathrm{Cov}(X, Y) = 0$ se deduce inmediatamente de la anterior propiedad y de 3.24. Sin embargo $\mathrm{Cov}(X, Y) = 0 \nRightarrow X, Y$ independientes
- $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ se deduce de la anterior propiedad y de 3.25.

Capítulo 4

Modelos de variables aleatorias: Distribuciones de probabilidad

A continuación pasamos a ver distintos modelos de variables aleatorias. Modelar una variable aleatoria es proponer una fórmula para su función de distribución (o su función de probabilidad o densidad). Cuando se da un modelo para una variable aleatoria se suele hablar de *distribución* de la variable aleatoria.

La elección del modelo se puede basar en:

- El comportamiento de los datos en muestras suficientemente grandes (idealmente infinitas).
- Una simplificación del mecanismo de generación de la variable aleatoria (como por ejemplo suponer en el lanzamiento de una moneda que el 50 % de las veces saldrá cara y el 50 % cruz.)

Un modelo bien elegido:

- Permite predecir la probabilidad de cada valor de la variable en el caso discreto, o de cualquier intervalo en el caso continuo.
- Facilita la estimación de los parámetros de una población, como son la media y la varianza.

4.1. Distribución Uniforme discreta

La variable aleatoria uniforme discreta X, puede tomar un número finito de valores equiprobables. Es decir, X toma los valores reales $x_1, x_2, ..., x_n$ con función de probabilidad:

$$p(x_i) = \frac{1}{n}, \quad i = 1, ..., n \tag{4.1}$$

En el caso $x_1 = 1, x_2 = 2, ..., x_n = n$, el valor medio y la varianza la variable aleatoria uniforme discreta son,

$$\mathbb{E}(X) = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_n + x_1}{2} \tag{4.2}$$

$$\mathrm{Var}(X) = \frac{n^2 - 1}{12} \tag{4.3}$$

Demostración:

- Para la esperanza tenemos:

$$\mathbb{E}(X) = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n}\frac{n(n+1)}{2} = \frac{n+1}{2} \tag{4.4}$$

- En el caso de la varianza:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{1}{n}\sum_{i=1}^{n} i^2 - \left(\frac{(n+1)}{2}\right)^2 \\
&= \frac{1}{n}\frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} \\
&= (n+1)\left(\frac{2n+1}{6} - \frac{n+1}{4}\right) = \frac{n^2-1}{12}
\end{aligned}$$

Ejemplo: En un dado de seis caras, el número que sale en el dado es una variable aleatoria uniforme discreta que toma los valores $1, 2, ..., 6$ con probabilidad $1/6$, y valor medio y varianza:

$$\begin{aligned}
\mathbb{E}(X) &= \frac{n+1}{2} = 3.5 \\
\mathrm{Var}(X) &= \frac{n^2-1}{12} = \frac{35}{12}.
\end{aligned}$$

4.2. Distribución Uniforme continua

La variable aleatoria uniforme continua X, puede tomar cualquier valor real en el intervalo $[a, b]$ siendo todos los valores del intervalo equiprobables. La función de densidad es:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \le x \le b \\ 0 & \text{en caso contrario} \end{cases} \tag{4.5}$$

Comandos de R: `dunif, punif, qunif`

El valor medio y la varianza la variable aleatoria uniforme continua son,

$$\mathbb{E}(X) = \frac{b+a}{2} \tag{4.6}$$

$$\mathrm{Var}(X) = \frac{(b-a)^2}{12}. \tag{4.7}$$

Demostración: Ejercicio.

Ejemplo: Un tren tiene programada su llegada en los próximos 5 minutos pero no disponemos de más información. El tiempo de espera es una variable uniforme en el intervalo [0,5]. El tiempo medio de espera y la varianza, en minutos y minutos al cuadrado respectivamente, son:

$$\begin{aligned} \mathbb{E}(X) &= \frac{b+a}{2} = \frac{5}{2} = 2.5 \text{ minutos} \\ \text{Var}(X) &= \frac{(b-a)^2}{12} = \frac{25}{12} \text{ minutos}^2. \end{aligned}$$

4.3. Distribución de Bernoulli

La variable aleatoria de Bernoulli X corresponde a un experimento binario, es decir, X toma los valores 1 y 0 con probabilidades p y $1-p$ respectivamente. Por lo tanto su función de probabilidad se puede escribir,

$$p(x) = p^x(1-p)^{1-x}, \quad x = 0, 1. \tag{4.8}$$

El valor medio y la varianza la variable aleatoria de Bernoulli son,

$$\mathbb{E}(X) = p \tag{4.9}$$
$$\text{Var}(X) = p(1-p). \tag{4.10}$$

Demostración:

$$\mathbb{E}(X) = \sum_i x_i p(x_i) = 0 \times (1-p) + 1 \times p = p$$

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}X^2 - (\mathbb{E}X)^2 = \sum_i x_i^2 p(x_i) - p^2 \\ &= 0^2(1-p) + 1^2 p = p(1-p). \end{aligned}$$

Ejemplo: Aplicamos un tratamiento a un enfermo sabiendo que la probabilidad de curación es $p = 0.25$. Definimos la variable aleatoria de Bernoulli X que puede tomar dos valores, $X = 1 \Rightarrow$ curación, $X = 0 \Rightarrow$ fracaso terapéutico. Por tanto tenemos,

$$\begin{aligned} p(0) &= \text{P}(X=0) = (0.25)^0(1-0.25)^{1-0} = 0.75 \\ p(1) &= \text{P}(X=1) = (0.25)^1(1-0.25)^{1-1} = 0.25. \end{aligned}$$

4.4. Distribución binomial

La variable aleatoria binomial es la suma de n variables aleatorias independientes de Bernoulli.

Repetimos n veces un experimento binario con dos posibles resultados cada vez, 1 = éxito y 0 = fracaso, con probabilidades p y $1-p$ respectivamente. La

variable aleatoria binomial X representa el número total de éxitos obtenidos en los n ensayos independientes.

La función de probabilidad de la variable binomial X, es:

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, ..., n. \tag{4.11}$$

El coeficiente $\binom{n}{x} = \frac{n!}{(n-x)!x!}$ aparece para sumar todas las formas de combinar los x éxitos y $n - x$ fracasos.

Comandos de R: `dbinom, pbinom, qbinom`

El valor medio y la varianza la variable aleatoria binomial son,

$$\mathbb{E}(X) = np \tag{4.12}$$
$$\text{Var}(X) = np(1-p). \tag{4.13}$$

Demostración: Sean $Y_1, Y_2, ..., Y_n$ variables de Bernoulli independientes $\Rightarrow$ $X = \sum_i Y_i$

$$\mathbb{E}(X) = \mathbb{E}\sum_i Y_i = \sum_i \mathbb{E}Y_i = np.$$

$$\text{Var}(X) = \text{Var}\sum_i Y_i = \sum_i \text{Var}Y_i = np(1-p)$$

Ejemplo: Aplicamos un tratamiento independientemente a 4 enfermos sabiendo que la probabilidad de curación en cada uno de ellos es $p = 0.25$. La variable aleatoria binomial representa el número de éxitos de los 4 pacientes tratados. Por tanto tenemos,

$$\begin{aligned}
p(0) &= \text{P}(X=0) = \binom{4}{0}(0.25)^0(1-0.25)^{4-0} = 0.316 \\
p(1) &= \text{P}(X=1) = \binom{4}{1}(0.25)^1(1-0.25)^{4-1} = 0.422 \\
p(2) &= \text{P}(X=2) = \binom{4}{2}(0.25)^2(1-0.25)^{4-2} = 0.211 \\
p(3) &= \text{P}(X=3) = \binom{4}{3}(0.25)^3(1-0.25)^{4-3} = 0.047 \\
p(4) &= \text{P}(X=4) = \binom{4}{4}(0.25)^4(1-0.25)^{4-4} = 0.004.
\end{aligned}$$

4.5. Distribución de Poisson

Una variable aleatoria de Poisson, X, expresa la probabilidad que un determinado número de eventos ocurran en un determinado intervalo de tiempo (o de espacio), dada una frecuencia media conocida e independientemente del tiempo discurrido (o espacio recorrido) desde el último evento. Su función de probabilidad es:

$$p(x) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!}, \quad x = 0, 1, ... \tag{4.14}$$

siendo λ un número real y positivo que representa el número medio de sucesos en el intervalo de tiempo (o de espacio).

COMANDOS DE R: `dpois, ppois, qpois`

La función de probabilidad 4.14 se puede obtener como el límite de la función de probabilidad Binomial 4.11 cuando $n \to \infty$, $p \to 0$ y considerando $\lambda = np$ constante.

$$\begin{aligned}\binom{n}{x} p^x (1-p)^{n-x} &= \frac{n(n-1)\cdots(n-x-1)}{x(x-1)\cdots 3 \cdot 2 \cdot 1} \left(\frac{\lambda}{n}\right)^x (1-p)^{-x} \left(1 - \frac{\lambda}{n}\right)^n \\ &= \frac{n}{n} \cdot \frac{n-1}{n} \cdots \frac{n-x-1}{n} \cdot \left(\frac{\lambda}{x!}\right)^x (1-p)^{-x} \left(1 - \frac{\lambda}{n}\right)^n.\end{aligned}$$

Cuando $n \to \infty$, $p \to 0$ los x primeros factores tienden a 1, el siguiente factor permanece igual, el siguiente tiende a 1, y sustituyendo $k = -\lambda/n$ o bien $n = -\lambda/k$, el último factor es:

$$(1+k)^n = \left[(1+k)^{1/k}\right]^{-\lambda} \to \mathrm{e}^{-\lambda}$$

por lo tanto con $np = \lambda$, fijo,

$$\lim_{\substack{n\to\infty \\ p\to 0}} \binom{n}{x} p^x (1-p)^{n-x} = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} \tag{4.15}$$

Tomando límites en los valores binomiales, se demuestra fácilmente que el valor medio y la varianza de la v.a. de Poisson son,

$$\mathbb{E}(X) = \lambda \tag{4.16}$$

$$\mathrm{Var}(X) = \lambda. \tag{4.17}$$

Ejemplo: En una población, el número medio de casos de leucemia es 5 al año. ¿Cuál es la probabilidad de que en 2012 haya 8 casos?

$$p(8) = \frac{5^8 \mathrm{e}^{-5}}{8!} = 0.065.$$

Hay númerosos casos en los que se puede aplicar la distribución de Poisson:

- Número de accidentes en un tramo de carretera
- Número de enfermos se atienden en urgencias cada día
- Número de infartos que se producen en un año
- Número de visitas a una página web

Es importante recalcar que los sucesos han de ser independientes par que podamos aplicar la distribución de Poisson correctamente.

4.6. Distribución normal

La distribución normal (o Gaussiana) juega un papel central en estadística ya que es un modelo aceptable para muchos tipos de datos biológicos, para errores de medida o desviaciones entre valores observados y teóricos. Algunos ejemplos de variables asociadas a fenómenos naturales que siguen el modelo de la normal son:

- Errores cometidos al medir ciertas magnitudes
- Nivel de ruido en telecomunicaciones
- Caracteres morfológicos de individuos como la estatura
- Caracteres fisiológicos como el efecto de un fármaco
- Caracteres sociológicos como el consumo de cierto producto por un mismo grupo de individuos
- Caracteres psicológicos como el cociente intelectual

La validez y utilidad del modelo se basan en el teorema central del límite que enunciamos más adelante.

La variable aleatoria normal, X, es continua, a diferencia de las variables aleatorias binomial y de Poisson son discretas. La función de densidad de la variable aleatoria normal de media $\mathbb{E}(X) = \mu$ y varianza $\mathrm{Var}(X) = \sigma^2$, es

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp - \left\{ \frac{(x-\mu)^2}{2\sigma^2} \right\}. \tag{4.18}$$

Comandos de R: `dnorm, pnorm, qnorm`

Es importante tener en cuenta las siguientes propiedades de la distribución normal:

- La notación $X \sim N(\mu, \sigma^2)$ se utiliza para expresar *la variable aleatoria X tiene distribución normal con media μ y varianza σ^2.*
- La función de distribución normal se puede calcular fácilmente con R usando el comando `pnorm`.

  Ejemplo: Supongamos una población de altura media 175 cm y desviación típica 10 cm, en la que queremos saber qué porcentaje de la población tiene una altura menor que 160 cm. Para calcular $\mathrm{P}(X \leq 160) = F(160)$ usamos el comando `pnorm(160, 175, 10)`, siendo el resultado $\mathrm{P}(X \leq 160) = 0.067$. Es decir que el porcentaje de la población que tiene una altura menor que 160 cm es el 6.7 %.

- Para calcular la función de distribución normal también pueden utilizarse tablas. En las tablas encontraremos los valores de la función de distribución de la variable normal tipificada, Z:

$$Z = \frac{X - \mu}{\sigma}, \tag{4.19}$$

que tiene media $\mu = 0$ y varianza $\sigma^2 = 1$, es decir $Z \sim N(0, 1)$. La función de distribución de Z se denota $\Phi(z)$.

- Las probabilidades de X se pueden calcular fácilmente a partir de las de Z

$$F(x) = P(X \leq x) = P(Z \leq \frac{x - \mu}{\sigma}) = \Phi(z)$$

- Muchas veces necesitaremos calcular los percentiles de una variable $X \sim N(\mu, \sigma^2)$. Llamaremos $x(p)$ al percentil p de X, es decir que $x(p)$ es un número tal que $F(x(p)) = p$.

  Ejemplo: Supongamos una población de altura media 175 cm y desviación típica 10 cm, en la que queremos saber la altura por debajo de la cual se encuentre el 5 % de la población. Dicha altura es el percentil $x(p)$ con $p = 0.05$, y se puede calcular con el comando `qnorm(0.05, 175, 10)`, siendo el resultado $x_{0.05} = 158.55$ cm.

- Los percentiles de $Z \sim N(0, 1)$ también se encuentran tabulados. Llamaremos $z(p)$ al percentil p de Z, es decir que $z(p)$ es un número tal que $\Phi(z) = p$. Es fácil calcular $x(p)$ a partir de $z(p)$ ya que $z(p) = \frac{x(p)-\mu}{\sigma}$.

- La distrubución normal es simétrica y por tanto:

$$P(X < x - \mu) = P(X > x + \mu)$$

Veamos ahora por qué razón la distribución normal aparece tan a menudo en fenómenos naturales.

Capítulo 5

Muestreo y Estimación Puntual

Un problema estándar en inferencia estadística es estimar un parámetro de una población. Como hemos visto para ello se extrae una muestra de tamaño n, y aplicamos un estimador para estimar el parámetro.

Por ejemplo si queremos estimar una media poblacional, un estimador adecuado es la media muestral $\overline{X}$.

Hablamos de estimación puntual cuando el estimador nos da un solo valor para estimar el parámetro.

Veamos más en detalle el proceso de muestreo y las características de los estimadores.

5.1. Muestreo aleatorio simple (sin reposición)

La idea es númerar cada unidad de la población y elegir dorsales por el método de la lotería hasta obtener el tamaño muestral deseado. Sin reposición quiere decir que, una vez extraída, la unidad no se devuelve a la población.

Para realizar el muestreo correctamente necesitaremos generar números aleatorios, bien mediante tablas o bien mediante un programa informático. Si el muestreo no se realiza al azar la calidad del estimador disminuye y podremos llegar a conclusiones erróneas.

Comandos de R: `sample`

Una alternativa válida al muestreo aleatorio simple es el muestreo sistemático, pero no vamos a describirlo en este curso.

5.2. Muestras y variables aleatorias

Matemáticamente, una muestra de tamaño n de una variable aleatoria X, es un conjunto de n realizaciones de la variable aleatoria X, es decir el conjunto $\{X_1, X_2, ..., X_n\}$. Cada una de las n realizaciones es una variable aleatoria con la misma distribución que X.

Una muestra aleatoria simple es aquélla en la que cada unidad de muestreo se extrae de forma independiente, es decir en la que las variables $\{X_1, X_2, ..., X_n\}$ son independientes en el caso de población infinita.

5.3. Media muestral

La media muestral definida en una muestra $\{X_1, X_2, ..., X_n\}$ de una variable aleatoria X, como

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad i = 1, ..., n \tag{5.1}$$

es también una variable aleatoria con media

$$\mathbb{E}(\overline{X}) = \mathbb{E}(X). \tag{5.2}$$

Demostración:

$$\begin{aligned} \mathbb{E}(\overline{X}) &= \mathbb{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\mathbb{E}(\sum_{i=1}^{n} X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mathbb{E}(X). \end{aligned}$$

Si la muestra es aleatoria simple y la población infinita, la varianza de $\overline{X}$ es,

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}(X)/n. \tag{5.3}$$

Demostración:

$$\begin{aligned} \mathrm{Var}(\overline{X}) &= \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \\ &= \frac{1}{n^2} n\mathrm{Var}(X) = \frac{\mathrm{Var}(X)}{n}. \end{aligned}$$

Además la **ley de los grandes números** afirma que $\overline{X}$ converge a $\mathbb{E}(X)$ cuando el tamaño muestral n tiende a infinito.

5.4. Teorema central del límite

La suma de un gran número de variables aleatorias independientes e idénticamente distribuidas (aunque no sean normales) se distribuye aproximadamente como una normal.

Como consecuencia de esto el teorema afirma que la media muestral de una muestra aleatoria simple de una variable aleatoria X, también se aproxima a una normal de media $\mathbb{E}(X)$ y varianza $\mathrm{Var}(X)/n$, si el tamaño muestral es suficientemente grande.

5.5. Calidad de un estimador

Hablando de forma general, si queremos estimar un parámetro θ mediante un estimador T_n, cometeremos un error que puede ser más o menos grande. La calidad de un estimador puede cuantificarse mediante el error cuadrático medio, que se compone de dos sumandos:

$$\mathbb{E}(T_n - \theta)^2 = \mathrm{Var}(T_n) + (\mathbb{E}(T_n - \theta))^2 \tag{5.4}$$

$$\text{Error cuadrático medio} = \text{Varianza} + \text{Sesgo}^2. \tag{5.5}$$

Un buen criterio de calidad es que el estimador sea consistente, es decir que converja al valor del parámetro cuando $n \to \infty$.

Por otra parte también es bueno que el estimador sea insesgado (es decir con sesgo igual a cero).

Veamos como ejemplos la estimación de la media y la varianza poblacionales.

5.6. Estimación de la media

Si extraemos una muestra de tamaño n, mediante muestreo aleatorio simple, utilizaremos el estimadormedia muestral, $\overline{X}_n$ (hemos añadido el subíndice para indicar que depende de n) para estimar la media poblacional $\mathbb{E}(X) = \mu$. El estimador $\overline{X}_n$ es insesgado ya que $\mathbb{E}(\overline{X}) = \mathbb{E}(X) = \mu$ y además es consistente, porque $\lim_{n\to\infty} \overline{X}_n = \mathbb{E}(X) = \mu$. Si no realizamos el muestreo correctamente podrían dejar de cumplirse ambas propiedades.

5.7. Estimación de la varianza

Si queremos estimar la varianza poblacional σ^2, considerando la misma muestra que en el apartado anterior, utilizaremos el estimador varianza muestral:

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2. \tag{5.6}$$

Este estimador es insesgado $\mathbb{E}(S_n^2) = \sigma^2$ y consistente $\lim_{n\to\infty} S_n^2 = \sigma^2$.

5.8. Generación de numeros aleatorios uniformes

La generación de numeros aleatorios es fundamental para el muestreo y las técnicas de simulación, que tiene numerosas aplicaciones en física, biología, química, ingeniería, economía, ...

En los ordenadores personales es fácil simular la generación de números aleatorios, mediante mecanismos de generación de números pseudoaleatorios, que, sin ser aleatorios ya que se generan a partir de un algoritmo, lo aparentan.

Las subrutinas RAN de los ordenadores generan estos números.

Comando de R: `runif`

5.8.1. Generadores congruenciales lineales

La forma más sencilla de generar una secuencia de números pseudoaleatorios r_i es mediante un generador congruencial lineal. Escogemos r_0 la semilla o valor inicial y dos constantes enteras a, m y generamos los siguientes números de la secuencia mediante la siguiente operación:

$$r_{i+1} = (a \times r_i) \text{mod } m \tag{5.7}$$

Si elegimos a, m adecuados, obtenemos una secuencia finita de números que parecen seleccionados aleatoriamente entre 0 y $m-1$. La longitud de la secuencia viene determinada por a, m. La secuencia se repite con periodo menor que m.

Ejemplo: La secuencia completa con $m = 37, a = 5$ y valor inicial $r_0 = 1$ es

$$1, 5, 25, 14, 33, 17, 11, 18, 16, 6, 30, 2, 10, 13, 28, 29, 34, 22,$$
$$36, 32, 12, 23, 4, 20, 26, 19, 21, 31, 7, 35, 27, 24, 9, 8, 3, 15.$$

Depués de esta secuencia el ciclo vuelve a repetirse.

Otra fórmula similar a la que hemos visto y que también es utilizada, es:

$$r_{i+1} = (a \times r_i + b) \text{mod } m, \tag{5.8}$$

siendo b otra constante entera.

Si queremos números aleatorios en el intervalo real $x \in [0, 1]$ basta con dividir entre m: $x_{i+1} = r_{i+1}/m$

Las constantes m, a, b han de elegirse con cuidado. m suele elegirse como el mayor número entero del ordenador es decir $m \sim 2^{32}$.

Un buen algoritmo con ciclo $\sim 2 \times 10^{19}$ es , por ejemplo, el generador estandar mínimo con $a = 16807, m = 2^{31} - 1, b = 0$.

Sin embargo un algoritmo similar con $a = 65539, m = 2^{31}$ genera números con correlaciones muy fuertes.

El generador estándar mínimo tiene la ventaja de ser sencillo y rápido sin embargo pueden conseguirse números con correlaciones menores barajando los números obtenidos mediante una segunda secuencia aleatoria. El uso de este método aumenta el ciclo y reduce las correlaciones aunque disminuye la velocidad de cálculo.

Capítulo 6

Estimación por intervalo. Test de hipótesis

En la sección anterior veiamos como estimar un parámetro mediante estimación puntual, es decir calculábamos un valor que se aproximaba al valor real del parámetro. En esta sección veremos la estimación por intervalo mediante intervalos de confianza, y veremos como contrastar dos hipótesis. En primer lugar veremos los intervalos de predicción para diferenciarlos de los intervalos de confianza.

6.1. Intervalos de predicción

Ejemplo introductorio: Vamos a realizar un análisis de sangre a un paciente y queremos saber si su nivel de potasio en sangre será anómalo o no. ¿Es posible calcular un intervalo tal que el valor de potasio que se mida para el paciente esté contenido en el intervalo con un 95 % de probabilidad?

La respuesta es sí, calculando un intervalo de predicción. Para ello necesitamos disponer de una muestra (es decir de muchos valores de potasio en sangre para otras personas) lo más grande posible, o bien conocer la media y la varianza de la población.

Un **intervalo de predicción** al nivel $1-\alpha$ para una observación de una variable aleatoria X, es un intervalo (u, v) tal que la probabilidad de que la observación pertenezca al intervalo es $1-\alpha$. Normalmente dicho intervalo tiene su centro en la media de X.

Veamos cómo calcular un intervalo de predicción (u, v) cuando la variable aleatoria X tenga una distribución normal. Consideramos el caso en el que conocemos la media poblacional μ y la varianza poblacional σ^2. Por simetría se debe cumplir $\mathrm{P}(X \leq u) = \alpha/2$ y $\mathrm{P}(X \leq v) = 1 - \alpha/2$. De modo que:

$$\begin{aligned} u &= x(\alpha/2) = \mu + \sigma \times z(\alpha/2) \\ v &= x(1-\alpha/2) = \mu + \sigma \times z(1-\alpha/2). \end{aligned} \tag{6.1}$$

En R tendríamos:

$$\begin{aligned} u &= \texttt{qnorm}(\alpha/2, \mu, \sigma) \\ v &= \texttt{qnorm}(1-\alpha/2, \mu, \sigma). \end{aligned}$$

6.2. Intervalos de confianza

Ejemplo introductorio: Para el mismo análisis de nivel de potasio en sangre, queremos encontrar una respuesta a una pregunta diferente a la anterior ¿Es posible calcular un intervalo tal que la media poblacional de potasio en sangre esté contenida en el intervalo con un 95 % de probabilidad?

La respuesta es de nuevo afirmativa, pero habría que calcular un intervalo de confianza y no uno de predicción. Para ello necesitamos de nuevo una muestra.

Un **intervalo de confianza** al nivel $1-\alpha$ para un parámetro θ, es un intervalo aleatorio (U, V), obtenido a partir de una muestra de la población, tal que la probabilidad de que el parámetro esté contenido en el intervalo es $1-\alpha$.

Un intervalo de confianza es similar al intervalo de predicción pero con una diferencia conceptual muy importante:

- Un intervalo de predicción es fijo para cada nivel de predicción si μ y σ^2 son conocidas, y se trata de predecir entre qué límites fijos se va a encontrar una observación aleatoria de la variable de interés. Es decir que u y v son fijos y la variable X no lo es.

- En el intervalo de confianza (U, V) los límites U y V son variables aleatorias que se calculan a partir de una muestra, y el parámetro θ es el que es una cantidad fija.

Veamos cómo calcular intervalos de confianza con R a partir de los datos de una muestra aleatoria simple.

6.2.1. Intervalo de confianza para una media de una población normal

Para calcular el intervalo de confianza para una media poblacional, μ, de una variable aleatoria normal, al nivel de confianza $1-\alpha$, usaremos el siguiente comando:

$$\texttt{t.test}(\texttt{datos}, \texttt{conf.level} = 1 - \texttt{alpha}) \tag{6.2}$$

Comando de R: `t.test`

Para calcular el intervalo a mano se utiliza el hecho de que la siguiente cantidad tiene distribución conocida:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}, \tag{6.3}$$

siendo t_{n-1} la distribución *t de student* con $n-1$ grados de libertad. Dicha distribución es simétrica con forma similar a la normal y de hecho tiende a la normal para n grande. La función de distribución y los percentiles $t_{n-1}(p)$ de la distribución t de student están tabulados y por supuesto también se pueden calcular con R.

Comandos de R: `dt, pt, qt`

La siguiente igualdad es evidente aplicando la definición de percentil:

$$\mathrm{P}\left(t_{n-1}(\alpha/2) < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{n-1}(1-\alpha/2)\right) = 1-\alpha. \tag{6.4}$$

Operando en las desigualdades y sabiendo que por simetría $t_{n-1}(1-\alpha/2) = -t_{n-1}(\alpha/2)$ es fácil construir un intervalo de confianza para la media, μ:

$$\mathrm{P}\left(\overline{X} - t_{n-1}(1-\alpha/2)\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{n-1}(1-\alpha/2)\frac{S}{\sqrt{n}}\right) = 1-\alpha, \tag{6.5}$$

por lo que

$$\begin{aligned} U &= \overline{X} - t_{n-1}(1-\alpha/2)\frac{S}{\sqrt{n}} \\ V &= \overline{X} + t_{n-1}(1-\alpha/2)\frac{S}{\sqrt{n}}, \end{aligned} \tag{6.6}$$

siendo (U, V) el intervalo de confianza buscado.

6.3. Contrastes de hipótesis

Un contraste de hipótesis es un test para decidir entre dos hipótesis en base a unos datos de una muestra. Las hipótesis que se contrastan se denominan hipótesis nula, H_0, e hipótesis alternativa, H_1. En los contrastes de hipótesis, asumiremos cierta la hipótesis nula mientras no haya evidencia estadística para rechazarla, por tanto, como resultado del test de hipótesis *rechazamos* H_0 o bien *no rechazamos* H_0. Es importante matizar que se habla de no rechazar y no de *aceptar* la hipótesis nula, ya que no podemos aceptar definitivamente una hipótesis solo porque nuestros datos no muestren evidencia para rechazarla. Por otra parte, el hecho de rechazar H_0 no implica tener que aceptar H_1 como verdad definitiva; solo implica que los datos aportan suficiente evidencia para declarar a H_0 falsa.

Sea cual sea el resultado de nuestro contraste, siempre tendremos una probabilidad no nula de equivocarnos.

El contraste de hipótesis se realiza comparando dos cantidades que vamos a definir a continuación:

- Error de tipo I, α, nivel de significación o falso positivo: es la probabilidad de rechazar la hipótesis nula cuando es cierta.

- valor-p, p: es la probabilidad de que, siendo cierta H_0, se obtenga un resultado más desfavorable que el observado.

Habitualmente α es elegida por el científico que realiza el contraste de hipótesis al diseñar el experimento. Un valor estándar es $\alpha = 0.05$, pero es importante saber que hay casos en los que un 5 % de probabilidad de equivocarse al rechazar H_0 es demasiado alto.

A la hora de diseñar un experimento también hay que tener en cuenta el error de tipo II, β o falso negativo, que es la probabilidad de no rechazar la hipótesis nula cuando es falsa.

Los pasos a seguir en un test de hipótesis son los siguientes:

1. Diseño del experimento: elegir α y definir la hipótesis nula y la alternativa.
2. Calcular el valor-p con los datos de la muestra.
3. Comparar α con el valor-p de modo que:
   - $p < \alpha \Rightarrow$ Rechazamos la hipótesis nula.
   - $p \geq \alpha \Rightarrow$ No rechazamos la hipótesis nula.

Este esquema simplificado lo usaremos repetidas veces hasta final de curso y veremos cómo calcular el valor-p en distintos casos.

En esta sección comparamos siempre los datos con un valor de referencia.

6.3.1. Contrastes de hipótesis para una media de una población normal

La hipótesis nula sería en este caso:

$$\mathrm{H}_0 : \mu = \mu_0 \tag{6.7}$$

donde comparamos la media poblacional μ con un valor de referencia μ_0. Recordemos que estamos suponiendo que la variable aleatoria en cuestión tiene distribución normal.

Veamos primero el caso en que la hipótesis alternativa es bilateral, es decir,

$$\mathrm{H}_1 : \mu \neq \mu_0. \tag{6.8}$$

Para calcular el valor-p utilizamos la cantidad t:

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}. \tag{6.9}$$

El valor-p es la probabilidad de que, siendo cierta H_0, se obtenga un resultado más desfavorable que el observado y por tanto es

$$p = \mathrm{P}(|t| \geq |t_{n-1}| \ \mid \ \mu = \mu_0). \tag{6.10}$$

Dada una muestra podemos calcular $\overline{X}$, S y $\sqrt{n}$, y además suponemos $\mu = \mu_0$ por lo que una tabla de la función de distribución de la t de student podemos hallar fácilmente el valor-p.

Todavía más sencillo es utilizar el comando de R ya visto para los intervalos de confianza:

$$\texttt{t.test}(\texttt{datos}, \texttt{mu} = \mu_0), \tag{6.11}$$

que nos devuelve el valor-p.

Comando de R: `t.test`

Si la hipótesis alternativa es unilateral, es decir:

$$\mathrm{H}_1 : \mu > \mu_0 \quad \text{ó} \quad \mu < \mu_0, \tag{6.12}$$

el razonamiento es el mismo y tenemos,

$$p = \mathrm{P}(t > t_{n-1} \mid \mu = \mu_0) \quad \text{si} \quad \mathrm{H}_1 : \mu > \mu_0 \tag{6.13}$$

y con R:

$$\texttt{t.test}(\texttt{datos}, \texttt{mu} = \mu_0, \texttt{alternative} = \text{"}\texttt{greater}\text{"}), \tag{6.14}$$

o bien

$$p = \mathrm{P}(t < t_{n-1} \mid \mu = \mu_0) \quad \text{si} \quad \mathrm{H}_1 : \mu < \mu_0 \tag{6.15}$$

siendo el comando para R:

$$\texttt{t.test}(\texttt{datos}, \texttt{mu} = \mu_0, \texttt{alternative} = \text{"}\texttt{less}\text{"}). \tag{6.16}$$

6.3.2. Relación entre contrastes de hipótesis e intervalos de confianza

Cuando la hipótesis alternativa es bilateral, es fácil verificar que las dos siguientes afirmaciones son equivalentes:

- Rechazamos $\mathrm{H}_0 : \mu = \mu_0$ al nivel de significación α.
- El intervalo de confianza al nivel $1 - \alpha$ no contiene a μ_0.

Capítulo 7

Comparación de medias entre dos grupos

Hasta ahora hemos considerado la estimación puntual y por intervalo de la media de una población, así como las correspondientes pruebas de hipótesis. En esta sección veremos métodos para comparar dos poblaciones. Distinguiremos dos casos:

- **Muestras independientes**: Se parte de una población de la que se toman dos muestras aleatorias e independientes; una se usa como control, y a la otra se le administra el tratamiento de interés. Se trata de ver si el tratamiento tiene ó no efecto. Más concretamente, se trata de ver si el tratamiento altera la media poblacional de la variable de interés.

  Ejemplo: Queremos compromar si el medicamento A es efectivo para reducir la tensión areterial. Para ello tomamos dos muestras de 20 personas y administramos el medicamento A a una de las muestras, mientras que la otra toma un placebo y sirve de control. Para comprobar si es efectivo el medicamento, comparamos la tensión de ambas muestras.

- **Muestras apareadas**: Se toma una muestra de una población y cada individuo de la muestra actúa a la vez como control y como tratado, en períodos de tiempo simultáneos o diferentes.

  Ejemplo: Queremos comprobar si el medicamento A es efectivo para reducir la tensión areterial. Para ello tomamos una muestra de 20 personas y medimos su tensión arterial. Después les administramos el medicamento A y tras el tratamiento volvemos a medir la tensión arterial. En este caso para comprobar si es efectivo el medicamento comprobamos los valores medidos antes y después del tratamiento en las mismas personas.

Veamos ambos casos por separado.

7.1. Muestras independientes

Llamamos X a la variable control e Y a la variable tratamiento. Ambas se suponen normales, $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ y con tamaño muestral n_X y n_Y respectivamente.

Queremos saber si el tratamiento modifica la media poblacional, es decir si las medias poblacionales son distintas $\mu_X \neq \mu_Y$ o lo que es lo mismo si la diferencia de medias, δ es distinta de cero, $\delta = \mu_X - \mu_Y \neq 0$.

Para ello podemos:

- Construir un intervalo de confianza a un nivel dado $1-\alpha$ para la diferencia de medias poblacionales δ. Si el intervalo contiene el valor 0, los datos no demuestran que el tratamiento sea efectivo al nivel de confianza $1-\alpha$.
- Contrastar las hipótesis $\mathrm{H}_0: \ \delta = 0$ y $\mathrm{H}_1: \ \delta \neq 0$ a un nivel de significación α.

Los métodos a utilizar varían dependiendo de si el tratamiento altera la varianza o no; es decir dependiendo de si las varianzas de X e Y son homogéneas, $\sigma_X^2 = \sigma_Y^2$ o bien heterogéneas $\sigma_X^2 \neq \sigma_Y^2$.

Veremos al final de la sección cómo contrastar las hipótesis $\mathrm{H}_0: \ \sigma_X^2 = \sigma_Y^2$ y $\mathrm{H}_1: \ \sigma_X^2 \neq \sigma_Y^2$.

7.1.1. Intervalo de confianza, varianzas homogéneas

Para calcular el intervalo de confianza al nivel $1-\alpha$ con R en este caso, basta con utilizar:

$$\texttt{t.test(datos\_X, datos\_Y, var.equal = TRUE, conf.level = 1 - alpha)} \quad (7.1)$$

Para el cálculo manual se puede demostrar que S^2 es un estimador adecuado para la varianza común $\sigma^2 = \sigma_X^2 = \sigma_Y^2$,

$$S^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2} \quad (7.2)$$

es un estimador adecuado para la varianza común $\sigma^2 = \sigma_X^2 = \sigma_Y^2$ y que la siguiente cantidad,

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{S^2\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim t_{n_X+n_Y-2} \quad (7.3)$$

tiene distribución t de student con $n_X + n_Y - 2$ grados de libertad.

Siguiendo los pasos de la sección anterior, podemos construir un intervalo de confianza (U, V) para la diferencia de medias $\delta = \mu_X - \mu_Y$:

$$\begin{aligned} U &= \overline{X} - \overline{Y} - t_{n_X+n_Y-2}(1-\alpha/2) \times S\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \\ V &= \overline{X} - \overline{Y} + t_{n_X+n_Y-2}(1-\alpha/2) \times S\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}. \end{aligned} \quad (7.4)$$

7.1.2. Contraste de hipótesis, varianzas homogéneas

Cómo hemos comentado se trata de contrastar las siguientes hipótesis para la diferencia de medias poblacionales δ en dos poblaciones normales con varianzas homogéneas.

$$\begin{aligned} \mathrm{H}_0 : \delta &= 0 \\ \mathrm{H}_0 : \delta &\neq 0 \end{aligned} \tag{7.5}$$

También aquí podemos usar el mismo comando que usamos para calcular el intervalo de confianza:

$$\texttt{t.test}(\texttt{datos\_X}, \texttt{datos\_Y}, \texttt{var.equal} = \texttt{TRUE}) \tag{7.6}$$

que nos devuelve el valor-p.
Para calcular el valor-p a mano, con la ayuda de tablas usamos la cantidad

$$t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{S^2 \left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim t_{n_X + n_Y - 2}, \tag{7.7}$$

y calculamos

$$p = \mathrm{P}(|t| \geq |t_{n_X+n_Y-2}| \ \mid \ \delta = 0). \tag{7.8}$$

Veamos ahora el caso en el que $\sigma_X^2 \neq \sigma_Y^2$.

7.1.3. Intervalo de confianza, varianzas heterogéneas

Para calcular el intervalo de confianza con R, basta con una pequeña modificación:

$$\texttt{t.test}(\texttt{datos\_X}, \texttt{datos\_Y}, \texttt{var.equal} = \texttt{FALSE}, \texttt{conf.level} = 1 - \texttt{alpha}) \tag{7.9}$$

Sin embargo en el cálculo manual tenemos que usar la aproximación de Welch:

$$\frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)}} \sim t_g, \tag{7.10}$$

en donde g es el entero más próximo a:

$$\frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\frac{S_X^4}{n_X^2(n_X-1)} + \frac{S_Y^4}{n_Y^2(n_Y-1)}}. \tag{7.11}$$

De este modo es fácil ver que el intervalo de confianza (U, V) es:

$$
\begin{aligned}
U &= \overline{X} - \overline{Y} - t_g(1-\alpha/2)\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}} \\
V &= \overline{X} - \overline{Y} + t_g(1-\alpha/2)\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}.
\end{aligned} \tag{7.12}
$$

7.1.4. Contraste de hipótesis, varianzas heterogéneas

De nuevo las hipótesis son:

$$
\begin{aligned}
\mathrm{H}_0 : \delta &= 0 \\
\mathrm{H}_1 : \delta &\neq 0
\end{aligned} \tag{7.13}
$$

y el cálculo del valor-p se puede realizar con el comando de R:

$$
\texttt{t.test}(\texttt{datos\_X}, \texttt{datos\_Y}, \texttt{var.equal} = \texttt{FALSE}) \tag{7.14}
$$

Si queremos calcular el valor-p a mano tendremos que usar la cantidad:

$$
t = \frac{(\overline{X} - \overline{Y}) - (\mu_X - \mu_Y)}{\sqrt{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)}} \sim t_g, \tag{7.15}
$$

y calcular

$$
p = \mathrm{P}(|t| \geq |t_{n_X+n_Y-2}| \ \mid \ \delta = 0). \tag{7.16}
$$

7.1.5. Comparación de varianzas de dos poblaciones normales mediante muestras independientes

Veamos ahora como decidir si las varianzas son homogéneas o heterogéneas. Las hipótesis son:

$$
\begin{aligned}
\mathrm{H}_0 : \sigma_1^2 &= \sigma_2^2 \\
\mathrm{H}_1 : \sigma_1^2 &\neq \sigma_2^2
\end{aligned} \tag{7.17}
$$

y el cálculo del valor-p se puede realizar con el comando de R:

$$
\texttt{t.test}(\texttt{datos\_X}, \texttt{datos\_Y}) \tag{7.18}
$$

El cálculo está basado en el hecho de que el cociente de varianzas sigue una distribución conocida:

$$
F = \frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-2}, \tag{7.19}
$$

siendo F_{n_1-1,n_2-1} la distribución F de Snedecor.
por tanto el valor-p será:

$$
p = \mathrm{P}(|F| \geq |F_{n_1-1,n_2-1}| \ \mid \ \mu_D = 0). \tag{7.20}
$$

7.2. Muestras apareadas

Para distinguir del caso de muestras independientes, cambiaremos ligeramente la notación. Llamamos X_1 a la variable control y X_2 a la variable tratamiento. Ambas se suponen normales, $X_1 \sim N(\mu_{X_1}, \sigma^2_{X_1})$, $Y \sim N(\mu_{X_2}, \sigma^2_{X_2})$ y con el mismo tamaño muestral $n = n_{X_1} = n_{X_2}$.

De nuevo queremos saber si el tratamiento modifica la media poblacional, es decir si las medias poblacionales son distintas $\mu_{X_1} \neq \mu_{X_2}$.

Para ello, a continuación vamos a calcular un intervalo de confianza para la diferencia de medias $\mu_D = \mu_{X_1} - \mu_{X_2}$, y realizar un contraste de hipótesis.

7.2.1. Intervalo de confianza

El intervalo de confianza de la diferencia de medias para muestras apareadas al nivel $1 - \alpha$, se puede calcular con R mediante:

$$\texttt{t.test}(\texttt{datos\_X1}, \texttt{datos\_X2}, \texttt{paired} = \texttt{TRUE}, \texttt{conf.level} = \texttt{1} - \texttt{alpha}) \quad (7.21)$$

Los cálculos a mano se basan en la variable diferencia, $D = X_2 - X_1$ que se supone normal $D \sim N(\mu_D, \sigma^2_D)$ de modo que

$$\frac{\overline{D} - \mu_D}{S_D/\sqrt{n_D}} \sim t_{n-1} \quad (7.22)$$

y por tanto el intervalo de confianza (U, V) sería

$$\begin{aligned} U &= \overline{D} - t_{n-1}(1 - \alpha/2) \times \frac{S_D}{\sqrt{n}} \\ V &= \overline{D} + t_{n-1}(1 - \alpha/2) \times \frac{S_D}{\sqrt{n}}. \end{aligned} \quad (7.23)$$

7.2.2. Contraste de hipótesis

Las hipótesis a contrastar son:

$$\begin{aligned} \mathrm{H}_0 : \mu_D &= 0 \\ \mathrm{H}_1 : \mu_D &\neq 0. \end{aligned} \quad (7.24)$$

Para calcular el valor-p con R basta con utilizar el mismo comando que usamos para el intervalo de confianza:

$$\texttt{t.test}(\texttt{datos\_X1}, \texttt{datos\_X2}, \texttt{paired} = \texttt{TRUE}) \quad (7.25)$$

mientras que para el cálculo mediante tablas hay que usar la cantidad:

$$t = \frac{\overline{D} - \mu_D}{S_D/\sqrt{n_D}} \sim t_{n-1} \quad (7.26)$$

y calcular

$$p = \mathrm{P}(|t| \geq |t_{n\_1}| \ \mid \ \mu_D = 0). \quad (7.27)$$

Capítulo 8

Regresión y Correlación

En esta sección estudiamos el grado de asociación entre dos variables X e Y.

Distinguiremos dos casos:

- **Regresión**: La variable X es controlada y de hecho no se considera variable aleatoria, mientras que la variable aleatoria Y es la variable respuesta. Se trata de estudiar la variación de Y para distintos valores prefijados de X, para luego poder predecir Y, sin necesidad de medirla, para nuevos valores de X.

  Ejemplos:

  - ¿Cómo varía la tensión arterial de un paciente en función de la dosis de medicamento?
  - ¿Cómo se relacionan la duración del embarazo y el peso al nacer? ¿Podemos predecir mejor el peso al nacer sabiendo la duración del embarazo?

- **Correlación**: Ninguna de las dos variables es controlada, ni puede ser predicha a priori por el observador. Se trata de estudiar la variación de Y cuando varía X, y viceversa, y de medir el grado de asociación entre ambas variables aleatorias. No se trata sin embargo de especular si una de las variables es causa o no de la otra.

  Ejemplos:

  - ¿Cómo se relacionan la edad y el peso?
  - ¿Cómo se relacionan el volumen pulmonar y la potencia desarrollada por un ciclista?

En cualquiera de los dos casos es importante comenzar por representar los datos gráficamente en un diagrama de dispersión.

8.1. Regresión lineal simple

Para estudiar la regresión, utilizaremos un modelo lineal simple en el que las variables X e Y se tienen la siguiente relación:

$$Y = a + bX + \varepsilon \tag{8.1}$$

en donde llamamos *recta de regresión* a la recta $a + bX$, y ε se denomina residuo y es una variable aleatoria.

Comando de R: `lm(y ~ x)`

En el modelo lineal simple además de asumir una relación lineal entre ambas variables, haremos las siguientes suposiciones sobre ε:

- La media poblacional de los residuos es cero $\mathbb{E}(\varepsilon) = 0$.
- La varianza residual es independiente de X, es decir $\text{var}(\varepsilon) = \sigma^2$.
- Los residuos de cada unidad de muestreo son independientes, es decir para una muestra $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ los residuos $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ son independientes entre sí.

Si se cumplen las suposiciones anteriores, se pueden hallar estimadores de los coeficientes de la recta por mínimos cuadrados. Minimizando los cuadrados de los residuos $\sum_{i=1}^{n} \varepsilon_i = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$, obtenemos las siguientes expresiones:

$$\begin{aligned} \widehat{b} &= \frac{S_{XY}}{S_X^2} \\ \widehat{a} &= \overline{Y} - b\overline{X}, \end{aligned} \tag{8.2}$$

en donde S_{XY} es la covarianza muestral $S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$.

Comando de R: `coefficients(lm(y ~ x))`

Es importante resaltar que $\widehat{a}$ y $\widehat{b}$ son estimadores de los valores poblacionales y desconocidos a y b.

La varianza residual se puede expresar de la siguiente manera:

$$\sigma^2 = \sigma_Y^2 (1 - R^2) \tag{8.3}$$

en donde

$$R^2 = \left(\frac{\sigma_{XY}}{\sigma_Y \sigma_Y} \right)^2 \tag{8.4}$$

es el *coeficiente de determinación*, y σ_{XY} es la covarianza poblacional.

En la ecuación 8.3 vemos que la varianza residual, σ^2 es menor que la varianza de Y, σ_Y^2 en un factor $(1 - R^2)$. Por tanto cuanto mayor sea R^2, más se reduce la varianza residual por lo que tendremos menor incertidumbre en el valor de Y dado un valor X.

El coeficiente de determinación, nos transmite en qué medida el conocimiento de X determina el conocimiento de Y. De modo que si $R = 0$ el conocimiento de X no aporta nada al conocimiento de Y ya que $\sigma^2 = \sigma_Y^2$ y no hay

relación entre las variables. Por contra si $R = 1$ el conocimiento de X determina completamente el conocimiento de Y ya que $\sigma^2 = 0$ y todos los puntos $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ formarían una recta. Cuanto mayor sea R^2 más determinado estará el conocimiento de Y por el conocimiento de X. Obviamente como $\sigma^2 > 0 \Rightarrow R^2 < 1$.

El coeficiente de determinación se puede estimar mediante el coeficiente de determinación muestral, r^2:

$$r^2 = \left(\frac{S_{XY}}{S_X S_Y} \right)^2. \tag{8.5}$$

Hay que tener en cuenta que el modelo lineal no siempre es válido, ya que la relación entre las variables X e Y puede ser curvilínea o bien puede no cumplirse alguna de las suposiciones sobre los residuos.

8.2. Correlación lineal

En este caso la función de regresión es una recta, como en el caso de la regresión lineal. Sin embargo X no es una variable controlada, y hay que medirla junto con Y para cada individuo de una muestra.

En lugar del coeficiente de determinación se usa el coeficiente de correlación lineal, ρ:

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}, \tag{8.6}$$

que se puede estimar mediante el coeficiente de correlación muestral, r:

$$r = \frac{s_{XY}}{s_X s_Y}. \tag{8.7}$$

Comando de R: `cor(x,y)`

El coeficiente de correlación tiene las siguientes propiedades:

- $-1 \leq \rho \leq 1$. Cuando $\rho = 1$, todos los puntos de la población se encuentran sobre una recta de pendiente positiva, y cuando $\rho = -1$, sobre una recta de pendiente negativa. Cuando no hay relación entre las variables X e Y, entonces $\rho = 0$

- El coeficiente de correlación mide el grado de asociación entre las dos variables y su signo indica si la correlación es positiva (Y crece cuando X aumenta) o negativa (Y decrece cuando X aumenta).

- Análogamente al caso de regresión se verifica que $\sigma^2 = \sigma_Y^2(1 - \rho^2)$

- Si la regresión no es lineal, ρ pierde valor informativo, y no debe utilizarse. Por ejemplo, si los puntos $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ están sobre una curva, ésta explica el 100 % de la variación de dichos puntos, y sin embargo no tiene por qué valer 1 ó -1.

8.3. Errores comunes en la interpretación

- Como se indica más arriba, el coeficiente de correlación es una medida de relación lineal entre dos variables aleatorias; si la regresión no es lineal, el coeficiente de correlación no es una medida adecuada de relación.
- Si una de las variables es controlada, se puede hablar de regresión, pero no de correlación.
- No se debe establecer una relación causa-efecto entre dos variables solo porque estén correlacionadas.

  Ejemplos

  - El número de helados vendidos por mes y el número de muertes por ahogamiento en el mar por mes pueden estar correlacionadas ya que ambas variables aumentan en verano, pero no debemos establecer una relación causa-efecto entre las variables.
  - El número de bomberos que se encuentran apagando un fuego está correlacionado con la magnitud del fuego pero no por ello el número de bomberos es la causa de que la magnitud sea más grande.
  - El número piratas en el mar ha ido disminuyendo con los años al igual que la temperatura media en la tierra aumentaba por el efecto invernadero, sin embargo no podemos decir que la disminución del número de piratas se deba al aumento de la temperatura.

Capítulo 9

Bibliografía

Para este curso recomendamos:

- Probability and Statistics with R. M. D. Ugarte, A. F. Militino y A. T. Arnholt. CRC Press. 2008.
- Elements of statistical inference. David V. Huntsberger, Patrick Billingsley. Dubuque, Iowa : WCB, cop. 1989.
- Discrete Probability. Hugh Gordon. Springer. 1997.
- Introducción a la teoría de probabilidades y sus aplicaciones. Feller, W. Limusa Wiley, México. 1988.
- Understanding Statistics. Ott, L. y Mendenhall, W. PWS-KENT, Boston. 1990.