Bioestadística. Curso 2014-2015
Capítulo 1

Carmen Mª Cadarso, Mª del Carmen Carollo, Xosé Luis Otero, Beatriz Pateiro

Contents

1 Introducción a la Bioestadística

La Bioestadística es uno de los campos científicos que más se ha desarrollado en las últimas décadas. La creciente atención que está recibiendo en la literatura médica especializada pone de manifiesto la importancia de esta disciplina y el hecho, cada vez más patente, de que los profesionales médicos han dado a la investigación en Bioestadística un puesto dominante dentro de su formación.
La estadística permite analizar situaciones en las que los componentes aleatorios contribuyen de forma importante en la variabilidad de los datos obtenidos. La variabilidad es uno de los aspectos más esenciales de nuestra vida. La consiguiente incertidumbre que genera dicha variabilidad es importante y en muchos campos, como el de la medicina, es fundamental contar con métodos que nos permitan cuantificar dicha incertidumbre y minimizar su impacto en las decisiones que tomemos.

Se podría definir la Bioestadística como la ciencia que maneja mediante métodos estadísticos la incertidumbre en el campo de la medicina y la salud. En medicina, los componentes aleatorios se deben, entre otros aspectos, al desconocimiento o a la imposibilidad de medir algunos determinantes de los estados de salud y enfermedad, así como a la variabilidad en las respuestas de los pacientes. La fuente más común de incertidumbre en la medicina es la variabilidad natural de carácter biológico que existe entre individuos. Además, la variabilidad entre laboratorios, observadores, instrumentación, etc. también son fuentes de incertidumbre a tener en cuenta.
Por supuesto la Bioestadística no sólo se centra en medir incertidumbres sino que se preocupa también del control de su impacto. Por otra parte el profesional de la medicina no solo se forma para atender al paciente, sino que tiene además una responsabilidad y obligación social con la colectividad. Debe por lo tanto conocer los problemas de salud que afectan a su comunidad, los recursos con que cuenta y sus posibles soluciones, para lo cual necesita conocer la Estadística de Salud Pública y aplicarla en el proceso de planificación, ejecución y evaluación de acciones colectivas de salud.

La Bioestadística es la ciencia que maneja mediante métodos estadísticos la incertidumbre en el campo de la medicina y la salud

El campo de la estadística tiene que ver con la recopilación, presentación, análisis y uso de datos para tomar decisiones y resolver problemas. Cualquier persona, tanto en su carrera profesional como en la vida cotidiana recibe información en forma de datos a través de periódicos, de la televisión y de otros medios.

Ejemplo 1: Un cardiólogo, que investiga un nuevo fármaco para rebajar el colesterol, desea conocer el consumo de grasas en varones adultos mayores de 40 años. ¿Cómo debe proceder?

Población: Es el universo de individuos al cual se refiere el estudio que se pretende realizar.

Muestra: Subconjunto de la población cuyos valores de la variable que se pretende analizar son conocidos.

Variable: Rasgo o característica de los elementos de la población que se pretende analizar.

Una muestra aleatoria es un subconjunto de casos o individuos de una población

En el Ejemplo 1, la población objeto de estudio sería la formada por todos los varones adultos mayores de 40 años. La variable de interés es el consumo de grasas. El cardiólogo podría pensar en analizar a todos los individuos de la población. Sin embargo, esto resulta inviable (y así ocurre en muchas otras situaciones prácticas debido al coste, al tiempo que requiere,...) Entonces se conformará con extraer una muestra. La muestra proporciona información sobre el objeto de estudio. Lo habitual en nuestro contexto es que en el procedimiento de extracción intervenga el azar. Por ejemplo, el

cardiólogo seleccionaría al azar a 100 varones adultos mayores de 40 años y estudiaría el consumo de grasas de cada uno de ellos.

Ejemplo 2: Se quiere analizar el tiempo que dedican al estudio semanal los alumnos del Grado en Medicina de esta Universidad. Para ello se pregunta a 50 alumnos de esta titulación.
Población: Todos los estudiantes del Grado en Medicina de esta Universidadd.
Variable: Número de horas de estudio semanal.
Muestra: 50 alumnos encuestados.

Ejercicio 1: Se desea estimar el porcentaje de albúmina en el suero proteico de personas sanas. Para ello se analizan muestras de 40 personas, entre 2 y 40 años de edad.
¿Cuál es la población objeto de estudio?
¿Cuál es la variable de interés?
¿Cuál es la muestra con la que se realiza el estudio?

Clasificamos las tareas vinculadas a la Estadística en tres grandes disciplinas:

Estadística Descriptiva. Se ocupa de recoger, clasificar y resumir la información contenida en la muestra.

Cálculo de Probabilidades. Es una parte de la matemática teórica que estudia las leyes que rigen los mecanismos aleatorios.

Inferencia Estadística. Pretende extraer conclusiones para la población a partir del resultado observado en la muestra.

La Inferencia Estadística tiene un objetivo más ambicioso que el de la mera descripción de la muestra (Estadística Descriptiva). Dado que la muestra se obtiene mediante procedimientos aleatorios, el Cálculo de Probabilidades es una herramienta esencial de la Inferencia Estadística.

2 Tipos de variables

Variables cualitativas: No aparecen en forma numérica, sino como categorías o atributos. Por ejemplo el sexo, color de ojos, profesión, resultado de un tratamiento, etc. Las variables cualitativas se clasifican a su vez en:

Cualitativas nominales: Miden características que no toman valores numéricos. A estas características se les llama modalidades. Por ejemplo, en la variable sexo las modalidades son hombre y mujer.

Cualitativas ordinales: Miden características que no toman valores numéricos pero sí presentan entre sus posibles valores una relación de orden. Por ejemplo, si se desea examinar el resultado de un tratamiento, las modalidades podrían ser: en remisión, mejorado, estable, empeorado. El nivel de estudios puede tomar los valores: sin estudios, primaria, secundaria, etc.

Es importante clasificar correctamente las variables de interés ya que los procedimientos que veremos a continuación dependerán del tipo de variable con que trabajemos

Variables cuantitativas: Toman valores numéricos porque son frecuentemente el resultado de una medición. Por ejemplo, el peso (kg.) de una persona, la estatura (m.), número de llamadas diarias a un servicio de urgencias, temperatura (°C) corporal, etc. Las variables cuantitativas se clasifican a su vez en:

Cuantitativas discretas: Toman un número discreto de valores (en el conjunto de números naturales). Por ejemplo el número de hijos de una familia, número de cigarrillos fumados por día, etc.

Cuantitativas continuas: Toman valores numéricos dentro de un intervalo real. Por ejemplo, la altura, el peso, concentración de un elemento, tiempo transcurrido hasta que se inicia una reacción alérgica a una picadura de insecto, etc.

3 Distribución de frecuencias

La primera forma de recoger y resumir la información contenida en la muestra es efectuar un recuento del número de veces que se ha observado cada uno de los distintos valores que puede tomar la variable. A eso le llamamos frecuencia. Daremos definiciones precisas del concepto de frecuencia en sus distintas formas de presentación a través de un ejemplo práctico.

Ejemplo 3: En la última hora han acudido al servicio de urgencias de un hospital ocho pacientes, cuyos datos de ingreso se encuentran resumidos en la siguiente tabla. Clasifica las variables recogidas (sexo, peso, estatura, temperatura, número de visitas previas al servicio de urgencias y dolor).

Sexo	Peso (kg.)	Estatura (m.)	Temperatura (°C)	Visitas	Dolor
M	63	1.74	38	0	Leve
M	58	1.63	36.5	2	Intenso
H	84	1.86	37.2	0	Intenso
M	47	1.53	38.3	0	Moderado
M	70	1.75	37.1	1	Intenso
M	57	1.68	36.8	0	Leve
H	87	1.82	38.4	1	Leve
M	55	1.46	36.6	1	Intenso

En primer lugar, definimos el **tamaño muestral**, al que denotamos por n, como el número de individuos o de observaciones en la muestra. En el Ejemplo 3, el tamaño muestral es $n = 8$.

En la mayor parte de procedimientos estadísticos es necesario manejar conjuntos de observaciones numéricas. Para representar de forma concisa los cálculos, se ha desarrollado una notación matemática abreviada. Por ejemplo, para designar la adición se usa la letra griega Σ

3.1 Descripción de variables cualitativas.

Supongamos que los distintos valores que puede tomar la variable son: $c_1, c_2, \ldots, c_m$.

Frecuencia absoluta: Se denota por n_i y representa el número de veces que ocurre el resultado c_i.

Frecuencia relativa: Se denota por f_i y representa la proporción de datos en cada una de las clases,

$$f_i = \frac{n_i}{n}.$$

La frecuencia relativa es igual a la frecuencia absoluta dividida por el tamaño muestral.

Frecuencia absoluta acumulada. Es el número de veces que se ha observado el resultado c_i o valores anteriores. La denotamos por $N_i = \sum_{c_j \leq c_i} n_j$.

Frecuencia relativa acumulada. Es la frecuencia absoluta acumulada dividida por el tamaño muestral. La denotamos por

$$F_i = \frac{N_i}{n} = \sum_{c_j \leq c_i} f_j.$$

Debemos observar que las frecuencias acumuladas sólo tienen sentido cuando es posible establecer una relación de orden entre los valores de la variable, esto es, cuando la variable es ordinal.
Las frecuencias se pueden escribir ordenadamente mediante una **tabla de frecuencias**, que adopta esta forma:

Para comprender y resumir un conjunto de datos es útil presentarlos en una tabla en la que aparezcan los valores posibles de la variable y el número de veces que cada valor se repite

c_i	n_i	f_i	N_i	F_i
c_1	n_1	f_1	N_1	F_1
c_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_m	n_m	f_m	N_m	F_m

Propiedades:

Frecuencias absolutas	$0 \leq n_i \leq n$	$\sum_{i=1}^{m} n_i = n$
Frecuencias relativas	$0 \leq f_i \leq 1$	$\sum_{i=1}^{m} f_i = 1$
Frecuencias absolutas acumuladas	$0 \leq N_i \leq n$	$N_m = n$
Frecuencias relativas acumuladas	$0 \leq F_i \leq 1$	$F_m = 1$

Claramente, la suma de las frecuencias absolutas es el número total de datos, n, y la suma de las frecuencias relativas es 1. Observa que el último valor de la distribución de frecuencias absolutas acumuladas coincide con el número de observaciones. Análogamente, el último valor de la distribución de frecuencias relativas acumuladas es uno.
La distribución de frecuencias acumuladas permite conocer la proporción de valores por debajo de cierto valor de la variable, o entre dos valores especificados, o por encima de cierta cantidad.

Como ejemplo, vamos a construir la tabla de frecuencias para la variable Dolor del Ejemplo 3. La variable Dolor es una variable cualitativa ordinal que presenta tres modalidades: leve, moderado e intenso. Tendríamos así la tabla de frecuencias:

c_i	n_i	f_i	N_i	F_i
Leve	3	0.375	3	0.375
Moderado	1	0.125	4	0.5
Intenso	4	0.5	8	1
	$n = 8$	$\sum f_i = 1$		

Interpreta los resultados obtenidos y comprueba que se verifican las propiedades de las frecuencias. ¿Qué porcentaje de pacientes que acudieron al servicio de urgencias sufren dolor intenso? ¿Cuántos pacientes acudieron al servicio de urgencias con dolor leve o moderado?

Ejercicio 2: Construye la tabla de frecuencias para el resto de variables cualitativas que aparecen en el Ejemplo 3.

Ejercicio 3: Con el objetivo de estudiar la influencia de la dureza del agua en ciertos trastornos gastrointestinales simples, un laboratorio determinó la dureza del agua de 10 muestras obteniendo los siguientes resultados:

Muestra	Dureza
1	Agua blanda
2	Agua blanda
3	Agua dura
4	Agua muy dura
5	Agua muy dura
6	Agua extremadamente dura
7	Agua blanda
8	Agua blanda
9	Agua dura
10	Agua muy dura

Construye la tabla de frecuencias relativas para la variable "Dureza del agua".

3.2 Descripción de variables cuantitativas.

3.2.1 Descripción de variables cuantitativas discretas.

Una variable cuantitativa discreta es una variable que toma un número finito o infinito numerable de valores posibles. La forma de resumir los datos observados de una variable cuantitativa discreta es similar a la forma de resumir datos de una variable cualitativa. Veremos como construir la tabla de frecuencias de una variable discreta a través de un ejemplo.

Considera ahora la variable Visitas del Ejemplo 3. Fíjate que la variable Visitas es discreta ya que puede tomar los valores 0,1,2,... (un número infinito numerable de valores). A continuación construimos la tabla de frecuencias:

Visitas	n_i	f_i	N_i	F_i
0	4	0.5	4	0.5
1	3	0.375	7	0.875
2	1	0.125	8	1

Fíjate en la información que nos ofrece la tabla de frecuencias. Observamos por ejemplo que el 87.5% de los pacientes registrados no habían acudido con anterioridad en más de una ocasión al servicio de urgencias. También observamos que sólo 1 paciente había acudido anteriormente en 2 ocasiones al servicio de urgencias (lo que representa un 12.5% del total de pacientes registrados).

Ejercicio 4: Consideremos una muestra de 200 familias en las que contamos el número de hijos. Supongamos que se han observado 50 familias sin hijos, 80 familias con un hijo, 40 familias con dos hijos, 20 familias con tres hijos y 10 familias con cuatro hijos. Construye la tabla de frecuencias correspondiente.

3.2.2 Descripción de variables cuantitativas continuas.

Para construir tablas de frecuencias de variables cuantitativas continuas es habitual agrupar los valores que puede tomar la variable en intervalos. De este modo contamos el número de veces que la variable cae en cada intervalo. A cada uno de estos intervalos le llamamos **intervalo de clase** y a su punto medio **marca de clase**. Por tanto, para la definición de las frecuencias y la construcción de la tabla de frecuencias sustituiremos los valores c_i por los intervalos de clase y las marcas de clase. Algunas consideraciones a tener en cuenta:

- *Número de intervalos a considerar:* Para adoptar esta decisión tendremos en cuenta:
  1. Cuantos menos intervalos tomemos, menos información se recoge.
  2. Cuantos más intervalos tomemos, más difícil es manejar las frecuencias.

  Aunque no hay unanimidad al respecto, un criterio bastante extendido consiste en tomar como número de intervalos el entero más próximo a $\sqrt{n}$.
- *Amplitud de cada intervalo:* Lo más común es tomar todos los intervalos de igual longitud.
- *Posición de los intervalos:* Los intervalos deben situarse allí donde se encuentran las observaciones y de forma contigua. Es aconsejable que los restos de intervalos en los extremos derecho e izquierdo del conjunto de observaciones sean similares.

Si una variable cuantitativa discreta toma muchos valores distintos puede ser conveniente una agrupación por intervalos como en el caso continuo

A continuación veremos un ejemplo práctico de cómo se construyen los intervalos y la tabla de frecuencias para variables cuantitativas continuas. En la resolución de los ejemplos será útil ordenar la muestra de observaciones y después calcular el **recorrido o rango**, que definimos como la diferencia entre el dato más grande y el más pequeño de la muestra. El recorrido se usa para obtener la amplitud de los intervalos. La ordenación facilita mucho también el recuento de las frecuencias en cada intervalo.

Considera la variable Peso del Ejemplo 3. En primer lugar vamos a ordenar los datos de la muestra de menor a mayor para que sea más sencillo el recuento de frecuencias.

- Muestra ordenada: $47, 55, 57, 58, 63, 70, 84, 87$.
- Recorrido$= 87 - 47 = 40$.
- Número de intervalos $\approx \sqrt{8} = 2.82 \approx 3$.

  Como $40/3 = 13.3$, podemos tomar 3 intervalos de amplitud 14 y así conseguimos contener toda la muestra y los extremos de los intervalos resultan manejables.

Intervalo de clase	Marca de clase				
$[L_i, L_{i+1})$	c_i	n_i	f_i	N_i	F_i
$[46, 60)$	53	4	0.5	4	0.5
$[60, 74)$	67	2	0.25	6	0.75
$[74, 88)$	81	2	0.25	8	1

Observamos, por ejemplo, que hay 2 pacientes con peso comprendido en el intervalo $[74, 88)$ y que el 75% de los pacientes atendidos pesan menos de 74 kg.

Ejercicio 5: En un estudio sobre trastornos de sueño se analizó el comportamiento de 10 varones cuyas edades se muestran a continuación:

$$52, 47, 51, 28, 64, 31, 22, 53, 29, 23$$

Calcula una tabla de frecuencias para la variable Edad organizando los datos en tres intervalos [20,35), [35,50), [50,65).

4 Representaciones gráficas

La representación gráfica de la información contenida en una tabla estadística es una manera de obtener una información visual clara y evidente de los valores asignados a la variable estadística. Existen multitud de gráficos adecuados a cada situación. Unos se emplean con variables cualitativas y otros con variables cuantitativas.

4.1 Representaciones gráficas de variables cualitativas

Diagrama de barras: Representaremos las frecuencias absolutas o relativas de variables cualitativas mediante un diagrama de barras. Para ello, situamos las modalidades de la variable en el eje de abscisas, respetando su orden si lo hubiera, y dibujamos barras verticales sobre ellas. Las alturas de las barras representan frecuencias absolutas, relativas o porcentajes.

En la Figura 1 se muestra el diagrama de barras de frecuencias absolutas para la variable Dolor del Ejemplo 3.

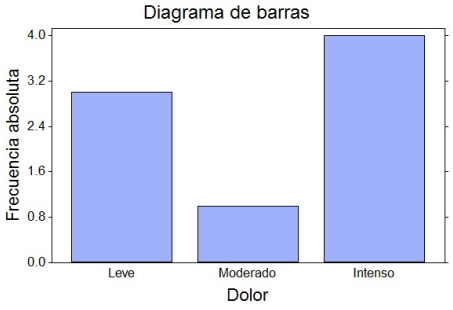


Figure 1: Diagrama de barras de frecuencias absolutas para la variable Dolor

Diagrama de sectores: Se obtiene dividiendo un círculo en tantos sectores como modalidades tome la variable. La amplitud de cada sector debe ser proporcional a la frecuencia del valor correspondiente.

En la Figura 2 se muestra el diagrama de sectores de la variable Dolor del Ejemplo 3.

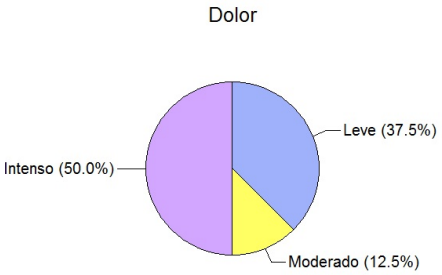


Figure 2: Diagrama de sectores para la variable Dolor

Ejercicio 6: Un laboratorio está desarrollando unas nuevas tiras de orina para detectar los niveles de acetona. Se realizan 50 pruebas de acetona en pacientes y se obtiene en 15 ocasiones el color naranja, 25 veces se obtiene el color amarillo y en 10 ocasiones resulta el color verde. Construye la tabla de frecuencias y representa las gráficas adecuadas para la variable Color de reacción.

4.2 Representaciones gráficas de variables cuantitativas

4.2.1 Representaciones gráficas de variables cuantitativas discretas

Representaremos los datos de variables cuantitativas discretas mediante diagramas de barras, al igual que hicimos con variables cualitativas. En la Figura 3 se muestra el diagrama de barras de frecuencias absolutas para la variable Visitas del Ejemplo 3.

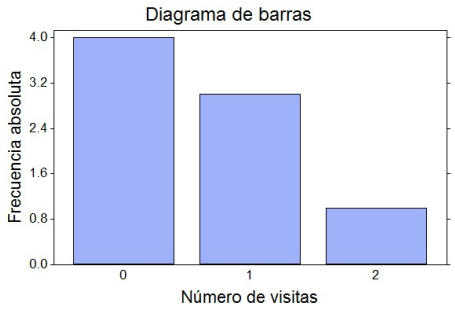


Figure 3: Diagrama de barras de frecuencias absolutas para la variable Dolor

4.2.2 Representaciones gráficas de variables cuantitativas continuas

Las frecuencias de una variable cuantitativa continua también se pueden representar gráficamente. Sin embargo, el diagrama de barras no parece adecuado para este caso, pues lo que debemos representar son frecuencias de intervalos contiguos.

Histograma: Es un gráfico para la distribución de una variable cuantitativa continua que representa frecuencias mediante áreas. El histograma se construye colocando en el eje de abscisas los intervalos de clase, como trozos de la recta real, y levantando sobre ellos rectángulos con **área proporcional a la frecuencia.**

Dibujamos en la Figura 4 el histograma correspondiente a la distribución de frecuencias obtenida para la variable Peso del Ejemplo 3.

- A diferencia del diagrama de barras, los rectángulos se dibujan contiguos.
- El aspecto del histograma cambia variando el número de clases y el punto donde empieza la primera clase.
- Cuanto mayor es el área de una clase, mayor es su frecuencia.
- El histograma ayuda a describir cómo es la distribución de la variable, si es simétrica (con un eje de simetría), bimodal (con dos máximos),...etc.

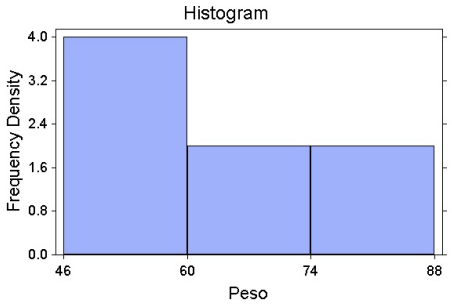


Figure 4: Histograma.

Formalmente, la altura de los rectángulos de un histograma debería representar la **densidad de frecuencia**, que es el cociente $f_i/(L_{i+1} - L_i)$. Así, el área total encerrada por el histograma sería igual a uno. Sin embargo, la mayoría de programas informáticos de estadística representan el histograma mediante rectángulos de altura igual a la frecuencia absoluta o relativa de cada intervalo como se muestra en la Figura 4

5 Medidas características: Medidas de posición y de dispersión

El objetivo fundamental de la estadística es extraer conclusiones sobre una población basándonos en la información obtenida en la muestra. Hasta ahora hemos visto como resumir esa información mediante tablas de frecuencias y representaciones gráficas que nos ayudan a visualizar la distribución de los datos. Estudiaremos ahora como calcular medidas que nos den una descripción muy resumida sobre alguna propiedad concreta del conjunto de datos. Por **medida** entendemos, pues, un número que se calcula sobre la muestra y que refleja cierta cualidad de la misma. El cálculo de estas medidas requiere efectuar operaciones con los valores que toma la variable. Por este motivo, a partir de ahora tratamos sólo con variables cuantitativas.

5.1 Medidas de posición

En esta sección estudiamos medidas que nos indican la posición que ocupa la muestra. La posición central son el objetivo de la media, la mediana y la moda. El estudio de posiciones no centrales se hará con los cuantiles.

Media aritmética: Sean $x_1, x_2, \ldots, x_n$ un conjunto de n observaciones de la variable X. Se define la media aritmética (o simplemente media) de estos valores como:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Ejemplo 4: Calculamos el peso medio de los pacientes de urgencias del Ejemplo 3.

$$\bar{x} = \frac{63 + 58 + 84 + \ldots + 55}{8} = 65.125 \text{ kg}.$$

Observamos que el peso medio es 65.125 kg. Fíjate que la unidad de medida de la media es la misma que la de los datos originales.

La media aritmética tiene interesantes propiedades:

Propiedades:

1. $\min(x_i) \leq \bar{x} \leq \max(x_i)$ y tiene las mismas unidades que los datos originales.
2. Es el centro de gravedad de los datos:
   $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$,
   $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \min_{a \in \mathbb{R}} \sum_{i=1}^{n} (x_i - a)^2$.
3. Si $y_i = a + bx_i \Rightarrow \bar{y} = a + b\bar{x}$.

Ejemplo 5: Se ha detectado un error en la báscula con la que se han pesado los pacientes del Ejemplo 3. La báscula estaba mal equilibrada y añadía a todos los pacientes 5 kg. a su peso real ¿Cuál es entonces el peso medio correcto de los pacientes?
Si X representa el peso que hemos medido con error, $Y = X - 5$ representaría el peso real de los pacientes. Para calcular el peso medio correcto no nos haría falta calcular de nuevo todos los pesos, ya que por las propiedades de la media (propiedad 3) sabemos que:

$$\bar{y} = \bar{x} - 5 = 60.125 \text{ kg}.$$

Efectivamente, los pesos reales serían 58, 53, 79, 42, 65, 52, 82, 50. Por lo tanto la media de los pesos reales sería:

$$\bar{y} = \frac{58 + 53 + 79 + \ldots + 50}{8} = 60.125 \text{ kg}.$$

Mediana: Una vez ordenados los datos de menor a mayor, se define la mediana como el valor de la variable que deja a su izquierda el mismo número de valores que a su derecha. Si hay un número impar de datos, la mediana es el valor central. Si hay un número par de datos, la mediana es la media de los dos valores centrales.

Ejemplo 6: Calculamos el peso mediano de los pacientes de urgencias del Ejemplo 3. En primer lugar ordenamos los datos de menor a mayor:

$$47, 55, 57, 58, 63, 70, 84, 87$$

Tenemos un número par de datos ($n = 8$) y por lo tanto la mediana será:

$$\text{Me} = \frac{58 + 63}{2} = 60.5 \text{ kg}.$$

Observa que la media y la mediana tendrán valores similares, salvo cuando haya valores atípicos o cuando la distribución sea muy asimétrica. La mediana es la medida de posición central más robusta (es decir, más insensible a datos anómalos).

Moda: Es el valor de la variable que se presenta con mayor frecuencia. A diferencia de las otras medidas, la moda también se puede calcular para variables cualitativas. Pero, al mismo tiempo, al estar tan vinculada a la frecuencia, no se puede calcular para variables continuas sin agrupación por intervalos de clase. Al intervalo con mayor frecuencia le llamamos **clase modal.**

Puede ocurrir que haya una única moda, en cuyo caso hablamos de distribución de frecuencias **unimodal.** Si hay más de una moda, diremos que la distribución es **multimodal.**

Ejemplo 7: Calculamos la moda de la variable Visitas del Ejemplo 3. Fíjate en la tabla de frecuencias y observa que la mayoría de los pacientes no habían acudido con anterioridad al servicio de urgencias. Por lo tanto,

$$\text{Moda} = 0.$$

Para la variable Peso del Ejemplo 3 nos fijamos también en la tabla de frecuencias.

$$\text{Intervalo modal} = [46, 60).$$

Cuantiles: Hemos visto que la mediana divide a los datos en dos partes iguales. Pero también tiene interés estudiar otros parámetros, llamados cuantiles, que dividen los datos de la distribución en partes iguales, es decir en intervalos que comprenden el mismo número de valores. En general, sea $p \in (0, 1)$. Se define el cuantil p como el número que deja a su izquierda una frecuencia relativa p. Observa que la mediana es el cuantil 0.5. Existen distintos métodos para calcular los cuantiles. Una posible forma de calcular el cuantil p consistiría en ordenar la muestra y tomar como cuantil el menor dato de la muestra (primero de la muestra ordenada) cuya frecuencia relativa acumulada es mayor que p.

Recuerda ordenar las observaciones de menor a mayor para calcular la mediana y el resto de cuantiles

Algunos órdenes de los cuantiles tienen nombres específicos. Así los **cuartiles** son los cuantiles de orden (0.25, 0.5, 0.75) y se representan por Q_1, Q_2, Q_3. Los cuartiles dividen la distribución en cuatro partes. Los **deciles** son los cuantiles de orden (0.1, 0.2,..., 0.9). Los **percentiles** son los cuantiles de orden $j/100$ donde j=1,2,...,99.

5.2 Medidas de dispersión

Las medidas de dispersión se utilizan para describir la variabilidad o esparcimiento de los datos de la muestra respecto a la posición central.

Recorrido o rango: $R = \max x_i - \min x_i$.

Recorrido intercuartílico: se define como la diferencia entre el cuartil tercero y el cuartil primero, es decir, $RI = Q_3 - Q_1$.

Varianza: Si hemos empleado la media como medida de posición, parece razonable tomar como medida de dispersión algún criterio de discrepancia de los puntos respecto a la media. Según hemos visto, la simple diferencia de los puntos y la media, al ponderarla, da cero. Por tanto, elevamos esas diferencias al cuadrado para que no se cancelen los sumandos positivos con los negativos. El resultado es la varianza, cuya definición se da a continuación.

Sean $x_1, x_2, \ldots, x_n$ un conjunto de n observaciones de la variable X. Se define la varianza muestral como:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Una medida de variabilidad más lógica sería $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Sin embargo, por motivos teóricos es preferible calcular la varianza muestral s^2 tal y como la hemos definido (con denominador $n-1$). Además así definido s^2 coincide con el valor que calculan la mayor parte de programas informáticos.

Propiedades:

1. $s^2_{a+X} = s^2_X$. La varianza no se ve afectada por cambios de localización.
2. $s^2_{b\cdot X} = b^2 \cdot s^2_X$. La varianza se mide en el cuadrado de la escala de la variable

---

Que una medida de dispersión no se vea afectada por cambios de localización, como ocurre con la varianza (propiedad 1), es una condición casi indispensable para admitirla como tal medida de dispersión. La dispersión de un conjunto de datos no se ve alterada por una mera traslación de los mismos.

Ejemplo 8: Calculamos la varianza del peso de los pacientes de urgencias del Ejemplo 3. Recuerda que $\bar{x} = 65.125$ kg.

$$s^2 = \frac{(63 - 65.125)^2 + (58 - 65.125)^2 + \ldots + (55 - 65.125)^2}{7} = 201.55 \text{ kg}^2.$$

Desviación típica: La propiedad 2 de la varianza nos da pie a calcular la raíz cuadrada de la varianza, obteniendo así una medida de dispersión que se expresa en la mismas unidades de la variable. Esta medida es la desviación típica, que en coherencia denotamos por s.

Ejemplo 9: Calculamos la desviación típica del peso de los pacientes de urgencias del Ejemplo 3.

$$s = \sqrt{201.55} = 14.197 \text{ kg}.$$

Coeficiente de variación: Si queremos una medida de dispersión que no dependa de la escala y que, por tanto, permita una comparación de las dispersiones relativas de varias muestras, podemos

utilizar el coeficiente de variación, que se define así:

$$CV = \frac{s}{\bar{x}}.$$

Por supuesto, para que se pueda definir esta medida es preciso que la media no sea cero. Es más, el coeficiente de variación sólo tiene sentido para variables que sólo tomen valores positivos y que no sean susceptibles de cambios de localización.

Ejemplo 10: Calculamos el coeficiente de variación del peso de los pacientes del Ejemplo 3.

$$CV = \frac{s}{\bar{x}} = \frac{14.197}{65.125} = 0.218$$

Ejercicio 7: Un estudio tiene como objetivo determinar la concentración de pH en muestras de saliva humana. Para ello se recogieron datos de 10 personas obteniéndose los siguientes resultados.

6.59 7.37 7.15 7.08 5.75 5.83 7.12 7.23 7.13 5.60

Calcular la media, mediana, desviación típica, cuartiles y rango intercuartílico.

Ejercicio 8: Realiza un análisis descriptivo completo de cada una de las variables del Ejemplo 3.

5.3 Medidas de forma

Las medidas de forma tratan de medir el grado de simetría y apuntamiento en los datos.

Coeficiente de asimetría de Fisher: Se define como

$$As_F = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{ns^3}.$$

La interpretación de este coeficiente es la siguiente: Si su valor es prácticamente cero se dice que los datos son simétricos. Si toma valores significativamente mayores que cero diremos que los datos son asimétricos a la derecha y si toma valores significativamente menores que cero diremos que son asimétricos a la izquierda.

Coeficiente de apuntamiento de Fisher: Mide el grado de concentración de una variable respecto a su medida de centralización usual (media). Se define como:

$$K_F = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4}.$$

Puesto que en Estadística el modelo de distribución habitual de referencia es el gausiano o normal y este presenta teóricamente un coeficiente de apuntamiento de 3, se suele tomar este valor como referencia. Así, si este coeficiente es menor que 3 diremos que los datos presentan una forma platicúrtica, si es mayor que 3 diremos que son leptocúrticos y si son aproximadamente 3 diremos que son mesocúrticos.

5.4 El diagrama de caja o Boxplot

La información obtenida a partir de las medidas de centralización, dispersión y forma se puede usar para realizar diagramas de caja (boxplots) que visualmente nos dan información sobre como están distribuidos los datos. El diagrama de caja consta de:

- una caja central que está delimitada por la posición de los cuartiles Q_1 y Q_3.
- Dentro de esa caja se dibuja la línea que representa la mediana (cuartil Q_2).
- De los extremos de la caja salen unas líneas (denominadas bigotes) que se extienden hasta los puntos $LI = \min\{x_i,\ \text{tal que } x_i \geq Q_1 - 1.5RI\}$ y $LS = \max\{x_i,\ \text{tal que } x_i \leq Q_3 + 1.5RI\}$. Es decir, LI es la menor de las observaciones que es mayor o igual que $Q_1 - 1.5RI$ y LS es la mayor de las observaciones que es menor o igual que $Q_3 + 1.5RI$. Estos límites representarían el rango razonable hasta el cual se pueden encontrar datos.
- Los datos que caen fuera de los bigotes se representan individualmente mediante "$*$" (datos atípicos moderados) y "o" (datos atípicos extremos).

La Figura 5 muestra los diagramas de caja para datos de Estatura agrupados por Sexo. Fíjate que en ambos sexos hay datos atípicos moderados (personas cuyas estaturas están fuera del rango "razonable" de valores determinado por el conjunto de observaciones de cada sexo).

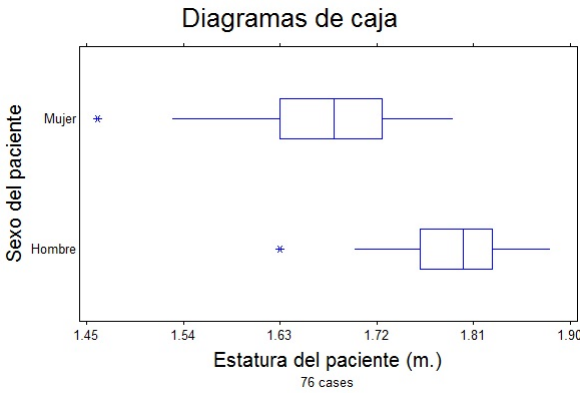


Figure 5: Diagramas de caja para la variable Estatura agrupada por Sexo.

Bioestadística. Curso 2014-2015
Capítulo 2

Carmen Mª Cadarso, Mª del Carmen Carollo, Xosé Luis Otero, Beatriz Pateiro

Contents

1 Introducción histórica

El objetivo de la Estadística es utilizar los datos para inferir sobre las características de una población a la que no podemos acceder de manera completa. En el tema anterior, hemos visto como realizar un análisis descriptivo de una muestra de datos. La **Probabilidad** es la disciplina científica que proporciona y estudia modelos para fenómenos aleatorios en los que interviene el azar y sirve de soporte teórico para la Estadística.

La Teoría de la Probabilidad surgió de los estudios realizados sobre los juegos de azar, que se remontan miles de años atrás.

Como primeros trabajos con cierto formalismo en Teoría de la Probabilidad cabe destacar los realizados por Cardano y Galilei (siglo XVI), aunque las bases de esta teoría fueron desarrolladas por Pascal y Fermat en el siglo XVII. De ahí en adelante grandes científicos han contribuido al desarrollo de la Probabilidad, como Bernouilli, Bayes, Euler, Gauss,... en los siglos XVIII y XIX. Será a finales del siglo XIX y principios del XX cuando la Probabilidad adquiera una mayor formalización matemática, debida en gran medida a la llamada Escuela de San Petesburgo en la que cabe destacar los estudios de Tchebychev, Markov y Liapunov.

2 Conceptos básicos

2.1 Experimento aleatorio

Cuando de un experimento podemos averiguar de alguna forma cuál va a ser su resultado *antes* de que se realice, decimos que el experimento es determinístico. Así, podemos considerar que las horas de salida del Sol, o la pleamar o bajamar son determinísticas, pues podemos leerlas en el periódico antes de que se produzcan. Por el contrario, no podemos encontrar en ningún medio el número premiado en la Lotería de Navidad *antes del sorteo*.

Nosotros queremos estudiar experimentos que no son determinísticos, pero no estamos interesados en todos ellos. Por ejemplo, no podremos estudiar un experimento del que, por no saber, ni siquiera sabemos por anticipado los resultados que puede dar. No realizaremos tareas de adivinación. Por ello definiremos experimento aleatorio como aquel que verifique ciertas condiciones que nos permitan un estudio riguroso del mismo.

Llamamos **experimento aleatorio** al que satisface los siguientes requisitos:

- Todos sus posibles resultados son conocidos de antemano.
- El resultado particular de cada realización del experimento es imprevisible.
- El experimento se puede repetir indefinidamente en condiciones idénticas.

Ejemplo 1: Ejemplos de experimentos aleatorios son:

- $\mathcal{E}_1$ =*Lanzar una moneda al aire*,
- $\mathcal{E}_2$ =*Lanzar dos veces una moneda*,
- $\mathcal{E}_3$ =*Lanzar dos monedas a la vez*,
- $\mathcal{E}_4$ =*Medir la temperatura corporal de un paciente*.

2.2 Espacio muestral. Sucesos.

Espacio muestral: Es el conjunto formado por todos los resultados posibles del experimento aleatorio. Lo denotamos por Ω.

Ejemplo 2: Si lanzamos una moneda, $\Omega = \{c, +\}$.

Suceso elemental: Es un suceso unitario. Está constituido por un solo resultado del experimento aleatorio.

Ejemplo 3: Si lanzamos un dado, $\Omega = \{1, 2, 3, 4, 5, 6\}$, los sucesos elementales son:

- A = El resultado es un 1 = $\{1\}$,
- B = El resultado es un 2 = $\{2\}$,
- ...,
- F = El resultado es un 6 = $\{6\}$.

Suceso: Cualquier subconjunto del espacio muestral.

Ejemplo 4: Si lanzamos un dado, $\Omega = \{1, 2, 3, 4, 5, 6\}$, podemos considerar muchos sucesos:

- A = El resultado es par= $\{2, 4, 6\}$,
- B = El resultado es menor que 3= $\{1, 2\}$,
- ...

Decimos que **ha ocurrido** un suceso cuando se ha obtenido alguno de los resultados que lo forman. El objetivo de la Teoría de la Probabilidad es estudiar con rigor los sucesos, asignarles probabilidades y efectuar cálculos sobre dichas probabilidades. Observamos que los sucesos no son otra cosa que conjuntos y por tanto, serán tratados desde la Teoría de Conjuntos. Recordamos las operaciones básicas y las dotamos de interpretación para el caso de sucesos.

Suceso seguro: Es el que siempre ocurre y, por tanto, es el espacio muestral, Ω.

Suceso imposible: Es el que nunca ocurre y, por tanto, es el vacío, $\emptyset$.

Unión.: Ocurre $A \cup B$ si ocurre al menos uno de los sucesos A o B.

Intersección: Ocurre $A \cap B$ si ocurren los dos sucesos A y B a la vez.

Complementario: Ocurre A^c si y sólo si no ocurre A.

Diferencia de sucesos: Ocurre $A \backslash B$ si ocurre A, pero no ocurre B. Por tanto, $A \backslash B = A \cap B^c$.

Sucesos incompatibles: Dos sucesos A y B se dicen incompatibles si no pueden ocurrir a la vez. Dicho de otro modo, que ocurra A y B es imposible. Lo escribimos como $A \cap B = \emptyset$.

Suceso contenido en otro: Diremos que A está contenido en B, y lo denotamos por $A \subset B$, si siempre que ocurra A también sucede B.

Ejemplo 5: Estudiamos el experimento aleatorio consistente en el lanzamiento de un dado, y consideramos los sucesos:

- A = El resultado es par = $\{2, 4, 6\}$,
- B = El resultado es múltiplo de tres = $\{3, 6\}$.

El suceso "que salga par y múltiplo de tres" se puede expresar como la intersección $A \cap B = \{2, 4, 6\} \cap \{3, 6\} = \{6\}$.
De la misma manera, el suceso "que salga par o múltiplo de tres" se puede expresar como la unión $A \cup B = \{2, 4, 6\} \cup \{3, 6\} = \{2, 3, 4, 6\}$.

Propiedades

Asociativa	$A \cup (B \cup C) = (A \cup B) \cup C$	$A \cap (B \cap C) = (A \cap B) \cap C$
Conmutativa	$A \cup B = B \cup A$	$A \cap B = B \cap A$
Distributiva	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
Neutro	$\emptyset$ para la unión	$A \cup \emptyset = A$
	Ω para la intersección	$A \cap \Omega = A$
Complementario	$A \cup A^c = \Omega$	$A \cap A^c = \emptyset$
Leyes de de Morgan	$(A \cup B)^c = A^c \cap B^c$	$(A \cap B)^c = A^c \cup B^c$

Ejercicio 1: Lanzamos un dado y consideramos los sucesos

- A = El resultado es par.
- B = El resultado es mayor que 2.

Indica cuáles son los sucesos $A \cup B$, $A \cap B$. ¿son los sucesos A y B incompatibles?, ¿son los sucesos A y A^c incompatibles?

3 Definiciones de probabilidad

El principal objetivo de un experimento aleatorio suele ser determinar con qué probabilidad ocurre cada uno de los sucesos elementales. ¿Pero cómo asignamos probabilidades a los sucesos?

3.1 Definición clásica o de Laplace

Nos encontramos ante un experimento, con su colección de sucesos, y nos preguntamos cómo tenemos que actuar para asignarle a cada suceso un número entre 0 y 1 que represente la probabilidad de que el suceso ocurra.
Cuando el espacio muestral es finito, el problema se reduce a asignar probabilidades a los sucesos elementales. Las probabilidades de los demás sucesos se obtendrán sumando las de los sucesos

elementales que lo componen (suma finita).
Sin duda el caso más fácil es aquél en el que no tenemos razones para suponer que unos sucesos sean más probables que otros.

Cuando, siendo el espacio muestral Ω finito, todos los sucesos elementales tienen la misma probabilidad, diremos que son **equiprobables** y podremos utilizar la conocida **Regla de Laplace**

$$P(A) = \frac{\text{casos favorables a } A}{\text{casos posibles}}.$$

Ejercicio 2: Lanzamos dos dados y sumamos sus puntuaciones. ¿Cuál es la probabilidad de obtener un 2? ¿Cuál es la probabilidad de obtener un 7?

3.2 Definición axiomática (Kolmogorov 1933)

Las dificultades que presenta la definición de probabilidad se han resuelto a principios del siglo XX mediante la utilización de una definición axiomática de la probabilidad.

Sea Ω el espacio muestral, y sea $\mathcal{P}(\Omega)$ el conjunto formado por todos los sucesos. Se define la **probabilidad** como una aplicación $P : \mathcal{P}(\Omega) \longrightarrow [0, 1]$ que cumple las siguientes condiciones:

- $P(\Omega) = 1$
  La probabilidad del suceso seguro es 1.
- $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$
  Si A y B son sucesos incompatibles, entonces la probabilidad de su unión es la suma de sus probabilidades.

A partir de la definición anterior se puede deducir que:

1. $P(\emptyset) = 0$
2. Si $A_1, A_2, \ldots, A_n$ son sucesos incompatibles dos a dos, se cumple
$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$
3. $P(A^c) = 1 - P(A)$
4. Si $A \subset B$, entonces $P(A) \leq P(B)$
5. Si A y B son dos sucesos cualesquiera (ya no necesariamente incompatibles) se cumple
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

En esta definición está basado todo el Cálculo de Probabilidades en el siglo XX.

4 Probabilidad condicionada

El concepto de probabilidad condicionada es uno de los más importantes en Teoría de la Probabilidad. La probabilidad condicionada pone de manifiesto el hecho de que las probabilidades cambian cuando la información disponible cambia. Por ejemplo, ¿Cuál es la probabilidad de sacar un 1 al lanzar un dado? ¿Y cuál es la probabilidad de sacar un 1 al lanzar un dado si sabemos que el resultado ha sido un número impar?

Ejemplo 6: Si lanzamos un dado, la probabilidad de obtener un 1 es 1/6, pero si disponemos de la información adicional de que el resultado obtenido ha sido impar entonces reducimos los casos posibles de 6 a 3 (sólo puede ser un 1, un 3 o un 5), con lo cual la probabilidad de obtener un 1 (sabiendo que el resultado ha sido impar) es 1/3.

Supongamos entonces que en el estudio de un experimento aleatorio nos interesa conocer la probabilidad de que ocurra un cierto suceso A pero dispongamos de información previa sobre el experimento: sabemos que el suceso B ha ocurrido. Está claro que ahora la probabilidad de A ya no es la misma que cuando no sabíamos nada sobre B.

La **probabilidad del suceso A condicionada al suceso B** se define:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}\ , \qquad \text{siendo} \qquad P(B) \neq 0$$

También se deduce de manera inmediata que $P(A \cap B) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B)$.

Ejemplo 7: Se ha realizado una encuesta en Santiago para determinar el número de lectores de La Voz y de El Correo. Los resultados fueron que el 35% de los encuestados lee La Voz, el 20% de los encuestados lee El Correo. Además, analizando las respuestas se concluye que el 5% de los encuestados lee ambos periódicos. Si se selecciona al azar un lector de El Correo, ¿cuál es la probabilidad de que lea también La Voz?
En primer lugar, vamos a ponerle nombre a los sucesos. Denotamos

- A= Es lector de La Voz.
- B= Es lector de El Correo.

Fíjate en que la información que nos da el problema es:

- $P(A) = 0.35$.
- $P(B) = 0.2$.
- $P(A \cap B) = 0.05$.

Lo que nos preguntan es un probabilidad condicionada. Sabiendo que una persona es lectora de El Correo, ¿Cuál es la probabilidad de que también sea lector de La Voz? Es decir, debemos calcular

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.05}{0.2} = 0.25.$$

5 Independencia de sucesos

Dos sucesos A y B son **independientes** si

$$P(A \cap B) = P(A) \cdot P(B)$$

Comentarios:

- Si $P(B) > 0$, A y B son independientes si y sólo si $P(A/B) = P(A)$, esto es, el conocimiento de la ocurrencia de B no modifica la probabilidad de ocurrencia de A.
- Si $P(A) > 0$, A y B son independientes si y sólo si $P(B/A) = P(B)$, esto es, el conocimiento de la ocurrencia de A no modifica la probabilidad de ocurrencia de B.
- No debemos confundir sucesos *independientes* con sucesos *incompatibles*: los sucesos incompatibles son los más dependientes que puede haber. Por ejemplo, si en el lanzamiento de una moneda consideramos los sucesos incompatibles 'salir cara' y 'salir cruz', el conocimiento de que ha salido *cara* nos da el máximo de información sobre el otro suceso: ya que ha salido cara es imposible que haya salido *cruz*.

Recuerda que los dos sucesos son incompatibles si $A \cap B = \emptyset$

- Si los sucesos A y B son independientes, también lo son los sucesos A y B^c; los sucesos A^c y B; y los sucesos A^c y B^c.

Ejercicio 3: Se estima que entre la población de Estados Unidos, el 55% padece de obesidad, el 20% es hipertensa, y el 60% es obesa o hipertensa. ¿Es independiente el que una persona sea obesa de que padezca hipertensión?

6 Teoremas clásicos: Regla del producto, ley de probabilidades totales y teorema de Bayes

En esta sección veremos tres teoremas muy importantes, tanto a nivel teórico como para la resolución de ejercicios. Los enunciaremos en su forma más general, aunque después veremos por medio de ejemplos que su aplicación no es complicada.

6.1 Regla del producto

Si tenemos los sucesos $A_1, A_2, \ldots, A_n$ tales que $P(A_1 \cap A_2 \cap \ldots \cap A_n) \neq 0$, entonces se cumple

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1) \cdot P(A_2/A_1) \cdot P(A_3/A_1 \cap A_2) \cdots P(A_n/A_1 \cap A_2 \cap \ldots \cap A_{n-1})$$

La regla del producto se utiliza en experimentos aleatorios que están formados por etapas consecutivas (de la 1 a la n) y nos permite calcular la probabilidad de que ocurra una concatenación (intersección) de sucesos a lo largo de las etapas (A_1 en la primera etapa y A_2 en la segunda etapa y ...y A_n en la etapa n). Esta probabilidad queda expresada como el producto de la probabilidad inicial $P(A_1)$ y las probabilidades en cada etapa condicionadas a las etapas anteriores, conocidas como *probabilidades de transición*.

Ejemplo 8: Un grupo de investigadores de un laboratorio trata de desarrollar una vacuna efectiva contra parásitos gastrointestinales. La vacuna en la que trabajan en la actualidad es capaz de matar en la primera aplicación al 80% de los parásitos gastrointestinales. Los parásitos supervivientes desarrollan resistencia y en cada aplicación posterior de la vacuna el porcentaje de parásitos muertos se reduce a la mitad del verificado en la aplicación inmediatamente anterior: así en la segunda aplicación muere el 40% de los parásitos supervivientes de la primera aplicación, en la tercera aplicación muere el 20%, etc.

a) ¿Cuál es la probabilidad de que un parásito sobreviva a dos aplicaciones de la vacuna?

b) ¿Cuál es la probabilidad de que un parásito sobreviva a tres aplicaciones de la vacuna?

Como siempre, en primer lugar vamos a vamos a ponerle nombre a los sucesos. Denotamos

- A_1= El parásito sobrevive a la primera aplicación de la vacuna.
- A_2= El parásito sobrevive a la segunda aplicación de la vacuna.
- A_3= El parásito sobrevive a la tercera aplicación de la vacuna,...

Fíjate en que la información que nos da el problema es:

- $P(A_1) = 0.2$.
- $P(A_2/A_1) = 0.6$.
- $P(A_3/A_1 \cap A_2) = 0.8$.

Aplicando la regla de la cadena podemos contestar a las dos preguntas del problema.

a) La probabilidad de que un parásito sobreviva a dos aplicaciones de la vacuna será

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2/A_1) = 0.2 \cdot 0.6 = 0.12.$$

b) La probabilidad de que un parásito sobreviva a tres aplicaciones de la vacuna será

$$P(A_1 \cap A_2 \cap A_3) = P(A_1) \cdot P(A_2/A_1) \cdot P(A_3/A_1 \cap A_2) = 0.2 \cdot 0.6 \cdot 0.8 = 0.096.$$

6.2 Ley de las probabilidades totales

El segundo teorema es la llamada **ley de las probabilidades totales**. Descompone la probabilidad de un suceso en la segunda etapa en función de lo que ocurrió en la etapa anterior. Previamente al enunciado de este teorema damos una definición.

Sistema completo de sucesos. Es una partición del espacio muestral, esto es, es una colección de sucesos $A_1, A_2, \ldots, A_n$ (subconjuntos del espacio muestral) verificando $A_1 \cup A_2 \cup \ldots \cup A_n = \Omega$ (son exhaustivos, cubren todo el espacio muestral) y además son incompatibles dos a dos (si se verifica uno de ellos, no puede a la vez ocurrir ninguno de los otros).

Ley de las probabilidades totales. Sea $A_1, A_2, \ldots, A_n$ un sistema completo de sucesos. Entonces se cumple que:

$$P(B) = P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2) + \cdots + P(A_n) \cdot P(B/A_n)$$

Ejemplo 9: Se sabe que una determinada enfermedad coronaria es padecida por el 7% de los fumadores y por el 2.5% de los no fumadores. Si en una población de 5.000 habitantes hay 600 fumadores, ¿cuál es la probabilidad de que una persona elegida al azar sufra dicha enfermedad?
En este caso:

- E = Sufre enfermedad coronaria.
- A_1 = Es fumador.
- A_2 = Es no fumador.

Fijate que A_1, A_2 un sistema completo de sucesos. La información que nos da el problema es:

- $P(E/A_1) = 0.07$.
- $P(E/A_2) = 0.025$.
- $P(A_1) = 600/5000 = 0.12$.
- $P(A_2) = 4400/5000 = 0.88$ (también se puede calcular como $P(A_2) = 1 - P(A_1)$ ya que son sucesos complementarios).

Entonces, por la ley de probabilidades totales

$$P(E) = P(A_1) \cdot P(E/A_1) + P(A_2) \cdot P(E/A_2) = 0.12 \cdot 0.07 + 0.88 \cdot 0.025 = 0.0304$$

6.3 Teorema de Bayes

Por último, tratamos el teorema de Bayes. Consideremos un experimento que se realiza en dos etapas: en la primera, tenemos un sistema completo de sucesos $A_1, A_2, \ldots, A_n$ con probabilidades $P(A_i)$ que denominamos *probabilidades a priori*. En una segunda etapa, ha ocurrido el suceso B y se conocen las probabilidades condicionadas $P(B/A_i)$ de obtener en la segunda etapa el suceso B cuando en la primera etapa se obtuvo el suceso A_i, $i = 1, \ldots, n$.
En estas condiciones el teorema de Bayes permite calcular las probabilidades $P(A_i/B)$, que son probabilidades condicionadas en sentido inverso. Reciben el nombre de *probabilidades a posteriori*, pues se calculan después de haber observado el suceso B.

Teorema de Bayes. En las condiciones anteriores,

$$P(A_i/B) = \frac{P(A_i) \cdot P(B/A_i)}{P(B)}$$

Además, aplicando en el denominador la ley de probabilidades totales:

$$P(A_i/B) = \frac{P(A_i) \cdot P(B/A_i)}{P(A_1) \cdot P(B/A_1) + P(A_2) \cdot P(B/A_2) + \cdots + P(A_n) \cdot P(B/A_n)}$$

Ejemplo 10: Volvamos al Ejemplo 9 y supongamos ahora que llega a nuestra consulta una persona que sufre la enfermedad coronaria citada. ¿Cuál es la probabilidad de que dicha persona sea fumadora?
En este caso nos están preguntando $P(A_1/E)$. Por el Teorema de Bayes,

$$P(A_1/E) = \frac{P(A_1) \cdot P(E/A_1)}{P(E)} = \frac{0.12 \cdot 0.07}{0.0304} = 0.2763$$

7 Pruebas diagnósticas: Sensibilidad y especificidad

La epidemiología es la ciencia que estudia la frecuencia de aparición de la enfermedad y de sus determinantes en la población

Las leyes de probabilidad que hemos visto hasta ahora son fundamentales en el campo de ciencias de la salud, en la evaluación de pruebas diagnósticas. Entendemos por prueba diagnóstica cualquier procedimiento que pretenda determinar en un paciente la presencia de cierta condición, supuestamente patológica, no susceptible de ser observada directamente. Antes de estudiar los procedimientos estadísticos que permiten evaluar la validez de las pruebas diagnósticas introduciremos dos conceptos muy importantes en epidemiología: el de prevalencia e incidencia de una enfermedad.

Prevalencia: proporción de individuos de la población que presentan la enfermedad.

Incidencia: medida del número de casos nuevos de una enfermedad en un período determinado. Podría considerarse como una tasa que cuantifica las personas que enfermarán en un periodo de tiempo.

A los médicos les interesa tener mayor capacidad para determinar sin equivocarse la presencia o ausencia de una enfermedad en un paciente a partir de los resultados (positivos o negativos) de pruebas o de los síntomas (presentes o ausentes) que se manifiestan.
También es importante conocer la probabilidad de obtención de resultados positivos o negativos de las pruebas y la probabilidad de la presencia o ausencia de un determinado síntoma en pacientes con o sin una determinada enfermedad.
Es importante tener en cuenta que las pruebas de detección no siempre son infalibles y que los procedimientos pueden dar **falsos positivos** o **falsos negativos**.

- Un **falso positivo** resulta cuando una prueba indica que el estado es positivo, cuando en realidad es negativo.
- Un **falso negativo** resulta cuando una prueba indica que el estado es negativo, cuando en realidad es positivo.

Para evaluar la utilidad de los resultados de una prueba, debemos contestar a las siguientes preguntas:

1. Dado que un individuo tiene la enfermedad, ¿qué probabilidad existe de que la prueba resulte positiva?

2. Dado que un individuo no tiene la enfermedad, ¿qué probabilidad existe de que la prueba resulte negativa?
3. Dada un resultado positivo de una prueba de detección, ¿qué probabilidad existe de que el individuo tenga la enfermedad?
4. Dada un resultado negativo de una prueba de detección, ¿qué probabilidad existe de que el individuo no tenga la enfermedad?

Relacionando estas ideas con los conceptos de probabilidad que hemos visto anteriormente, definiremos los siguientes sucesos:

- + = El resultado de la prueba diagnóstica es positivo.
- − = El resultado de la prueba diagnóstica es negativo.
- E = El paciente tiene la enfermedad.
- S = El paciente no tiene la enfermedad.

Definimos entonces los siguientes conceptos, que responderán a las preguntas 1 y 2.

Sensibilidad: La sensibilidad de una prueba es la probabilidad de un resultado positivo de la prueba dada la presencia de la enfermedad. Se trata, por lo tanto, de una probabilidad condicionada, la de que el resultado de la prueba sea positivo condicionada a que el paciente sufre la enfermedad.

$$\text{Sensibilidad} = P(+/E)$$

Especificidad: La especificidad de una prueba es la probabilidad de un resultado negativo de la prueba dada la ausencia de la enfermedad. Se trata, por lo tanto, de una probabilidad condicionada, la de que el resultado de la prueba sea negativo condicionada a que el paciente está sano.

$$\text{Especificidad} = P(-/S)$$

Para responder a las preguntas 3 y 4, definimos:

Valor predictivo positivo: El valor predictivo positivo de una prueba es la probabilidad de que un individuo tenga la enfermedad, dado que el individuo presenta un resultado positivo en la prueba de detección. Se trata, de nuevo, de una probabilidad condicionada.

$$\text{Valor predictivo positivo} = P(E/+)$$

Valor predictivo negativo: El valor predictivo negativo de una prueba es la probabilidad de que un individuo esté sano, dado que el individuo presenta un resultado negativo en la prueba de detección.

$$\text{Valor predictivo negativo} = P(S/-)$$

El valor predictivo positivo de una prueba puede obtenerse a partir del conocimiento de la sensibilidad y especificidad de la prueba y de la probabilidad de la enfermedad aplicando la regla de Bayes:

$$P(E/+) = \frac{P(E)\cdot P(+/E)}{P(+)} = \frac{P(E)\cdot P(+/E)}{P(E)\cdot P(+/E) + P(S)\cdot P(+/S)}.$$

Del mismo modo, el valor predictivo negativo de una prueba puede obtenerse también por la regla de Bayes.

Ejercicio 4: [Bioestadística. Daniel, W. W. (2006)] La siguiente tabla muestra los resultados de la evaluación de prueba de detección en la que participaron una muestra aleatoria de 650 individuos con la enfermedad y una segunda muestra aleatoria independiente de 1200 individuos sin la enfermedad.

	Enfermedad	
Resultado	Presente	Ausente
Positivo	490	70
Negativo	160	1130

a) Calcula la sensibilidad de la prueba.

b) Calcula la especificidad de la prueba.

c) Si la tasa de enfermedad en la población en general es 0.002, ¿cuál es el valor predictivo positivo de la prueba?

Bioestadística. Curso 2014-2015
Capítulo 3

Carmen Mª Cadarso, Mª del Carmen Carollo, Xosé Luis Otero, Beatriz Pateiro

Contents

1 Introducción

En el tema de Estadística Descriptiva hemos estudiado variables, entendiéndolas como mediciones que se efectúan sobre los individuos de una muestra. Así, la Estadística Descriptiva nos permitía analizar los distintos valores que tomaban las variables sobre una muestra ya observada. Se trataba, pues, de un estudio posterior a la realización del experimento aleatorio.
En este tema trataremos las variables situándonos antes de la realización del experimento aleatorio. Por tanto, haremos uso de los conceptos del tema anterior (Probabilidad), mientras que algunos desarrollos serán análogos a los del tema de Estadística Descriptiva.

2 Variable aleatoria

De manera informal, una **variable aleatoria** es un valor numérico que corresponde al resultado de un experimento aleatorio. Por ejemplo, una variable X como resultado de lanzar una moneda al aire puede tomar el valor 1 si el resultado es cara y 0 si es cruz. De este modo, escribiremos, por ejemplo, $P(X = 1) = 0.5$. Otro ejemplo de variable aleatoria, Y, puede ser el resultado de medir en °C la temperatura corporal de adultos varones sanos. Cuando se han tomado muchísimas observaciones (infinitas), se puede llegar a la conclusión, por ejemplo, que la probabilidad de que la temperatura corporal sea inferior a 36.8 °C es igual a 0.8, lo que escribimos con $P(Y < 36.8) = 0.8$.

Definición 1. *Llamamos* **variable aleatoria** *a una aplicación del espacio muestral asociado a un experimento aleatorio en $\mathbb{R}$, que a cada resultado de dicho experimento le asigna un número real, obtenido por la medición de cierta característica.*

$$\begin{array}{rccc} X: & \Omega & \longrightarrow & \mathbb{R} \\ & \omega & \longrightarrow & X(\omega) \end{array}$$

Denotamos la variable aleatoria por una letra mayúscula. El conjunto imagen de esa aplicación es el conjunto de valores que puede tomar la variable aleatoria, que serán denotados por letras minúsculas. Las variables aleatorias son equivalentes a las variables que analizábamos en el tema de Estadística Descriptiva. La diferencia es que en el tema de Estadística Descriptiva se trabajaba sobre una muestra de datos y ahora vamos a considerar que disponemos de toda la población (lo cual es casi siempre imposible en la práctica). Ahora vamos a suponer que podemos calcular las probabilidades de todos los sucesos resultantes de un experimento aleatorio.

Al igual que en el tema de Estadística Descriptiva, las variables aleatorias se pueden clasificar en discretas y continuas

De modo idéntico a lo dicho en el tema de Descriptiva, podemos clasificar las variables aleatorias en **discretas** y **continuas** en función del conjunto de valores que pueden tomar. Así, una variable aleatoria será discreta si dichos valores se encuentran separados entre sí. Por tanto será representable por conjuntos discretos. Una variable aleatoria será continua cuando el conjunto de valores que puede tomar es un intervalo.

2.1 Variables aleatorias discretas.

Una variable aleatoria es **discreta** cuando toma una cantidad numerable (que se pueden contar) de valores. Por ejemplo, el número de caras al lanzar dos veces una moneda o el número de pacientes con enfermedades articulares en centros de salud.
Si X es una variable discreta, su distribución viene dada por los valores que puede tomar y las probabilidades de que aparezcan. Si $x_1 < x_2 < .. < x_n$ son los posibles valores de la variable X, las

diferentes probabilidades de que ocurran estos sucesos,

$$\begin{aligned} p_1 &= P(X = x_1), \\ p_2 &= P(X = x_2), \\ &\vdots \\ p_n &= P(X = x_n). \end{aligned}$$

constituyen la distribución de X.

Definición 2. *La función* $P(X = x)$ *se denomina* **función de probabilidad o función de masa**.

La función de probabilidad se puede representar análogamente al diagrama de barras.

Ejercicio 1: Se lanza dos veces una moneda equilibrada. Sea X la variable que expresa el número de caras en los dos lanzamientos. Halla y representa la función de probabilidad de X.

Ejercicio 2: Sea X la variable aleatoria que expresa número de pacientes con enfermedades articulares en centros de salud con las siguientes probabilidades:

x_i	0	1	2	3	4	5	6	7
p_i	0.230	0.322	0.177	0.155	0.067	0.024	0.015	0.01

Comprueba que se trata efectivamente de una función de probabilidad y represéntala.

Definición 3. *La* **función de distribución** *de una variable aleatoria se define como:*

Calcularemos para variables aleatorias discretas su función de masa y su función de distribución

$$\begin{aligned} F: \quad \mathbb{R} &\longrightarrow \mathbb{R} \\ x_0 &\longrightarrow F(x_0) = P(X \leq x_0) \end{aligned}$$

El diagrama de barras de frecuencias acumuladas para variables discretas del tema 1 se puede reinterpretar en términos de probabilidades y da lugar a lo que recibe el nombre de **función de distribución,** $F(x)$, definida para cada punto x_0 como la probabilidad de que la variable aleatoria tome un valor menor o igual que x_0,

$$F(x_0) = P(X \leq x_0).$$

La función de distribución es siempre no decreciente y verifica que,

$$\begin{aligned} F(-\infty) &= 0, \\ F(+\infty) &= 1. \end{aligned}$$

Suponiendo que la variable X toma los valores $x_1 < x_2 < \ldots < x_n$, los puntos de salto de la función

de distribución vienen determinados por:

$$\begin{aligned} F(x_1) &= P(X \leq x_1) = P(X = x_1) \\ F(x_2) &= P(X \leq x_2) = P(X = x_1) + P(X = x_2) \\ &\vdots \\ F(x_n) &= P(X \leq x_n) = P(X = x_1) + \ldots + P(X = x_n) = 1 \end{aligned}$$

Obsérva la función de distribución es igual a uno en el máximo de todos los valores posibles.

Ejercicio 3: Calcular la función de distribución de la variable X en el Ejercicio 1.

Ejercicio 4: Calcular la función de distribución de la variable X en el Ejercicio 2.

Ejercicio 5: Calcula la probabilidad de que el número de caras sea al menos 1 en el Ejercicio 1.

Ejercicio 6: Calcula la probabilidad de que el número de pacientes con enfermedades articulares sea menor o igual que 4 y la probabilidad de que haya más de dos pacientes de este tipo en un centro de salud con la información del Ejercicio 2.

3 Medidas características de una variable aleatoria discreta.

Los conceptos que permiten resumir una distribución de frecuencias utilizando valores numéricos pueden utilizarse también para describir la distribución de probabilidad de una variable aleatoria. Las definiciones son análogas a las introducidas en el tema 1.

3.1 Media o esperanza.

Se define la **media poblacional** o **esperanza** de una variable aleatoria discreta como la media de sus posibles valores $x_1, x_2, \ldots, x_k$ ponderados por sus respectivas probabilidades $p_1, p_2, \ldots, p_k$, es decir,

$$\mu = \mathbb{E}(X) = x_1 p_1 + x_2 p_2 + \ldots + x_k p_k = \sum_{i=1}^{k} x_i p_i .$$

Ejercicio 7: Calcula la media de pacientes con enfermedades articulares del Ejercicio 2.

La interpretación de la media o esperanza es el valor esperado al realizar el experimento con la variable aleatoria. Además, la media puede verse también como el valor central de la distribución de probabilidad.

3.2 Varianza.

Se define la **varianza poblacional** de una variable aleatoria discreta con valores $x_1, x_2, ..., x_k$ como la media ponderada de las desviaciones a la media al cuadrado,

$$\sigma^2 = \text{Var}(X) = \sum_{i=1}^{k} (x_i - \mu)^2 \, p_i.$$

Ejercicio 8: Calcula la varianza de pacientes con enfermedades articulares del Ejercicio 2.

La interpretación de la varianza es la misma que para un conjunto de datos: es un valor no negativo que expresa la dispersión de la distribución alrededor de la media. Además, se puede calcular la **desviación típica poblacional** σ como la raíz cuadrada de la varianza. Los valores pequeños de σ indican concentración de la distribución alrededor de la esperanza y valores grandes corresponden a distribuciones más dispersas.

4 Principales modelos de distribuciones discretas

Estudiaremos ahora distribuciones de variables aleatorias que han adquirido una especial relevancia por ser adecuadas para modelizar una gran cantidad de situaciones. Presentaremos modelos de variables discretas y caracterizaremos estas distribuciones mediante la distribución de probabilidad. Calcularemos también los momentos (media y varianza) y destacaremos las propiedades de mayor utilidad.

4.1 Distribución de Bernoulli

En muchas ocasiones nos encontramos ante experimentos aleatorios con sólo dos posibles resultados: Éxito y fracaso (cara o cruz en el lanzamiento de una moneda, ganar o perder un partido, aprobar o suspender un examen, una prueba diagnóstica da positivo o negativo...). Se pueden modelizar estas situaciones mediante la variable aleatoria

$$X = \begin{cases} 1 & \text{si Éxito} \\ 0 & \text{si Fracaso} \end{cases}$$

Lo único que hay que conocer es la probabilidad de éxito, p, ya que los valores de X son siempre los mismos y la probabilidad de fracaso es $q = 1 - p$.

Definición 4. *Si denotamos por p a la probabilidad de éxito, entonces diremos que la variable X tiene* **distribución de Bernoulli de parámetro p**, *y lo denotamos* $X \in \text{Bernoulli}(p)$. *La distribución de probabilidad de* $X \in$ *Bernoulli(p) viene dada por*

X	0	1
$P(X = x_i)$	$1 - p$	p

Por tanto, la probabilidad de éxito p determina plenamente la distribución de Bernoulli. La media y la varianza de una Bernoulli(p) son:

- $\mu = p$.

- $\sigma^2 = p \cdot (1 - p)$.

Como ejemplo, la Figura 1 muestra la función de masa de una variable con distribución de Bernoulli para $p = 0.8$.

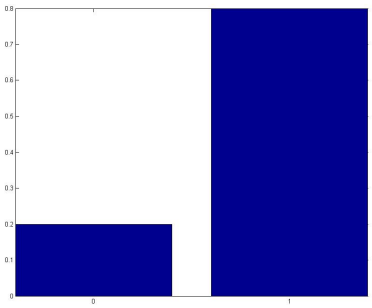


Figure 1: Función de masa de una Bernoulli(0.8).

4.2 Distribución binomial

Empezando con una prueba de Bernoulli con probabilidad de éxito p, vamos a construir una nueva variable aleatoria al repetir n veces la prueba de Bernoulli.

Ejemplo 1: Supongamos que lanzamos un dado normal 5 veces y queremos determinar la probabilidad de que exactamente en 3 de esos 5 lanzamientos salga el 6.
Cada lanzamiento es independiente de los demás y podemos considerarlo como un ensayo de Bernoulli, donde el éxito es sacar un 6 ($p = 1/6$). Lo que hacemos es repetir el experimento 5 veces y queremos calcular la probabilidad de que el número de éxitos sea igual a 3 (es decir, obtener 3 éxitos y 2 fracasos)

La distribución binomial sirve para modelizar situaciones en las que nos interesa contar el número de éxitos en n repeticiones de una prueba de Bernoulli con probabilidad de éxito p

La variable aleatoria **binomial** X es el número de éxitos en n repeticiones de una prueba de Bernoulli con probabilidad de éxito p. Debe cumplirse:

- Cada prueba individual puede ser un éxito o un fracaso.
- La probabilidad de éxito, p, es la misma en cada prueba.
- Las pruebas son independientes. El resultado de una prueba no tiene influencia sobre los resultados siguientes.

Definición 5. *La variable aleatoria X que representa el número de éxitos en n intentos independientes, siendo la probabilidad de éxito en cada intento p, diremos que tiene* **distribución binomial de parámetros n y p**.*Lo denotamos $X \in \text{Binomial}(n, p)$ o $X \in \text{Bin}(n, p)$. La distribución binomial es discreta y toma los valores $0, 1, 2, 3, \ldots, n$ con probabilidades*

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \qquad si \;\; k \in \{0, 1, 2, \ldots, n\}$$

donde el coeficiente binomial

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

*representa el número de subconjuntos diferentes de k elementos que se pueden definir a partir de un total de n elementos (***combinaciones** *de n elementos tomados de k en k).*

La media y la varianza de una Bin(n, p) son:

- $\mu = n \cdot p$.
- $\sigma^2 = n \cdot p \cdot (1 - p)$.

Como ejemplo, la Figura 2 muestra las funciones de masa de una variable con distribución binomial de parámetros $n = 5$ y $p = 1/6$ y una variable con distribución binomial de parámetros $n = 60$ y $p = 1/6$.

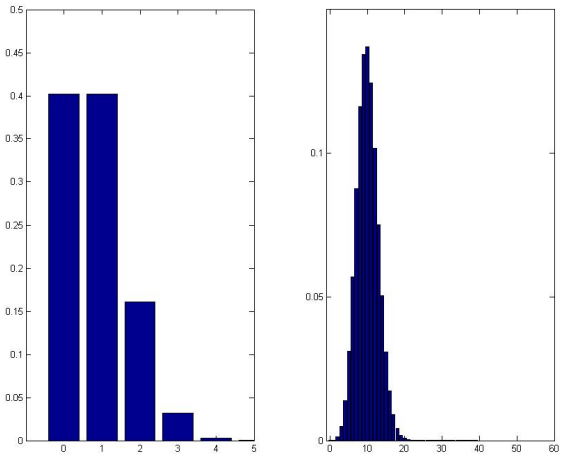


Figure 2: En la izquierda, función de masa de una Bin$(5, 1/6)$. En la derecha, función de masa de una Bin$(60, 1/6)$.

4.3 Distribución de Poisson

En muchas circunstancias (llamadas a una centralita telefónica, átomos que pueden emitir una radiación, ...) el número de individuos susceptibles de dar lugar a un éxito es muy grande. Para modelizar estas situaciones mediante una distribución binomial tendremos problemas al escoger el parámetro n (demasiado grande o incluso difícil de determinar) y al calcular la distribución de probabilidad (la fórmula resulta inviable). Sin embargo, se ha observado que si mantenemos constante la media $\mathbb{E}(X) = np$ y hacemos $n \to \infty$, la distribución de probabilidad de la binomial tiende a una nueva distribución, que llamaremos de Poisson de parámetro $\lambda = np$.

Definición 6. *Una variable aleatoria X tiene* **distribución de Poisson de parámetro** λ*, y lo denotamos* $X \in$ Poisson(λ)*, si es discreta y*

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!} \qquad si \;\; k \in \{0, 1, 2, 3, \ldots\}$$

La media y la varianza de la Poisson de parámetro λ son:

- $\mu = \lambda$

- $\sigma^2 = \lambda$

Como ejemplo, la Figura 3 muestra las funciones de masa de una variable con distribución de Poisson de parámetro $\lambda = 2$ y una variable con distribución de Poisson de parámetro $\lambda = 15$.

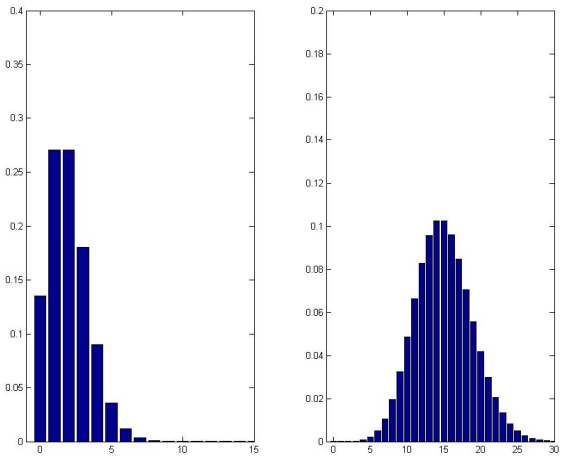


Figure 3: En la izquierda, función de masa de una Poisson(2). En la derecha, función de masa de una Poisson(15).

En la práctica usaremos la distribución de Poisson como aproximación de la distribución binomial cuando n sea grande y p pequeño, en base al límite que hemos visto. Usaremos el siguiente criterio:

- Si $n > 50$, $p < 0.1$ entonces la distribución binomial de parámetros n y p puede ser aproximada por una Poisson de parámetro $\lambda = np$.

Ejemplo 2: La probabilidad de que una persona se desmaye en un concierto es $p = 0.005$. ¿Cuál es la probabilidad de que en un concierto al que asisten 3000 personas se desmayen 18? La variable $X =$*Número de personas que se desmayan en el concierto* sigue una distribución Bin$(3000, 0.005)$. Queremos calcular

$$P(X = 18) = \begin{pmatrix} 3000 \\ 18 \end{pmatrix} \cdot 0.005^{18} \cdot 0.995^{2982}.$$

Estos valores están fuera de las tablas de la binomial y son difíciles de calcular, por eso es preferible aproximar por una Poisson de parámetro $\lambda = np = 3000 \cdot 0.005 = 15$. Entonces:

$$P(X = 18) \approx P(\text{Poisson}(15) = 18) = e^{-15}\frac{15^{18}}{18!} = 0.07061.$$

Ejercicio 9: Se sabe que la probabilidad de que un individuo reaccione desfavorablemente tras la inyección de una vacuna es de 0.002. Determina la probabilidad de que en un grupo de 2000 personas vacunadas haya como mucho tres que reaccionen desfavorablemente.

Aunque la distribución de Poisson se ha obtenido como forma límite de una distribución Binomial, tiene muchas aplicaciones sin conexión directa con las distribuciones binomiales. Por ejemplo, la distribución de Poisson puede servir como modelo del número de éxitos que ocurren durante un intervalo de tiempo o en una región específica.
Definimos el **proceso de Poisson** como un experimento aleatorio que consiste en contar el número de ocurrencias de determinado suceso en un intervalo de tiempo, verificando:

La distribución de Poisson sirve como aproximación de la distribución binomial Bin(n, p) cuando n es grande y p pequeño y también es adecuada para modelizar situaciones en las que nos interesa contar el número de ocurrencias de un determinado suceso en un intervalo de tiempo

- El número medio de sucesos por unidad de tiempo es constante. A esa constante la llamamos **intensidad del proceso**.
- Los números de ocurrencias en subintervalos disjuntos son independientes.

En un proceso de Poisson, consideremos $X =$"número de ocurrencias en un subintervalo". Entonces X tiene distribución de Poisson, cuyo parámetro es proporcional a la longitud del subintervalo.

Ejemplo 3: El número de nacimientos en un hospital constituye un proceso de Poisson con intensidad de 21 nacimientos por semana. ¿Cuál es la probabilidad de que se produzcan al menos tres nacimientos la próxima semana?

$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)] \\ &= 1 - \left[e^{-21} \frac{21^0}{0!} + e^{-21} \frac{21^1}{1!} + e^{-21} \frac{21^2}{2!} \right]. \end{aligned}$$

Bioestadística. Curso 2014-2015
Capítulo 4

Carmen Mª Cadarso, Mª del Carmen Carollo, Xosé Luis Otero, Beatriz Pateiro

Contents

1 Introducción

En el capítulo anterior hemos estudiado variables aleatorias discretas. Recuerda que una variable aleatoria es un valor numérico que corresponde al resultado de un experimento aleatorio. Podemos clasificar las variables aleatorias en discretas y continuas en función del conjunto de valores que pueden tomar. Estudiaremos en este tema variables aleatorias continuas y nos centraremos en un modelo de distribución continua (la distribución normal) que ha adquirido una especial relevancia por ser adecuada para modelizar una gran cantidad de situaciones prácticas.

2 Variables aleatorias continuas

Una variable aleatoria es **continua** cuando puede tomar cualquier valor en un intervalo. Por ejemplo, el peso de una persona o el contenido de paracetamol en un lote de pastillas.
El estudio de las variables continuas es más sutil que el de las discretas. Recordemos que la construcción del histograma es más delicado que el del diagrama de barras ya que depende de la elección de las clases.
Se ha comprobado en la práctica que tomando más observaciones de una variable continua y haciendo más finas las clases, el histograma tiende a estabilizarse en una curva suave que describe la distribución de la variable (véase la Figura 1). Esta función, $f(x)$, se llama **función de densidad** de la variable X. La función de densidad constituye una idealización de los histogramas de frecuencia o un **modelo** del cual suponemos que proceden las observaciones.

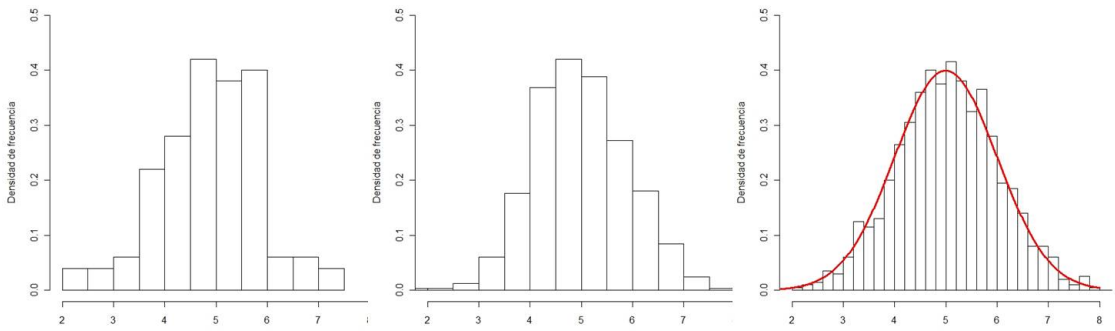


Figure 1: Histograma de la capacidad (en ml.) de $n = 100$, $n = 500$ y $n = 1000$ jeringas producidas por la empresa Clinic, que se dedica a la venta de material clínico. Tomando más observaciones y haciendo más finas las clases, el histograma tiende a estabilizarse en una curva suave (en rojo) que describe la distribución de la variable.

Definición 1. *Llamamos* **función de densidad** *de una variable aleatoria continua X a una aplicación $f : \mathbb{R} \to \mathbb{R}$ no negativa y tal que*

$$P(X \leq x_0) = \int_{-\infty}^{x_0} f(x)\, dx$$

De lo anterior se deduce que cualquier función es función de densidad si y sólo si verifica:

1. $f(x) \geq 0 \quad \forall x \in \mathbb{R}$
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1.$

Cualquier función que verifique estas dos propiedades es una función de densidad. La función de densidad se interpreta como el histograma. Sus valores más altos corresponden a las zonas más probables y viceversa. Por ejemplo, la densidad de la variable $X =$ "Capacidad en ml. de una jeringa producida por la empresa Clinic" de la Figura 1 indica que lo más probable es que la capacidad de una jeringa esté en el intervalo $[4, 6]$. Con menos probabilidad la capacidad de la jeringa estará en los intervalos $[2, 4]$ y $[6, 8]$ y será prácticamente imposible que la capacidad supere los 8 ml. o que sea menor de 2 ml.
Del mismo modo que el histograma representa frecuencias mediante áreas, análogamente, la función de densidad expresa probabilidades por áreas. La probabilidad de que una variable X sea menor que un determinado valor x_0 se obtiene calculando el área de la función de densidad hasta el punto x_0, es decir,

$$P(X \leq x_0) = \int_{-\infty}^{x_0} f(x)\, dx,$$

y análogamente, la probabilidad de que la variable tome un valor entre x_0 y x_1 es,

$$P(x_0 \leq x \leq x_1) = \int_{x_0}^{x_1} f(x)\, dx.$$

Es erróneo entender la función de densidad como la probabilidad de que la variable tome un valor específico, pues esta siempre es cero para cualquier variable continua ya que el área que queda encima de un punto es siempre cero. Por ejemplo, la probabilidad de que la capacidad de una jeringa producida por la empresa Clinic sea exactamente un 5.2 ml. es cero. Sin embargo, la probabilidad de que la capacidad de una jeringa esté en el intervalo $[5.1, 5.3]$, es el área encerrada por la función de densidad en ese intervalo. De esto deducimos que, para variables continuas,

$$P(x_0 < x < x_1) = P(x_0 \leq x \leq x_1) = P(x_0 < x \leq x_1) = P(x_0 \leq x < x_1).$$

Ejemplo 1: Se ha comprobado que el tiempo de vida (en años) de cierto tipo de marcapasos es una variable continua con función de densidad:

$$f(t) = \begin{cases} \frac{1}{16} e^{-t/16}, & \text{si } t > 0, \\ 0, & \text{en otro caso.} \end{cases}$$

¿Cuál es la probabilidad de que a una persona a la que se le ha implantado este marcapasos se le deba reimplantar otro antes de 20 años?
Sea T= tiempo de vida del marcapasos implantado a la persona. La función de densidad $f(t)$ aparece representada en la Figura 2 (izquierda). La probabilidad de que a una persona a la que se le ha implantado este marcapasos se le deba reimplantar otro antes de 20 años se calcula como $P(T \leq 20)$. Debemos calcular el área encerrada por la función hasta $t = 20$ (Figura 2 derecha). Es decir,

$$P(T \leq 20) = \int_{-\infty}^{20} f(t)dt = \int_{0}^{20} \frac{1}{16} e^{-t/16} dt = 1 - e^{-20/16} = 0.71349.$$

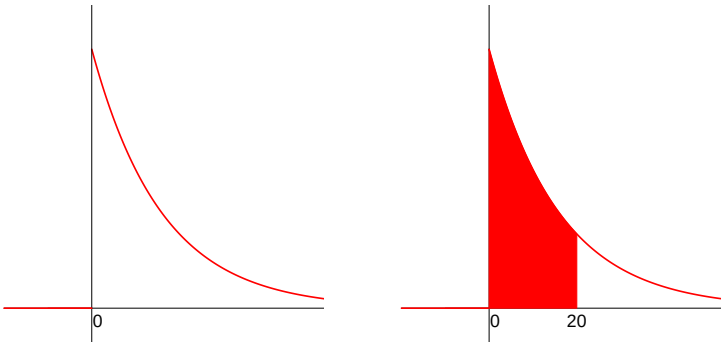


Figure 2: A la izquierda, función de densidad $f(t)$ del Ejemplo 1. A la derecha, el área en rojo representa $P(T \leq 20)$.

Calcularemos para variables aleatorias continuas su función de densidad y su función de distribución

La **función de distribución** para una variable aleatoria continua se define como en el caso discreto por,

$$F(x_0) = P(X \leq x_0),$$

y por tanto,

$$F(x_0) = P(X \leq x_0) = \int_{-\infty}^{x_0} f(x)\, dx,$$

La función de distribución de una variable continua es también no decreciente y verifica que,

$$\begin{aligned} F(-\infty) &= 0, \\ F(+\infty) &= 1. \end{aligned}$$

Además, podemos obtener la función de densidad a partir de la de distribución calculando su derivada:

$$f(x) = F'(x).$$

3 Medidas características de una variable aleatoria continua

Los conceptos que permiten resumir una distribución de frecuencias utilizando valores numéricos pueden utilizarse también para describir la distribución de probabilidad de una variable aleatoria.

3.1 Media o esperanza

Se define la **media poblacional** o **esperanza** de una variable aleatoria continua como,

$$\mu = \mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x)\, dx.$$

Ejemplo 2: ¿Cuál es la vida media de un marcapasos del tipo descrito en el Ejemplo 1?
Recuerda que T= tiempo de vida de un marcapasos. Aplicando la definición anterior,

$$\mu = \mathbb{E}(T) = \int_{-\infty}^{\infty} t f(t)\, dt = \int_{0}^{+\infty} \frac{t}{16} e^{-t/16} dt = 16.$$

Es decir, la vida media de un marcapasos del tipo descrito en el Ejemplo 1 es 16 años. Para hacer la integral hemos aplicado integración por partes.

La interpretación de la media o esperanza es el valor esperado al realizar el experimento con la variable aleatoria. Además, la media puede verse también como el valor central de la distribución de probabilidad.

3.2 Varianza

Se define la **varianza** de una variable aleatoria como

$$\sigma^2 = \text{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx.$$

Ejemplo 3: ¿Cuál es la varianza del tiempo de vida de un marcapasos del Ejemplo 1?
Aplicando la definición de varianza, y teniendo en cuenta que hemos calculado anteriormente que $\mu = 16$, se tiene

$$\sigma^2 = \text{Var}(T) = \int_{-\infty}^{\infty} (t-16)^2 f(t)\, dt = \int_{0}^{\infty} (t-16)^2 \frac{1}{16} e^{-t/16} dt = 256.$$

De nuevo hemos utilizado integración por partes.

La interpretación de la varianza es la misma que para un conjunto de datos: es un valor no negativo que expresa la dispersión de la distribución alrededor de la media. Además, se puede calcular la **desviación típica poblacional** σ como la raíz cuadrada de la varianza. Los valores pequeños de σ indican concentración de la distribución alrededor de la esperanza y valores grandes corresponden a distribuciones más dispersas.

4 Principales modelos de distribuciones continuas: La distribución normal

La distribución normal es la más importante y de mayor uso de todas las distribuciones continuas de probabilidad. Por múltiples razones se viene considerando la más idónea para modelizar una gran diversidad de mediciones de la Medicina, Física, Química o Biología.
La normal es una familia de variables que depende de dos parámetros, la media y la varianza. Dado que todas están relacionadas entre si mediante una transformación muy sencilla, empezaremos estudiando la denominada **normal estándar** para luego definir la familia completa.

4.1 La distribución normal estándar N(0,1)

Definición 2. *Una variable aleatoria continua Z se dice que tiene* **distribución normal estándar**, *y lo denotamos $Z \in N(0,1)$, si su función de densidad viene dada por:*

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \qquad si \;\; z \in \mathbb{R}$$

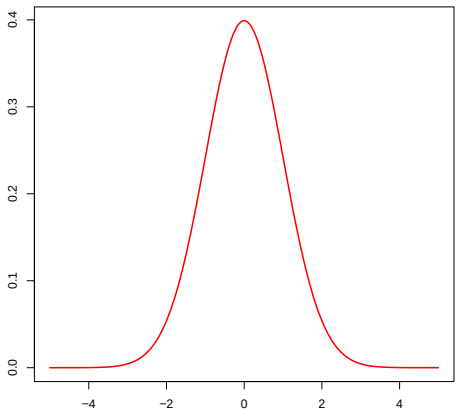


Figure 3: Función de densidad $f(z)$ para $Z \in N(0,1)$.

Propiedades: (Ver Figura 3)

1. $Z \in N(0,1)$ toma valores en toda la recta real. $(f(z) > 0 \quad \forall z \in \mathbb{R})$
2. f es simétrica en torno a cero. $(Si \; Z \in N(0,1) \; entonces \; -Z \in N(0,1))$
3. Si $Z \in N(0,1)$ entonces $\mathbb{E}(Z) = 0$ y $\sigma = 1$.

Supongamos que $Z \in N(0,1)$ y que queremos calcular $P(Z \leq z_0)$. Debemos de tener en cuenta que:

- La probabilidad inducida vendrá dada por el área bajo la densidad.
- Como no existe una expresión explícita para el área existen tablas con algunas probabilidades ya calculadas.
- Las tablas que nosotros utilizaremos proporcionan el valor de la función de distribución, $\Phi(z) = P(Z \leq z)$, de la normal estándar para valores positivos de z, donde z está aproximado hasta el segundo decimal.

Ejemplo 4: Supongamos que $Z \in N(0,1)$. ¿Cómo calcularías $P(Z \leq 1.03)$?

$$P(Z \leq 1.03) = \int_{-\infty}^{1.03} f(z)dz = \int_{-\infty}^{1.03} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$$

Para calcular $P(Z \leq 1.03)$, en el eje de las x marcamos el valor de Z (en este caso $z = 1.03$) e indicamos la probabilidad como el área que queda debajo de la campana de Gauss. (ver Figura 4).
Buscaremos $P(Z \leq 1.03)$ en la tabla en el cruce entre la fila correspondiente a 1.0 y la columna correspondiente a 0.03. Así obtenemos $P(Z \leq 1.03) = 0.8465$.

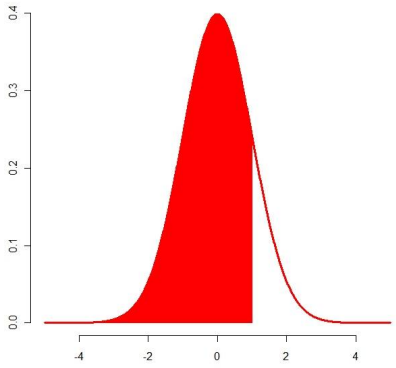


Figure 4: En rojo $P(Z \leq 1.03)$, para $Z \in N(0,1)$.

Ejercicio 1: Supongamos que $Z \in N(0,1)$. Calcula usando las tablas de la normal estándar:

- $P(Z \leq 1.64)$.
- $P(Z > 1)$.
- $P(Z \leq -0.53)$.
- $P(Z > -1.23)$.
- $P(-1.96 \leq Z \leq 1.96)$.
- $P(-1 \leq Z \leq 2)$.
- ¿Cuánto vale aproximadamente $P(Z > 4.2)$?

Ejercicio 2: Sea Z una variable aleatoria con distribución N(0,1). Halla los valores z_0 tales que

- $P(Z \leq z_0) = 0.87$.
- $P(Z > z_0) = 0.05$.
- $P(Z > z_0) = 0.975$.
- $P(|Z| > z_0) = 0.01$.

4.2 La distribución normal N(μ,σ)

Efectuando un cambio de localización y escala sobre la normal estándar, podemos obtener una distribución con la misma forma pero con la media y desviación típica que queramos.

Definición 3. *Si $Z \in N(0,1)$ entonces*

$$X = \mu + \sigma Z \in N(\mu, \sigma)$$

y diremos que X tiene **distribución normal de media μ y desviación típica σ**.

Así, la función de densidad de X tendrá la misma forma de campana, será simétrica en torno a la media μ. La función de densidad de una $N(\mu, \sigma)$ (ver Figura 5) es

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathbb{R}.$$

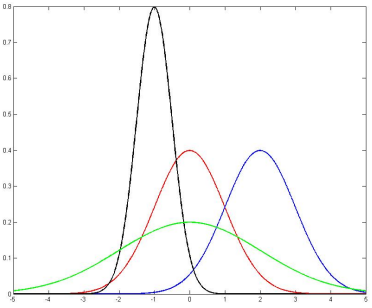


Figure 5: Funciones de densidad de variables normales con distintas medias y varianzas. En rojo densidad de una $N(0,1)$.

Podemos responder a cualquier pregunta sobre probabilidades de una distribución normal estandarizando y luego utilizando la tabla normal estándar. Para estandarizar un valor, réstale la media de la distribución y luego divídelo por la desviación típica.

En la práctica sólo disponemos de la tabla de la distribución normal estándar. Para efectuar cálculos sobre cualquier distribución normal hacemos la transformación inversa, esto es, le restamos la media y dividimos por la desviación típica. A este proceso le llamamos **estandarización** de una variable aleatoria.

$$\text{Si } X \in N(\mu, \sigma) \text{ entonces } Z = \frac{X - \mu}{\sigma} \in N(0,1).$$

Debemos observar que la estandarización se puede aplicar a cualquier variable aleatoria, tenga o no distribución normal. Al estandarizar una variable aleatoria, obtendremos otra (variable estandarizada) con media cero y desviación típica uno.

Ejemplo 5: Supongamos que $X \in N(5,2)$. ¿Cómo calcularías $P(X \leq 1)$?

$$P(X \leq 1) = P\left(\frac{X-5}{2} \leq \frac{1-5}{2}\right) = P\left(Z \leq -2\right)$$

donde $Z = \frac{X-5}{2} \in N(0,1)$. Entonces, consultando las tablas de la normal estándar, obtenemos que

$$P(X \leq 1) = P\left(Z \leq -2\right) = 0.02275.$$

Bioestadística. Curso 2014-2015
Capítulo 5

Carmen Mª Cadarso, Mª del Carmen Carollo, Xosé Luis Otero, Beatriz Pateiro

Contents

1 Introducción

Como ya hemos comentado en otras ocasiones, nuestro objetivo es el estudio de una población y sus características. Llamaremos **parámetro** a una característica numérica que nos interese conocer de la población. Por ejemplo, podrían ser parámetros de interés la presión sistólica media de una población, su nivel de colesterol medio o la proporción de pacientes que responden satisfactoriamente a un medicamento para la diabetes. Generalmente es difícil obtener la información de toda la población y por eso en la práctica contaremos con una muestra representativa de dicha población. En el capítulo 1 hemos estudiado conceptos básicos de Estadística Descriptiva, que nos proporcionaban herramientas para resumir, ordenar y extraer los aspectos más relevantes de la información de la muestra. En el capítulo 2 hemos fijado las bases para trabajar con incertidumbres o probabilidades. Ahora, tras estudiar los principales modelos de variables aleatorias en los capítulos 3 y 4, podremos empezar a hacer inferencia sobre la población de interés basándonos en lo que observamos en una muestra. No nos conformaremos con describir unos datos contenidos en una muestra sino que pretendemos extraer conclusiones para la población de la que fueron extraídos. A esta última tarea la llamamos **Inferencia Estadística**. Dependiendo de los objetivos, podremos clasificar las labores de inferencia en dos grandes categorías: la primera, en la que el interés se centra en estimar o aproximar el valor de un parámetro (por ejemplo, la proporción de pacientes que responden a un determinado medicamento para la diabetes) y la segunda, en la que el interés se centra en contrastar posibles valores de un parámetro (por ejemplo, determinar si el nivel de colesterol medio en hombres es superior al nivel de colesterol medio en mujeres).

2 Conceptos básicos.

Veamos algunas definiciones básicas en Inferencia Estadística. Algunas de ellas ya las hemos introducido en los temas anteriores.

Población. Es el conjunto homogéneo de individuos sobre los que se estudian una o varias características observables. Por ejemplo, la población de un país de la cual nos interesa la proporción de vacunados de gripe A.
Muestra. Es un subconjunto extraído de la población, al que podemos observar. Múltiples razones nos imposibilitan observar toda la población. Por ese motivo, extraemos una muestra y con ella obtenemos información sobre toda la población.
Tamaño de la población o de la muestra. Es el número de individuos que los forman, en cada caso.

Debemos hacer una primera distinción, al hablar de Inferencia, según la naturaleza del problema que se plantee:

1. **Inferencia paramétrica:** cuando se conoce la forma de la distribución de probabilidad e interesa averiguar el parámetro o parámetros de los que depende. Por ejemplo, sabemos que el nivel de colesterol en hombres es Normal e interesa conocer la media μ y la desviación típica σ. A su vez, dentro de la Inferencia Paramétrica vamos a distinguir distintos problemas:

   (a) **Estimación Puntual**. Consiste en aventurar un valor, calculado a partir de la muestra, que esté lo más próximo posible al verdadero parámetro. Por ejemplo, la media muestral puede ser un estimador razonable de la media poblacional y la proporción muestral de la proporción poblacional.

(b) **Intervalos de Confianza**. Dado que la estimación puntual conlleva un cierto error, construímos un intervalo que con alta probabilidad contenga al parámetro. La amplitud del intervalo nos da idea del margen de error de nuestra estimación.

(c) **Contrastes de Hipótesis**. Se trata de responder a preguntas muy concretas sobre la población, y se reducen a un problema de decisión sobre la veracidad de ciertas hipótesis. Por ejemplo, nos podemos preguntar si la proporción de vacunados de gripe A en una población supera el 90%, lo cual limitaría de manera considerable el riesgo epidémico en la población.

2. **Inferencia no Paramétrica:** cuando no se sabe la forma de la distribución poblacional. También se pueden plantear las tareas de estimación, intervalos de confianza y contrastes de hipótesis, aunque las técnicas estadísticas son diferentes.

3 Planteamiento general del problema de inferencia paramétrica

Como ya hemos comentado en el capítulo 3, una medida o característica observada en un individuo se denomina variable aleatoria. Por ejemplo, una variable aleatoria sería el nivel de colesterol. El valor de la variable cambia de individuo a individuo. Otros ejemplos sería la presencia o ausencia de determinada enfermedad, la presión sistólica, etc.
Generalmente se asume que la distribución de la variable de interés X pertenece a cierta familia de distribuciones como por ejemplo la binomial o la normal. Esta familia depende de uno o varios parámetros como la probabilidad de éxito p en el caso de la binomial, la media μ y la varianza σ^2 en el caso de la normal, etc. Usualmente es imposible o muy costoso obtener los valores de la variable de interés sobre todos los individuos de la población para poder determinar así el parámetro que determina la distribución. En la práctica solo contamos con una muestra representativa y tendremos que estimar los parámetros de la población en base a valores aproximados a partir de la muestra.

- Una **muestra aleatoria simple** de tamaño n está formada por n variables $X_1, X_2, \cdots, X_n$ independientes y con la misma distribución que X.
- Llamamos **realización muestral** a los valores concretos que tomaron las n variables aleatorias después de la obtención de la muestra.
- Un **estadístico** es una función de la muestra aleatoria, y por tanto nace como resultado de cualquier operación efectuada sobre la muestra. Es también una variable aleatoria y por ello tendrá una cierta distribución, que se denomina **distribución del estadístico en el muestreo**.
- Para resolver el problema de estimación puntual, esto es, para aventurar un valor del parámetro poblacional desconocido, escogemos el valor que ha tomado un estadístico calculado sobre nuestra realización muestral. Al estadístico escogido para tal fin le llamamos **estimador** del parámetro. Al valor obtenido con una realización muestral concreta se le llama **estimación**.

El problema radica, por lo tanto, en elegir un "buen estimador", es decir, una función de la muestra con buenas propiedades.
En general, un buen estimador de un parámetro poblacional (media, proporción de individuos que presentan cierta característica,...) va a ser el correspondiente parámetro muestral (media de la muestra, proporción de individuos que presentan la característica en la muestra,...).

4 Teorema Central del Límite

El siguiente resultado nos permitirá calcular la distribución en el muestreo de muchos estadísticos de interés. El denominado Teorema Central del Límite que afirma que, si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes y con la misma distribución X, donde X tiene media μ y varianza σ^2, entonces para n grande, la variable

$$\frac{X_1 + X_2 + \ldots + X_n}{n}$$

es aproximadamente normal con media μ y varianza σ^2/n. Formalmente:

Teorema 1 (Teorema central del límite)**.** *Sea $X_1, X_2, \ldots, X_n, \ldots$ una sucesión de variables aleatorias independientes y con la misma distribución, con media μ y varianza σ^2 todas ellas.*
Sea $S_n = X_1 + \ldots + X_n$. Entonces,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1).$$

Equivalentemente,

$$\frac{X_1 + X_2 + \ldots + X_n}{n} \xrightarrow{d} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

5 Distribuciones asociadas con la normal

Además del modelo normal, existen otros modelos que desempeñan un papel importante en la inferencia estadística. Entre ellos se encuentran las distribuciones χ^2 y t de Student.

5.1 La distribución χ^2

La distribución Chi-cuadrado (o ji-cuadrado) con n grados de libertad χ^2_n es un modelo de variable aleatoria continua. En la Figura 1 se representa la función de densidad de variables χ^2 para diferentes grados de libertad.

Propiedades:

1. La variable Chi-cuadrado toma valores $[0, +\infty)$.
2. La distribución Chi-cuadrado es asimétrica.

5.2 La distribución t de Student

La distribución t de Student con k grados de libertad es un modelo de variable aleatoria continua. En la Figura 2 se representa la función de densidad de variables t de Student para diferentes grados de libertad junto con la densidad de una N(0,1).

Propiedades:

1. La variable t de Student toma valores en toda la recta real.
2. La distribución t de Student es simétrica en torno al origen.
3. $t_k \xrightarrow{d} N(0,1)$ cuando $k \to \infty$.

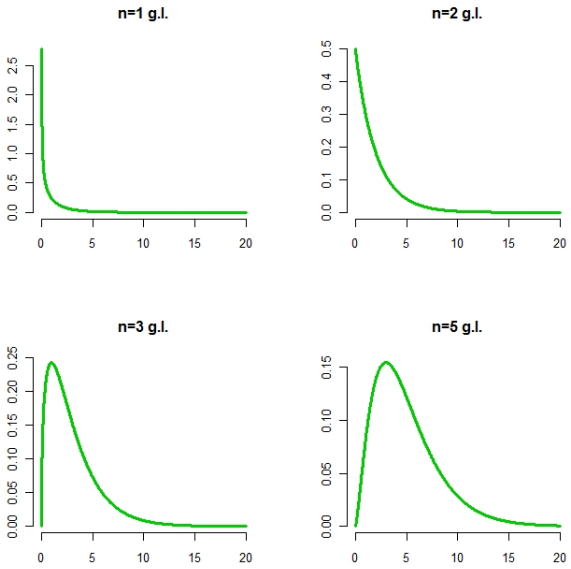


Figure 1: En verde densidades de variables χ^2_n para distintos valores de n.

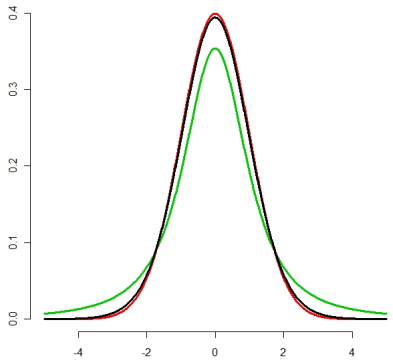


Figure 2: En verde densidad de una t de Student con 2 grados de libertad, en rojo densidad de una N(0,1) y en negro densidad de una t de Student con 20 grados de libertad

Al igual que ocurría con la distribución normal, calcularemos probabilidades y cuantiles de estas distribuciones a través de tablas.

Bioestadística. Curso 2014-2015
Capítulo 6

Carmen Mª Cadarso, Mª del Carmen Carollo, Xosé Luis Otero, Beatriz Pateiro

Contents

1 Introducción

En el capítulo anterior hemos presentado los conceptos básicos de la inferencia estadística. Además, hemos clasificado las labores de inferencia en dos grandes categorías: la estimación, que se centra en estimar o aproximar el valor de un parámetro desconocido y el contraste de hipótesis, que se centra en decidir sobre la veracidad de ciertas hipótesis acerca de los valores del parámetro desconocido. En este capítulo profundizaremos en los problemas de estimación, tanto en la estimación puntual como en la construcción de intervalos de confianza. Para todas estas labores será fundamental conocer los estadísticos adecuados para cada parámetro y sus distribuciones.

2 Estimación puntual

Como comentamos en el capítulo anterior, la estimación puntual de un parámetro desconocido θ consiste en aproximar su valor a partir de una muestra. Para resolver el problema de estimación puntual escogemos el valor que ha tomado un estadístico $\hat{\theta}$ calculado sobre nuestra realización muestral. Recordamos que un estadístico es una variable aleatoria y por ello tendrá una cierta distribución.

Definición 1. *Diremos que un estimador $\hat{\theta}$ para un parámetro poblacional θ es* **insesgado** *si*

$$\mathbb{E}\left(\hat{\theta}\right) = \theta.$$

Que un estimador sea insesgado es una buena propiedad. También nos interesará que la dispersión del estimador sea pequeña y que disminuya al aumentar el tamaño muestral.

2.1 Estimación puntual de una proporción

El primer problema práctico de inferencia que vamos a afrontar consiste en obtener información sobre la proporción de individuos con cierta característica en una población, mediante la extracción de una muestra aleatoria simple. Consideramos la variable

$$X = \begin{cases} 1 & \text{, si el individuo presenta la característica de interés,} \\ 0 & \text{, si el individuo no presenta la característica de interés.} \end{cases}$$

Así definida X tiene distribución Bernoulli(p), donde el parámetro p es precisamente la proporción (que desconocemos) de individuos con la característica en la población.

La muestra está formada por n variables $X_1, \ldots, X_n$ independientes y con la misma distribución que X. El estimador razonable para p es la proporción muestral

$$\hat{p} = \frac{\text{número de individuos con la característica en la muestra}}{n} = \frac{X_1 + \cdots + X_n}{n}.$$

Observamos en primer lugar que $\mathbb{E}(\hat{p}) = p$ y, por lo tanto, $\hat{p}$ es insesgado. Ahora que sabemos que $\hat{p}$ está centrado en torno a p, nos interesa que su dispersión sea pequeña. En nuestro caso

$$\text{Var}\left(\hat{p}\right) = \frac{p(1-p)}{n}$$

y $\lim_{n\to\infty} \text{Var}(\hat{p}) = 0$. Esto significa que al aumentar el tamaño muestral el estimador se aproxima al parámetro poblacional, lo cual también es deseable.

Distribución del estadístico proporción muestral $\hat{p}$. Hemos definido el estadístico

$$\hat{p} = \frac{X_1 + \cdots + X_n}{n}.$$

Además $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes y con la misma distribución X, donde X tiene media p y varianza $p(1-p)$. Entonces, por el Teorema Central del Límite, para n grande, $\hat{p}$ es aproximadamente normal con media p y varianza $p(1-p)/n$. Estandarizando,

$$\frac{\hat{p} - p}{\sqrt{\frac{p\,(1-p)}{n}}} \sim N(0,1).$$

2.2 Estimación puntual de la media y la varianza.

Consideramos ahora el problema de inferencia en una población normal. En esta situación disponemos de una muestra aleatoria simple $X_1, \ldots, X_n$ formada por n variables aleatorias independientes y con la misma distribución $N(\mu, \sigma)$. El problema de inferencia consiste en averiguar los parámetros μ (media poblacional) y σ (desviación típica poblacional).

2.2.1 Estimación puntual de la media

Como estimador natural para la media de la población, μ, proponemos la media muestral:

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}.$$

Se cumple que:

- La media de $\bar{X}$ es $\mathbb{E}(\bar{X}) = \mu$.
- La varianza de $\bar{X}$ es $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$.

De esto se deduce que la media muestral es un estimador insesgado de la media poblacional y que su varianza es la poblacional dividida por n. Por tanto, la dispersión decrece tendiendo a cero cuando el tamaño muestral aumenta.

Distribución del estadístico media muestral $\bar{X}$. Por la propiedad de aditividad de la distribución normal y dado que $\bar{X}$ es la suma de n variables independientes, entonces la media muestral tiene distribución normal

$$\bar{X} \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

2.2.2 Estimación puntual de la varianza

Estimaremos la varianza de la población σ^2 por medio de la varianza muestral

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2.$$

En muchos textos S^2 se denomina cuasivarianza muestral

Se puede comprobar que así definido S^2 es un estimador insesgado de la varianza σ^2 de la población.

3 Intervalos de confianza

La estimación puntual resulta incompleta en el siguiente sentido: ¿qué seguridad tenemos de que un estadístico se aproxime al verdadero valor del parámetro? Para poder dar respuesta a esta cuestión construimos intervalos de confianza, que permiten precisar la incertidumbre existente en la estimación.

Definición 2. *Un* **intervalo de confianza** *es un intervalo construido en base a la muestra y, por tanto, aleatorio, que contiene al parámetro con una cierta probabilidad, conocida como* **nivel de confianza**. *Sea θ el parámetro desconocido y L_1 y L_2 los extremos del intervalo (que son estadísticos ya que se definen en base a la muestra). Se dice que $[L_1, L_2]$ tiene un nivel de confianza $1-\alpha$, siendo $\alpha \in [0,1]$, si $P(L_1 \leq \theta \leq L_2) \geq 1-\alpha$.*

El método que usaremos para construir intervalos de confianza se denomina método pivotal

El nivel de confianza con frecuencia se expresa en porcentaje. Así, un intervalo de confianza del 95% es un intervalo de extremos aleatorios que contiene al parámetro desconocido con una probabilidad de 0.95.

3.1 Intervalo de confianza para la media de una población normal

Consideramos ahora el problema de construcción de un intervalo de confianza para la media μ en una población normal. En esta situación disponemos de una muestra aleatoria simple $X_1, \ldots, X_n$ formada por n variables aleatorias independientes y con la misma distribución $N(\mu, \sigma)$.

3.1.1 Intervalo de confianza para la media con varianza conocida

Supongamos que queremos construir un intervalo de confianza para la media μ y que conocemos la varianza de la población σ^2. La distribución de la media muestral permite obtener como pivote

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \in N(0,1).$$

Este estadístico (pivote) nos servirá para construir un intervalo de confianza con nivel de confianza $1-\alpha$ para la media μ cuando la varianza σ^2 es conocida. Sea $z_{\alpha/2}$ el valor tal que $P(Z > z_{\alpha/2}) = \alpha/2$, siendo $Z \in N(0,1)$ (ver Figura 1). Entonces:

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha.$$

Equivalentemente,

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha.$$

Así, el intervalo de confianza con nivel de confianza $1-\alpha$ para la media μ cuando la varianza σ^2 es conocida será:

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

3.1.2 Intervalo de confianza para la media con varianza desconocida

En la práctica no es habitual conocer la varianza de la variable de interés. Supongamos que queremos construir un intervalo de confianza para la media μ y que desconocemos la varianza de la población.

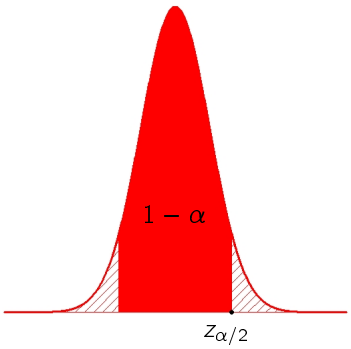


Figure 1: Denotamos $z_{\alpha/2}$ el número real tal que $P(Z > z_{\alpha/2}) = \alpha/2$, siendo $Z \in N(0,1)$.

Usaremos como estadístico (pivote) en este caso

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

Recuerda que

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Se cumple que:

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \in t_{n-1},$$

es decir, la distribución del estadístico es una t de Student con $n-1$ grados de libertad. Este estadístico (pivote) nos servirá para construir un intervalo de confianza con nivel de confianza $1-\alpha$ para la media μ cuando la varianza σ^2 es desconocida. Sea $t_{\alpha/2}$ el valor tal que $P(T > t_{\alpha/2}) = \alpha/2$, donde T es una variable t de Student con $n-1$ grados de libertad (ver Figura 2). Entonces:

$$P\left(-t_{\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_{\alpha/2}\right) = 1 - \alpha.$$

Equivalentemente,

$$P\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Así, el intervalo de confianza con nivel de confianza $1-\alpha$ para la media μ cuando la varianza σ^2 es desconocida será:

$$\left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}\ ,\ \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right).$$

Ejercicio 1: En un estudio sobre trastornos del sueño se evaluó el número de horas de sueño de 8 individuos seleccionados al azar. Los resultados se muestran a continuación.

6.9, 7.6, 6.5, 6.2, 7.8, 7.0, 5.5, 7.6.

A partir de esta muestra, estima la media y la desviación típica del número de horas de sueño de la población. Suponiendo normalidad, determina un intervalo de confianza para el número medio de horas de sueño con una confianza del 95%.

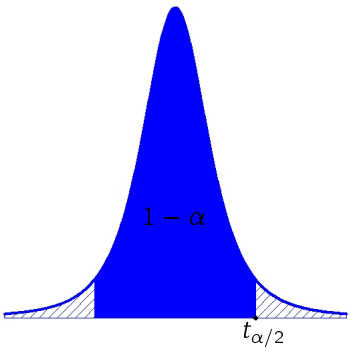


Figure 2: Denotamos $t_{\alpha/2}$ el número real tal que $P(T_k > t_{\alpha/2}) = \alpha/2$, siendo T una variable t de Student (con los grados de libertad correspondientes).

3.2 Intervalo de confianza para la diferencia de medias de poblaciones normales

El objetivo de este apartado es construir intervalos de confianza no sólo para la media de una característica de la población sino también para la comparación de dos poblaciones mediante su media.

3.2.1 Intervalo de confianza para $\mu_1 - \mu_2$. Muestras independientes, varianzas conocidas

Supongamos dos poblaciones en las que se estudian dos características distribuidas normalmente con medias desconocidas μ_1 y μ_2, respectivamente. Estamos interesados en construir un intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ a partir de dos muestras:

En muestras independientes, los individuos donde se han obtenido las mediciones de la población 1 son distintos de los individuos donde se han obtenido las mediciones de la población 2

- Una muestra formada por n_1 variables independientes y con la misma distribución $N(\mu_1, \sigma_1)$.
- Una muestra formada por n_2 variables independientes y con la misma distribución $N(\mu_2, \sigma_2)$.

Suponemos que las muestras son independientes, es decir, los individuos donde se han obtenido las mediciones de la población 1 son distintos de los individuos donde se han obtenido las mediciones de la población 2. Suponemos además que las varianzas σ_1^2 y σ_2^2 son conocidas. Entonces, utilizaremos como estadístico

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \in N(0,1).$$

El intervalo de confianza de nivel $1 - \alpha$ para la diferencia de medias $\mu_1 - \mu_2$ será:

$$\left((\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \ , \ (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right).$$

3.2.2 Intervalo de confianza para $\mu_1 - \mu_2$. Muestras independientes, varianzas desconocidas e iguales.

En muchas aplicaciones los valores de σ_1^2 y σ_2^2 son desconocidos y por lo tanto es necesario estimarlos. No obstante, puede suceder que pese a ser desconocidas podamos suponer que ambas varianzas son iguales. Consideremos entonces dos poblaciones en las que se estudian dos características distribuidas normalmente con medias desconocidas μ_1 y μ_2, respectivamente. Estamos interesados en construir un intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ a partir de dos muestras:

- Una muestra formada por n_1 variables independientes y con la misma distribución $N(\mu_1, \sigma_1)$.
- Una muestra formada por n_2 variables independientes y con la misma distribución $N(\mu_2, \sigma_2)$.

Suponemos que las muestras son independientes y que las varianzas σ_1^2 y σ_2^2 son desconocidas pero iguales. Entonces, utilizaremos como estadístico

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \in t_{n_1+n_2-2}.$$

En el estadístico anterior,

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

representa el estimador adecuado para la varianza de las dos poblaciones. El intervalo de confianza de nivel $1 - \alpha$ para la diferencia de medias $\mu_1 - \mu_2$ será entonces:

$$\left((\bar{X}_1 - \bar{X}_2) - t_{\alpha/2}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \ , \ (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \right).$$

El valor $t_{\alpha/2}$ se obtiene de una distribución t de Student con $n_1 + n_2 - 2$ grados de libertad.

Ejercicio 2: El Verapamil y el Nitroprusside son dos productos utilizados para reducir la hipertensión. Para compararlos, unos pacientes son tratados con Verapamil y otros con Nitroprusside. Los resultados obtenidos se muestran en la siguiente tabla, donde:

- X_1 =reducción (en mmHg) de la presión arterial de un paciente con Verapamil.
- X_2 =reducción (en mmHg) de la presión arterial de un paciente con Nitroprusside.

X_1	10	15	18	23	12	16	
X_2	15	10	19	9	14	12	18

Admitiendo normalidad y sabiendo que ambas variables tienen la misma desviación típica, construye un intervalo de confianza de nivel 95% para la diferencia de medias de la reducción de presión arterial.

3.2.3 Intervalo de confianza para la diferencia de medias. Muestras apareadas

En ocasiones nos interesará comparar dos métodos o tratamientos. En ese caso es natural que los individuos donde se aplican los tratamientos sean los mismos. Consideremos el siguiente ejemplo.

Ejemplo 1: Se quiere estudiar los efectos del abandono de la bebida sobre la presión sistólica en individuos alcohólicos. Para ello se mide la presión sistólica en 10 individuos alcohólicos antes y después de 2 meses de haber dejado al bebida.

Sujeto	1	2	3	4	5	6	7	8	9	10
X_1 presión antes	140	165	160	160	175	190	170	175	155	160
X_2 presión después	145	150	150	160	170	175	160	165	145	170

En muestras apareadas, se evalúan características diferentes en los mismos individuos

Cuando X_1 y X_2 representan características diferentes de la misma población y se quieren evaluar sus diferencias, conviene tomar muestras apareadas. Así, se obtiene el valor de las características X_1 y X_2 sobre los mismos individuos de la población. Se supone que las muestras se han obtenido de poblaciones normales $X_1 \in N(\mu_1, \sigma_1)$ y $X_2 \in N(\mu_2, \sigma_2)$ pero teniendo en cuenta que ahora X_1 y X_2 no son independientes. En esta situación consideraremos la variable $D = X_1 - X_2$ que sigue una distribución normal con media $\mu_D = \mathbb{E}(X_1 - X_2) = \mu_1 - \mu_2$ y varianza $\sigma_D^2 = \text{Var}(X_1 - X_2)$.
En la práctica tendremos dos muestras de tamaño n observadas en los mismos individuos, es decir,

- $X_{11}, \ldots, X_{1n} \in N(\mu_1, \sigma_1)$.
- $X_{21}, \ldots, X_{2n} \in N(\mu_2, \sigma_2)$.

Construimos la muestra $D_1 = X_{11} - X_{21}, \ldots, D_n = X_{1n} - X_{2n}$ y estimaremos

- μ_D mediante $\bar{D}$.
- σ_D^2 mediante S_D^2.

Como estadístico pivote utilizaremos

$$\frac{\bar{D} - (\mu_1 - \mu_2)}{S_D/\sqrt{n}} \in t_{n-1}.$$

El intervalo de confianza de nivel $1-\alpha$ para la diferencia de medias $\mu_1 - \mu_2$ será entonces:

$$\left(\bar{D} - t_{\alpha/2}\frac{S_D}{\sqrt{n}}\ ,\ \bar{D} + t_{\alpha/2}\frac{S_D}{\sqrt{n}}\right).$$

El valor $t_{\alpha/2}$ se obtiene en este caso de una distribución t de Student con $n-1$ grados de libertad.

Ejemplo 1: Volviendo al ejemplo sobre los efectos del abandono de la bebida sobre la presión sistólica en individuos alcohólicos,

Sujeto	1	2	3	4	5	6	7	8	9	10
X_1 presión antes	140	165	160	160	175	190	170	175	155	160
X_2 presión después	145	150	150	160	170	175	160	165	145	170
Diferencias D_i	-5	15	10	0	5	15	10	10	10	-10

Por lo tanto

$$\bar{D} = \frac{-5 + 15 + \ldots + 10 - 10}{10} = 6.$$

$$S_D^2 = \frac{(-5-6)^2 + \ldots + (-10-6)^2}{9} = 71.111.$$

$$S_D = \sqrt{71.11} = 8.4327.$$

El intervalo de confianza de nivel 95% para la diferencia $\mu_1 - \mu_2$ de la presión sistólica media será entonces:

$$\left(\bar{D} - t_{\alpha/2}\frac{S_D}{\sqrt{n}}\ ,\ \bar{D} + t_{\alpha/2}\frac{S_D}{\sqrt{n}}\right) = \left(6 - 2.26\frac{8.4327}{\sqrt{10}}\ ,\ 6 + 2.26\frac{8.4327}{\sqrt{10}}\right) = (-0.0266, 12.0266).$$

En este caso el valor $t_{\alpha/2}$ se obtiene de una distribución t de Student con $n - 1 = 9$ grados de libertad.

3.3 Intervalo de confianza para la proporción

Construimos ahora un intervalo de confianza para p. Nos basamos en la proporción muestral $\hat{p}$. Recuerda que, si n es grande, la distribución de $\hat{p}$ se puede aproximar por la normal y

$$\frac{\hat{p}-p}{\sqrt{\frac{p\,(1-p)}{n}}} \sim N(0,1).$$

Entonces

$$1-\alpha = P\left(-z_{\alpha/2} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) = P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$$

De la expresión anterior se deduciría un intervalo de confianza para p con nivel de confianza $1-\alpha$, que estaría centrado en $\hat{p}$ y tendría radio $z_{\alpha/2}\sqrt{p\,(1-p)/n}$. Sin embargo, la desviación típica de $\hat{p}$ es $\sqrt{p\,(1-p)/n}$ que, por depender de la proporción poblacional p, es desconocida. Por este motivo, tenemos que tomar $\sqrt{\hat{p}\,(1-\hat{p})/n}$ como estimador de la desviación típica de $\hat{p}$ y usarlo para construir el intervalo de confianza, que será:

$$\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right).$$

Ejercicio 3: Una empresa farmacéutica quiere comercializar un medicamento para cierta dolencia. Para probar si su medicamento es eficaz, lo administra a 100 pacientes, de los cuales 50 presentan mejoría. Construye un intervalo de confianza para la proporción de pacientes de la población que mejoran al tomar el medicamento, con una confianza del 99%.

3.4 Intervalo de confianza para la diferencia de proporciones

En algunas ocasiones estamos interesados en estimar la diferencia de proporciones $p_1 - p_2$ de dos poblaciones. Tenemos así dos muestras:

- Una muestra formada por n_1 variables independientes de la población 1.
- Una muestra formada por n_2 variables independientes de la población 2.

Suponemos que las muestras son independientes (los individuos donde se han obtenido las mediciones de la población 1 son distintos de los individuos donde se han obtenido las mediciones de la población 2). En este caso, para tamaños muestrales grandes,

$$\frac{(\hat{p_1}-\hat{p_2})-(p_1-p_2)}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} \sim N(0,1)$$

y el intervalo de confianza de nivel $1-\alpha$ para la diferencia de proporciones $p_1 - p_2$ será:

$$\left((\hat{p_1}-\hat{p_2}) - z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}\ ,\ (\hat{p_1}-\hat{p_2}) + z_{\alpha/2}\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}\right).$$

4 Resumen de las distribuciones de estadísticos en el muestreo

A continuación presentamos un resumen de los principales estadísticos que hemos visto a lo largo de este capítulo y sus distribuciones en el muestreo. Para los problemas de inferencia sobre la media o diferencia de medias estamos suponiendo que las poblaciones son normales. Las distribuciones de los estadísticos para la proporción o diferencia de proporciones son aproximadas y por lo tanto válidas para tamaños muestrales grandes.

Problema de inferencia	**Estadístico estandarizado**	**Distribución**
Media con varianza conocida	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$	$N(0,1)$
Media con varianza desconocida	$\frac{\bar{X}-\mu}{S/\sqrt{n}}$	t_{n-1}
Diferencia de medias. Muestras independientes con varianzas conocidas	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	$N(0,1)$
Diferencia de medias. Muestras independientes con varianzas desconocidas pero iguales	$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_P^2}{n_1}+\frac{S_P^2}{n_2}}}$	$t_{n_1+n_2-2}$
Diferencia de medias. Muestras apareadas	$\frac{\bar{D}-(\mu_1-\mu_2)}{S_D/\sqrt{n}}$	t_{n-1}
Proporción (muestras grandes)	$\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$	$N(0,1)$
Diferencia de proporciones (muestras grandes)	$\frac{\hat{p}_1-\hat{p}_2-(p_1-p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$	$N(0,1)$

Bioestadística. Curso 2014-2015
Capítulo 7

Carmen Mª Cadarso, Mª del Carmen Carollo, Xosé Luis Otero, Beatriz Pateiro

Contents

1 Introducción

Los procedimientos de inferencia que hemos realizado hasta ahora se resumen en dos: la estimación puntual y los intervalos de confianza. Con la estimación puntual se obtienen valores concretos que sirven de estimaciones de los parámetros poblacionales de interés, por ejemplo, estimamos la media poblacional, μ, con la media muestral, $\bar{x}$. Con los intervalos de confianza se obtienen regiones aleatorias que contienen a los parámetros de interés con cierta probabilidad, por ejemplo, el intervalo de confianza con nivel de confianza $1-\alpha$ para la media μ de una población normal es $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, cuando la desviación σ es conocida. La otra gran tarea de la Inferencia Estadística consiste en responder a preguntas muy concretas sobre la población. Por ejemplo, ¿podemos asumir que el nivel medio de colesterol es 200?, ¿la prevalencia del infarto de miocardio es mayor que 0.03?, ¿el nivel de colesterol promedio es el mismo en varones que en mujeres? Como veremos se plantean en términos de unas hipótesis que debemos aceptar o rechazar. Y esta decisión la tomaremos en base a una realización muestral. Cuando los datos muestrales discrepen mucho de la hipótesis rechazaremos la hipótesis.

2 Planteamiento y resolución de un contraste de hipótesis.

Se tiene una hipótesis de trabajo y una muestra de observaciones, y se trata de decidir si la hipótesis planteada es compatible con lo que se puede aprender del estudio de los valores muestrales, es decir, decidir si la muestra que se obtuvo está de acuerdo con la hipótesis de trabajo.

2.1 Hipótesis nula e hipótesis alternativa

Llamaremos **hipótesis nula**, y la denotamos por H_0, a la que se da por cierta antes de obtener la muestra. Goza de *presunción de inocencia*.

Rechazaremos la hipótesis nula H_0 en favor de H_1 si encontramos pruebas significativas en los datos a favor de H_1

Llamaremos **hipótesis alternativa**, y la denotamos por H_1 a lo que sucede cuando no es cierta la hipótesis nula. Por gozar la hipótesis nula de presunción de inocencia, es en la hipótesis alternativa donde recae la carga de la prueba.

Ejemplo 1: Si nos preguntamos si podemos asumir que el nivel medio de colesterol (μ) es 200, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu = 200 \\ H_1: \ \mu \neq 200 \end{cases}$$

La hipótesis nula $H_0: \ \mu = 200$ sólo será rechazada si existe evidencia en los datos para afirmar que $\mu \neq 200$ (hipótesis alternativa).

Además,

- Una **hipótesis simple** es la que está constituida por un único punto.
- Si la hipótesis consta de más de un punto la llamaremos **hipótesis compuesta**.

Ejemplo 2: Si queremos determinar si la prevalencia del infarto de miocardio (p) es mayor que 0.03, el contraste planteado sería:

$$\begin{cases} H_0: \ p \leq 0.03 \\ H_1: \ p > 0.03 \end{cases}$$

Ahora la hipótesis nula $H_0: \ p \leq 0.03$ es compuesta.

Ejemplo 3: Si nos preguntamos si podemos asumir que el nivel medio de colesterol es el mismo en varones que en mujeres, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu_1 = \mu_2 \\ H_1: \ \mu_1 \neq \mu_2 \end{cases} \text{ o equivalentemente } \begin{cases} H_0: \ \mu_1 - \mu_2 = 0 \\ H_1: \ \mu_1 - \mu_2 \neq 0 \end{cases}$$

siendo μ_1 el nivel medio de colesterol de los hombres y μ_2 el nivel medio de colesterol de las mujeres.

Ejemplo 4: Si queremos determinar si el nivel medio de colesterol es menor en varones que en mujeres, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu_1 \geq \mu_2 \\ H_1: \ \mu_1 < \mu_2 \end{cases} \text{ o equivalentemente } \begin{cases} H_0: \ \mu_1 - \mu_2 \geq 0 \\ H_1: \ \mu_1 - \mu_2 < 0 \end{cases}$$

2.2 Error de tipo I y error de tipo II

Volvemos al problema de decisión que supone el contraste de hipótesis. La siguiente tabla refleja los posibles aciertos o errores en un contraste.

		Decisión	
		No se rechaza H_0	Se rechaza H_0
Realidad	H_0 es verdadera	Decisión correcta	Error tipo I
	H_0 es falsa	Error tipo II	Decisión correcta

Observamos que se puede tomar una decisión correcta o errónea.

- Llamamos **error de tipo I** al que cometemos cuando rechazamos la hipótesis nula, siendo cierta.
- El **error de tipo II** es el que cometemos cuando aceptamos la hipótesis nula, siendo falsa.

2.3 Nivel de significación y potencia de un test

Ya que cualquier decisión tomada al hacer un contraste estará basada sobre información parcial de una población, debemos de tener en cuenta la probabilidad de tomar una decisión incorrecta.

- Probabilidad del error de tipo I: Se denota por α y se denomina **nivel de significación**.

$$\alpha = P(\text{Rechazar } H_0 / H_0 \text{ es cierta})$$

- Probabilidad del error de tipo II: se denota por β.

$$\beta = P(\text{No rechazar } H_0 / H_0 \text{ es falsa})$$

El nivel de significación es la probabilidad rechazar H_0 cuando H_0 es cierta. Son comunes los niveles de significación del 0.05, 0.01 y 0.1

- La probabilidad de detectar que una hipótesis es falsa se denomina **potencia**.

$$\text{Potencia} = P(\text{Rechazar } H_0 / H_0 \text{ es falsa}) = 1 - \beta$$

La potencia la probabilidad rechazar H_0 cuando H_0 es falsa

Debemos adoptar un criterio que, en base a la muestra, nos permita decidir si rechazamos o no la hipótesis nula. Sería deseable hacerlo de tal manera que las probabilidades de ambos tipos de error fueran tan pequeñas como fuera posible. Sin embargo, con una muestra de tamaño fijo, disminuir la probabilidad del error de tipo I, conduce a incrementar la probabilidad del error de tipo II. Piensa que la forma de minimizar la probabilidad del error de tipo I (el nivel de significación) es mediante un criterio que acepte H_0 la mayor parte de las veces. Sin embargo, así se incrementa la probabilidad del error de tipo II, es decir, disminuye la potencia del test. Una forma de proceder ante un problema con dos objetivos como es éste, consiste en fijar el nivel de significación y escoger el criterio que nos proporcione la mayor potencia posible.

2.4 Región crítica de un test

Al fijar un nivel de significación, α, se obtiene implícitamente una división en dos regiones del conjunto de posibles valores del estadístico de contraste:

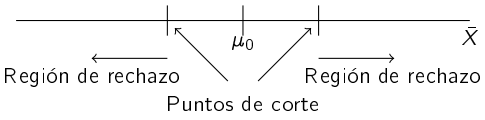
- La **región de rechazo** o región crítica que tiene probabilidad α (bajo H_0).
- La **región de aceptación** que tiene probabilidad $1 - \alpha$ (bajo H_0).

Si el valor del estadístico cae en la región de aceptación, no existen razones suficientes para rechazar la hipótesis nula con un nivel de significación α, y el contraste se dice **estadísticamente no significativo**, es decir no existe evidencia a favor de H_1.
Si el valor del estadístico cae en la región de rechazo, los datos no son compatibles con H_0 y la rechazamos. Entonces se dice que el contraste es **estadísticamente significativo**, es decir existe evidencia estadísticamente significativa a favor de H_1.

Ejemplo 5: **Contraste bilateral.** Si estamos interesados en determinar si la media μ de una variable difiere significativamente de un valor dado μ_0, el contraste planteado sería:

$$\begin{cases} H_0 : \ \mu = \mu_0 \\ H_1 : \ \mu \neq \mu_0 \end{cases}$$

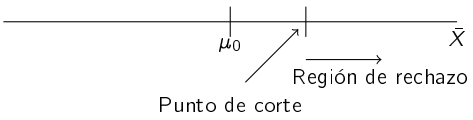
Dados los valores de una muestra, parece claro que deberíamos rechazar H_0 si $\bar{X}$ está "lejos" de μ_0 **en ambas direcciones**. Es decir, tendríamos una región crítica como se muestra a continuación:



Ejemplo 6: **Contraste unilateral por la derecha.** Si estamos interesados en determinar si la media μ de una variable es significativamente mayor que un valor dado μ_0, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu \leq \mu_0 \\ H_1: \ \mu > \mu_0 \end{cases}$$

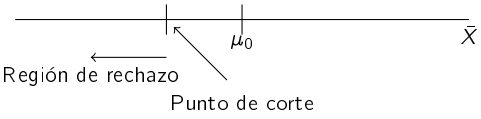
Dados los valores de una muestra, parece claro que deberíamos rechazar H_0 si $\bar{X}$ está "lejos" de μ_0 **en una sola dirección**. Es decir, tendríamos una región crítica como se muestra a continuación:



Ejemplo 7: **Contraste unilateral por la izquierda.** Si estamos interesados en determinar si la media μ de una variable es significativamente menor que un valor dado μ_0, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu \geq \mu_0 \\ H_1: \ \mu < \mu_0 \end{cases}$$

Dados los valores de una muestra, parece claro que deberíamos rechazar H_0 si $\bar{X}$ está "lejos" de μ_0 **en una sola dirección**. Es decir, tendríamos una región crítica como se muestra a continuación:



2.5 El p-valor de un contraste

A medida que el nivel de significación α disminuye es más difícil rechazar la hipótesis nula (manteniendo los mismos datos). Dado un estadístico de contraste, hay un valor de α a partir del cual ya no podemos rechazar H_0. A dicho valor se le se le llama el p-valor del contraste y se denota por p. Si el nivel de significación es menor que p ya no se rechaza H_0. En resumen:

- Si $\alpha < p$ no podemos rechazar H_0 a nivel α.
- Si $\alpha > p$ podemos rechazar H_0 a nivel α.

2.6 Etapas en la resolución de un contraste de hipótesis

Resumiendo, las etapas en la resolución de un contraste de hipótesis son:

1. Especificar las hipótesis nula H_0 y alternativa H_1.
2. Elegir un estadístico de contraste apropiado, para medir la discrepancia entre la hipótesis y la muestra.

3. Fijar el nivel de significación α en base a cómo de importante se considere rechazar H_0 cuando realmente es cierta.
4. Al fijar un nivel de significación, α, se obtiene implícitamente una división en dos regiones del conjunto de posibles valores del estadístico de contraste:
   - La región de rechazo o región crítica que tiene probabilidad α (bajo H_0).
   - La región de no rechazo que tiene probabilidad $1-\alpha$ (bajo H_0).
5. Si el valor del estadístico cae en la región de rechazo, los datos no son compatibles con H_0 y la rechazamos. Entonces se dice que el contraste es estadísticamente significativo, es decir existe evidencia estadísticamente significativa a favor de H_1.
6. Si el valor del estadístico cae en la región de aceptación, no existen razones suficientes para rechazar la hipótesis nula con un nivel de significación α, y el contraste se dice estadísticamente no significativo, es decir no existe evidencia a favor de H_1.

3 Relación con los intervalos de confianza

Consideramos ahora un contraste bilateral. Por ejemplo,

$$\begin{cases} H_0 : \ \mu = \mu_0 \\ H_1 : \ \mu \neq \mu_0 \end{cases}$$

Según hemos comentado anteriormente, una vez que tenemos una muestra deberíamos rechazar H_0 si $\bar{X}$ está "lejos" de μ_0 en ambas direcciones.

- Rechazamos $H_0 : \mu = \mu_0$ con una significación α si μ_0 no pertenece al intervalo de confianza para μ de nivel $1-\alpha$.

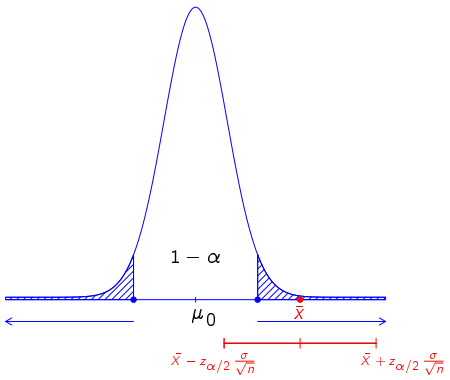


Figure 1: Relación entre contraste de hipótesis bilateral e intervalo de confianza. Si la hipótesis nula $H_0 : \ \mu = \mu_0$ es cierta, la distribución de $\bar{X}$ es normal con media μ_0. El nivel de significación α es igual al área rayada y nos sirve para definir la región crítica del test $H_0 : \ \mu = \mu_0$. Dada una muestra, calculamos el valor de $\bar{X}$. Si dicho valor pertenece a la región crítica (como en este ejemplo), rechazamos H_0 con significación α. Equivalentemente, si construimos el intervalo de confianza para μ de nivel $1-\alpha$ (en rojo) observamos que μ_0 no pertenece al intervalo.

Bioestadística. Curso 2014-2015
Capítulo 8

Carmen Mª Cadarso, Mª del Carmen Carollo, Xosé Luis Otero, Beatriz Pateiro

Contents

1 Introducción

En el capítulo anterior presentamos los conceptos básicos para el planteamiento y resolución de un contraste de hipótesis. Recordamos que los contrastes de hipótesis nos permitían responder a preguntas muy concretas sobre la población. En este capítulo veremos como llevar a cabo los contrastes de hipótesis en la práctica. Estudiaremos cuáles son los estadísticos de contraste adecuados dependiendo del parámetro al que haga referencia el test y veremos cómo construir la región crítica en cada caso.

2 Contrastes sobre la media de una población normal

Queremos contrastar hipótesis relativas a la media de una población normal $N(\mu, \sigma)$. Para ello, tomamos una muestra aleatoria simple $X_1, \ldots, X_n \in N(\mu, \sigma)$ independientes.

2.1 Contrastes sobre la media con varianza conocida

Supongamos que la varianza σ^2 es conocida. Se desea contrastar una hipótesis relativa a la media μ.

Contraste bilateral. Si queremos determinar si la media es significativamente distinta de cierto valor dado μ_0, entonces el contraste planteado sería un contraste bilateral

$$\begin{cases} H_0: \ \mu = \mu_0 \\ H_1: \ \mu \neq \mu_0 \end{cases}$$

Si la hipótesis nula $H_0 : \mu = \mu_0$ es cierta, entonces

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \in N(0,1).$$

El sentido común nos aconseja rechazar la hipótesis nula de que la media poblacional es μ_0 cuando la media muestral sea muy distinta de μ_0. Para respetar además un nivel de significación α prefijado, rechazamos la hipótesis nula $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$ si

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq z_{\alpha/2}.$$

Recuerda que $z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo $Z \in N(0,1)$, ver Figura 1.

Contraste unilateral por la derecha. Supongamos ahora que queremos determinar si la media μ es significativamente mayor que un valor dado μ_0. Entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu \leq \mu_0 \\ H_1: \ \mu > \mu_0 \end{cases}$$

Dados los valores de una muestra, parece claro que deberíamos rechazar H_0 si $\bar{X}$ es considerablemente mayor que μ_0. Rechazamos la hipótesis nula $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$ si

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \geq z_{\alpha}.$$

Contraste unilateral por la izquierda. Supongamos ahora que queremos determinar si la media μ es significativamente menor que un valor dado μ_0. Entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu \geq \mu_0 \\ H_1: \ \mu < \mu_0 \end{cases}$$

Dados los valores de una muestra, parece claro que deberíamos rechazar H_0 si $\bar{X}$ es considerablemente menor que μ_0. Rechazamos la hipótesis nula $H_0 : \mu \geq \mu_0$ frente a $H_1 : \mu < \mu_0$ si

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq -z_\alpha.$$

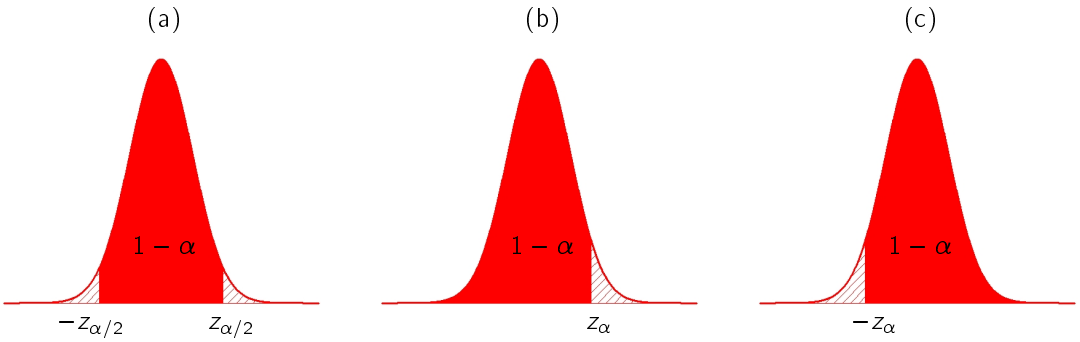


Figure 1: Densidad de una N(0,1). Regiones de aceptación y rechazo del estadístico $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$. (a) Contraste bilateral. (b) Contraste unilateral por la derecha. (c) Contraste unilateral por la izquierda.

Ejemplo 1: Según fuentes estadísticas, en la actualidad la edad media de las madres primerizas en España es de 29.3 años. Se considera una muestra de 10 madres primerizas de Portugal. Sus edades son:

30 28 27 28 28 28 24 23 31 30

Asumimos que la edad de las madres primerizas en Portugal sigue una distribución normal con una desviación típica de 2 años. Para una significación del 5%, ¿podemos concluir que la edad media de las madres primerizas en Portugal difiere de la de España? Calcula el p-valor del contraste.

Si denotamos por μ la edad media de la madres primerizas en Portugal, el contraste se plantea como un contraste bilateral de la forma:

$$\begin{cases} H_0: \ \mu = 29.3 \\ H_1: \ \mu \neq 29.3 \end{cases}$$

Rechazaremos la hipótesis nula si encontramos evidencia en los datos de que la la edad media de la madres primerizas en Portugal difiere de la de España.

La media muestral calculada a partir de los datos es $\bar{X} = 27.7$ y, por lo tanto, el estadístico de contraste será:

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{27.7 - 29.3}{2/\sqrt{10}} = -2.53.$$

Rechazaremos la hipótesis nula para una significación $\alpha = 0.05$ si el valor del estadístico de contraste es menor que $-z_{\alpha/2}$ o mayor que $z_{\alpha/2}$. Buscamos en la tabla de la N(0,1) el valor que deja a su izquierda una probabilidad 0.975 y obtenemos que $z_{\alpha/2} = 1.95$, ver Figura 2 (a). Como conclusión, rechazamos H_0 para un nivel de significación del 5%. Es decir, la edad media de las madres primerizas en Portugal es significativamente distinta de la de España.
El p-valor del contraste será (ver Figura 2 (b))

$$p\text{-valor} = 2 \cdot P(Z \leq -2.53) = 2 \cdot (1 - 0.99427) = 0.01146.$$

Es decir, rechazamos H_0 para niveles de significación α que verifiquen $\alpha \geq 0.01146$. Si el nivel de significación α es menor que el p-valor, entonces no podemos rechazar H_0 a nivel α. Por ejemplo, para $\alpha = 0.01$ no rechazamos H_0 y concluiríamos que, a un nivel del 1%, la edad media de las madres primerizas en Portugal no es significativamente distinta de la de España.

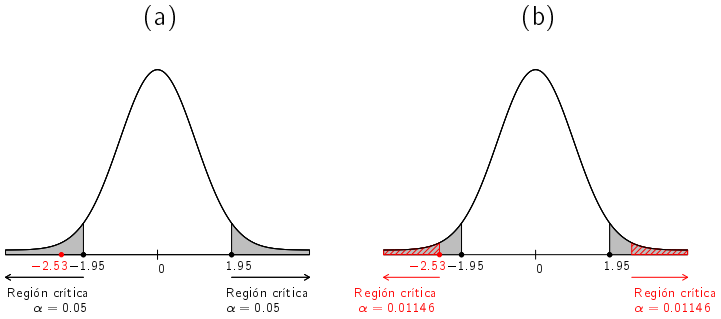


Figure 2: (a) Región crítica del contraste $H_0 : \mu = 29.3$ frente a $H_0 : \mu \neq 29.3$ del Ejemplo 1 para $\alpha = 0.05$. El estadístico del contraste pertenece a la región crítica y, por lo tanto, se rechaza la hipótesis nula H_0. (b) El p-valor del contraste 0.01146 se corresponde con el área rayada.

2.2 Contrastes sobre la media con varianza desconocida

Supongamos ahora que queremos contrastar hipótesis relativas a la media μ pero desconocemos la varianza σ^2. Podemos repetir toda la argumentación anterior con la salvedad de que cuando la varianza es desconocida, no podemos usar σ^2 y en su lugar debemos emplear un estimador adecuado como la varianza muestral S^2.

Contraste bilateral. Si queremos determinar si la media es significativamente distinta de cierto valor dado μ_0, entonces el contraste planteado sería un contraste bilateral

$$\begin{cases} H_0: \ \mu = \mu_0 \\ H_1: \ \mu \neq \mu_0 \end{cases}$$

Si la hipótesis nula $H_0 : \mu = \mu_0$ es cierta, entonces

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \in t_{n-1}.$$

El sentido común nos aconseja rechazar la hipótesis nula de que la media poblacional es μ_0 cuando la media muestral sea muy distinta de μ_0. Para respetar además un nivel de significación α prefijado, rechazamos la hipótesis nula $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$ si

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \leq -t_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq t_{\alpha/2}.$$

Recuerda que $t_{\alpha/2}$ denota el punto tal que $P(T > t_{\alpha/2}) = \alpha/2$ siendo T una variable t de Student con $n-1$ grados de libertad, ver Figura 3.

Contraste unilateral por la derecha. Supongamos ahora que queremos determinar si la media μ es significativamente mayor que un valor dado μ_0. Entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu \leq \mu_0 \\ H_1: \ \mu > \mu_0 \end{cases}$$

Dados los valores de una muestra, parece claro que deberíamos rechazar H_0 si $\bar{X}$ es considerablemente mayor que μ_0. Rechazamos la hipótesis nula $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$ si

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq t_{\alpha}.$$

Contraste unilateral por la izquierda. Supongamos ahora que queremos determinar si la media μ es significativamente menor que un valor dado μ_0. Entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu \geq \mu_0 \\ H_1: \ \mu < \mu_0 \end{cases}$$

Dados los valores de una muestra, parece claro que deberíamos rechazar H_0 si $\bar{X}$ es considerablemente menor que μ_0. Rechazamos la hipótesis nula $H_0 : \mu \geq \mu_0$ frente a $H_1 : \mu < \mu_0$ si

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \leq -t_{\alpha}.$$

En la Figura 3 se muestran las regiones de aceptación y rechazo de los contrastes sobre la media de una población con varianza desconocida.

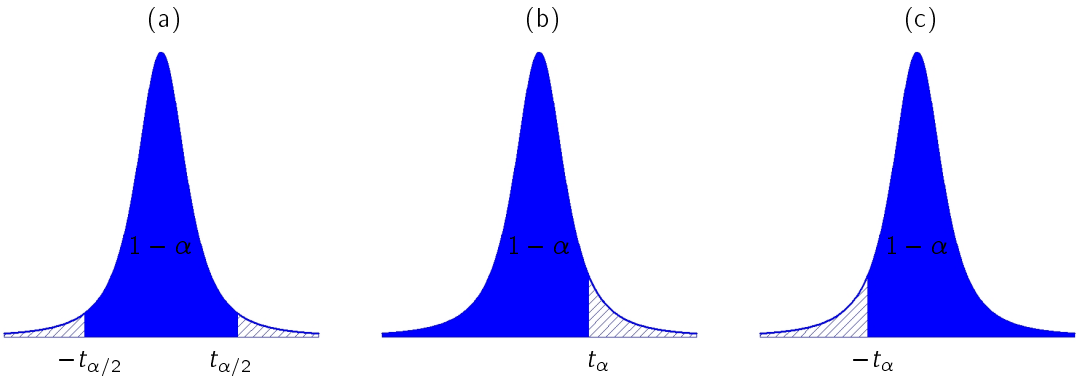


Figure 3: Densidad de una t de student con $n-1$ grados de libertad. Regiones de aceptación y rechazo del estadístico $\frac{\bar{X}-\mu_0}{S/\sqrt{n}}$. (a) Contraste bilateral. (b) Contraste unilateral por la derecha. (c) Contraste unilateral por la izquierda.

Ejemplo 2: La amilasa es un enzima que ayuda a digerir los carbohidratos. Se produce principalmente en el páncreas y en las glándulas salivales. Se ha medido el nivel de amilasa en sangre de 23 pacientes, obteniéndose una media muestral de 45 unidades por litro y una desviación típica muestral de 10 unidades por litro. Asumimos que el nivel de amilasa sigue una distribución normal. Para un nivel de significación $\alpha = 0.05$, ¿es el nivel medio de amilasa significativamente mayor que 40 unidades por litro? ¿Y para $\alpha = 0.01$?

Si denotamos por μ el nivel medio de amilasa, el contraste se plantea como un contraste unilateral de la forma:

$$\begin{cases} H_0: \ \mu \leq 40 \\ H_1: \ \mu > 40 \end{cases}$$

En este caso la varianza es desconocida y el estadístico de contraste será:

$$\frac{\bar{X}-\mu_0}{S/\sqrt{n}} = \frac{45-40}{10/\sqrt{23}} = 2.3979.$$

Rechazaremos la hipótesis nula para una significación $\alpha = 0.05$ si el valor del estadístico de contraste es mayor que t_α. Buscamos en la tabla de la t de Student con $n-1 = 22$ grados de libertad el valor que deja a su izquierda una probabilidad 0.95 y obtenemos que $t_\alpha = 1.72$. Como conclusión, rechazamos H_0 para un nivel de significación del 5%. Es decir, el nivel medio de amilasa es significativamente mayor que 40 unidades por litro.
Rechazaremos la hipótesis nula para una significación $\alpha = 0.01$ si el valor del estadístico de contraste es mayor que t_α, donde ahora t_α es el valor que en una t de Student con $n-1 = 22$ grados de libertad deja a su izquierda una probabilidad 0.99. Observamos que $t_\alpha = 2.51$ y por lo tanto no rechazamos H_0 para un nivel de significación del 1%.

3 Contrastes sobre las medias de dos poblaciones normales

3.1 Muestras independientes, varianzas conocidas

Consideremos ahora el siguiente modelo general. Tenemos dos poblaciones normales, con sus respectivas medias y varianzas: $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$ y queremos contrastar hipótesis que comparen sus medias, μ_1 y μ_2. Extraemos:

- Una muestra formada por n_1 variables independientes y con la misma distribución $N(\mu_1, \sigma_1)$.
- Una muestra formada por n_2 variables independientes y con la misma distribución $N(\mu_2, \sigma_2)$.

Suponemos que las muestras son independientes, es decir, los individuos donde se han obtenido las mediciones de la población 1 son distintos de los individuos donde se han obtenido las mediciones de la población 2. Suponemos además que las varianzas σ_1^2 y σ_2^2 son conocidas.

Contraste bilateral. Si nos preguntamos si podemos asumir que la media es la misma en ambas poblaciones, entonces el contraste planteado sería un contraste bilateral

$$\begin{cases} H_0: \ \mu_1 = \mu_2 \\ H_1: \ \mu_1 \neq \mu_2 \end{cases} \text{ o equivalentemente } \begin{cases} H_0: \ \mu_1 - \mu_2 = 0 \\ H_1: \ \mu_1 - \mu_2 \neq 0 \end{cases}$$

Si la hipótesis nula $H_0 : \mu_1 = \mu_2$ es cierta, entonces

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \in N(0,1).$$

Siguiendo el mismo razonamiento que en casos anteriores, fijado un nivel de significación α, rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq -z_{\alpha/2} \ \text{ ó } \ \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_{\alpha/2}.$$

De nuevo, $z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo $Z \in N(0,1)$.

Contraste unilateral por la derecha. Supongamos ahora que queremos determinar si la media μ_1 es significativamente mayor que μ_2. Entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu_1 \leq \mu_2 \\ H_1: \ \mu_1 > \mu_2 \end{cases} \text{ o equivalentemente } \begin{cases} H_0: \ \mu_1 - \mu_2 \leq 0 \\ H_1: \ \mu_1 - \mu_2 > 0 \end{cases}$$

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $\mu_1 > \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_\alpha$$

Contraste unilateral por la izquierda. Supongamos ahora que queremos determinar si la media μ_1 es significativamente menor que μ_2. Entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu_1 \geq \mu_2 \\ H_1: \ \mu_1 < \mu_2 \end{cases} \quad \text{o equivalentemente} \quad \begin{cases} H_0: \ \mu_1 - \mu_2 \geq 0 \\ H_1: \ \mu_1 - \mu_2 < 0 \end{cases}$$

Rechazamos la hipótesis nula $H_0 : \mu_1 \geq \mu_2$ frente a $\mu_1 < \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq -z_\alpha$$

Ejemplo 3: ¿Es la talla media de los niños (V) de 3 años mayor que la de las niñas (M) de la misma edad? Las desviaciones típicas poblacionales (en cm.) son conocidas ($\sigma_V = 4.6$, $\sigma_M = 4.5$). Medimos la talla de $n_V = 60$ niños y $n_M = 61$ niñas y obtenemos los siguientes resultados muestrales:

$$\bar{X}_V = 97.1 \text{ cm.}, \quad \bar{X}_V = 94.8 \text{ cm.}$$

Denotamos por μ_V la talla media de los niños y por μ_M la talla media de las niñas. El contraste planteado sería:

$$\begin{cases} H_0: \ \mu_V \leq \mu_M \\ H_1: \ \mu_V > \mu_M \end{cases} \quad \text{o equivalentemente} \quad \begin{cases} H_0: \ \mu_V - \mu_M \leq 0 \\ H_1: \ \mu_V - \mu_M > 0 \end{cases}$$

El estadístico de contraste en este caso será:

$$\frac{\bar{X}_V - \bar{X}_M}{\sqrt{\frac{\sigma_V^2}{n_V} + \frac{\sigma_M^2}{n_M}}} = \frac{97.1 - 94.8}{\sqrt{\frac{4.6^2}{60} + \frac{4.5^2}{61}}} = 2.7797 \approx 2.78.$$

Rechazaremos la hipótesis nula para una significación $\alpha = 0.05$ si el valor del estadístico de contraste es mayor que z_α. Buscamos en la tabla de la N(0,1) el valor que deja a su izquierda una probabilidad 0.95 y obtenemos que $z_\alpha = 1.64$, ver Figura 4 (a). Como conclusión, rechazamos H_0 para un nivel de significación del 5%. Es decir, la talla media de los niños de 3 años es significativamente mayor que la de las niñas de la misma edad.
El p-valor del contraste será (ver Figura 4 (b))

$$p\text{-valor} = P(Z \geq 2.78) = 1 - 0.997282 = 0.002718.$$

Es decir, la talla media de los niños de 3 años es significativamente mayor que la de las niñas de la misma edad para cualquier nivel de significación α que verifique $\alpha \geq 0.002718$.

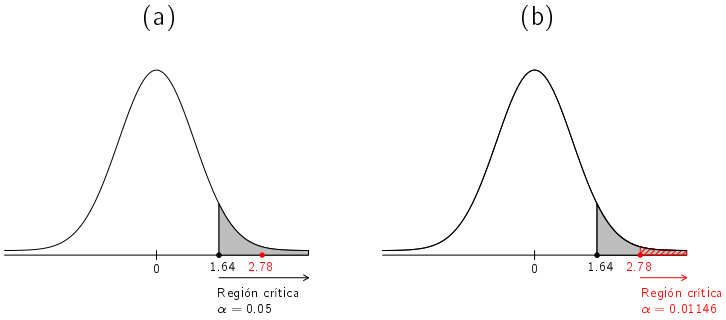


Figure 4: (a) Región crítica del contraste $H_0 : \mu_V \leq \mu_M$ frente a $H_0 : \mu_V > \mu_M$ del Ejemplo 3 para $\alpha = 0.05$. El estadístico del contraste pertenece a la región crítica y, por lo tanto, se rechaza la hipótesis nula H_0. (b) El p-valor del contraste 0.002718 se corresponde con el área rayada.

3.2 Muestras independientes, varianzas desconocidas e iguales

Como ya hemos comentado en el capítulo de intervalos de confianza, en la práctica los valores de σ_1^2 y σ_2^2 suelen ser desconocidos y por lo tanto es necesario estimarlos. No obstante, puede suceder que pese a ser desconocidas podamos suponer que ambas varianzas son iguales. Supongamos entonces que disponemos de:

- Una muestra formada por n_1 variables independientes y con la misma distribución $N(\mu_1, \sigma_1)$.
- Una muestra formada por n_2 variables independientes y con la misma distribución $N(\mu_2, \sigma_2)$.

Suponemos que las muestras son independientes y que las varianzas σ_1^2 y σ_2^2 son desconocidas pero iguales. Si suponemos que las varianzas de las dos poblaciones son iguales ya hemos visto que el mejor estimador de la varianza es:

$$S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

Recuerda que en la ecuación anterior, S_1^2 y S_2^2 denotan la varianza muestral de la primera y segunda población, respectivamente.

Contraste bilateral. Si nos preguntamos si podemos asumir que la media es la misma en ambas poblaciones, entonces el contraste planteado sería un contraste bilateral

$$\begin{cases} H_0 : \ \mu_1 = \mu_2 \\ H_1 : \ \mu_1 \neq \mu_2 \end{cases} \quad \text{o equivalentemente} \quad \begin{cases} H_0 : \ \mu_1 - \mu_2 = 0 \\ H_1 : \ \mu_1 - \mu_2 \neq 0 \end{cases}$$

Si la hipótesis nula $H_0 : \mu_1 = \mu_2$ es cierta, entonces

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_P^2}{n_1} + \frac{S_P^2}{n_2}}} \in t_{n_1+n_2-2}.$$

Fijado un nivel de significación α, rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \leq -t_{\alpha/2} \quad \text{ó} \quad \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \geq t_{\alpha/2}$$

Ahora $t_{\alpha/2}$ denota el punto tal que $P(T > t_{\alpha/2}) = \alpha/2$ siendo T una t de Student con $n_1 + n_2 - 2$ grados de libertad.

Contraste unilateral por la derecha. Supongamos ahora que queremos determinar si la media μ_1 es significativamente mayor que μ_2. Entonces, el contraste planteado sería:

$$\begin{cases} H_0 : \ \mu_1 \leq \mu_2 \\ H_1 : \ \mu_1 > \mu_2 \end{cases} \text{ o equivalentemente } \begin{cases} H_0 : \ \mu_1 - \mu_2 \leq 0 \\ H_1 : \ \mu_1 - \mu_2 > 0 \end{cases}$$

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $\mu_1 > \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \geq t_{\alpha}$$

Contraste unilateral por la izquierda. Supongamos ahora que queremos determinar si la media μ_1 es significativamente menor que μ_2. Entonces, el contraste planteado sería:

$$\begin{cases} H_0 : \ \mu_1 \geq \mu_2 \\ H_1 : \ \mu_1 < \mu_2 \end{cases} \text{ o equivalentemente } \begin{cases} H_0 : \ \mu_1 - \mu_2 \geq 0 \\ H_1 : \ \mu_1 - \mu_2 < 0 \end{cases}$$

Rechazamos la hipótesis nula $H_0 : \mu_1 \geq \mu_2$ frente a $\mu_1 < \mu_2$ si

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \leq -t_{\alpha}$$

Ejercicio 1: El Verapamil y el Nitroprusside son dos productos utilizados para reducir la hipertensión. Para compararlos, unos pacientes son tratados con Verapamil y otros con Nitroprusside. Los resultados obtenidos se muestran en la siguiente tabla, donde:

- X_1 ="reducción de la presión arterial de un paciente tratado con Verapamil"
- X_2 ="reducción de la presión arterial de un paciente tratado con Nitroprusside"

X_1	10	15	18	23	12	16	
X_2	15	10	19	9	14	12	18

Las variables X_1 y X_2 están medidas en mm. Admitiendo normalidad y sabiendo que ambas variables tienen la misma desviación típica, ¿se puede aceptar que la reducción media de hipertensión es la misma con ambos tratamientos?

3.3 Muestras apareadas

Como hemos visto en el capítulo de intervalos de confianza, en muchas ocasiones nos interesa comparar dos métodos o tratamientos. En ese caso es natural que los individuos donde se aplican los tratamientos sean los mismos. Cuando X_1 y X_2 representan características diferentes de la misma población y se quieren evaluar sus diferencias, conviene tomar muestras apareadas. Así, se obtiene el valor de las características X_1 y X_2 sobre los mismos individuos de la población. Se supone que las muestras se han obtenido de poblaciones normales $X_1 \in N\left(\mu_1, \sigma_1^2\right)$ y $X_2 \in N\left(\mu_2, \sigma_2^2\right)$ pero teniendo en cuenta que ahora X_1 y X_2 no son independientes. En esta situación considerábamos la variable $D = X_1 - X_2$.

Contraste bilateral. Si nos preguntamos si podemos asumir que la media es la misma en ambas poblaciones, entonces el contraste planteado sería un contraste bilateral

$$\begin{cases} H_0: \ \mu_1 = \mu_2 \\ H_1: \ \mu_1 \neq \mu_2 \end{cases} \quad \text{o equivalentemente} \quad \begin{cases} H_0: \ \mu_1 - \mu_2 = 0 \\ H_1: \ \mu_1 - \mu_2 \neq 0 \end{cases}$$

Si la hipótesis nula $H_0 : \mu_1 = \mu_2$ es cierta, entonces

$$\frac{\bar{D}}{S_D/\sqrt{n}} \in t_{n-1}.$$

Fijado un nivel de significación α, rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

$$\frac{\bar{D}}{S_D/\sqrt{n}} \leq -t_{\alpha/2} \quad \text{ó} \quad \frac{\bar{D}}{S_D/\sqrt{n}} \geq t_{\alpha/2}$$

siendo $t_{\alpha/2}$ el punto tal que $P(T > t_{\alpha/2}) = \alpha/2$ en una t de Student con $n-1$ grados de libertad.

Contraste unilateral por la derecha. Supongamos ahora que queremos determinar si la media μ_1 es significativamente mayor que μ_2. Entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu_1 \leq \mu_2 \\ H_1: \ \mu_1 > \mu_2 \end{cases} \quad \text{o equivalentemente} \quad \begin{cases} H_0: \ \mu_1 - \mu_2 \leq 0 \\ H_1: \ \mu_1 - \mu_2 > 0 \end{cases}$$

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $\mu_1 > \mu_2$ si

$$\frac{\bar{D}}{S_D/\sqrt{n}} \geq t_{\alpha}$$

Contraste unilateral por la izquierda. Supongamos ahora que queremos determinar si la media μ_1 es significativamente menor que μ_2. Entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ \mu_1 \geq \mu_2 \\ H_1: \ \mu_1 < \mu_2 \end{cases} \quad \text{o equivalentemente} \quad \begin{cases} H_0: \ \mu_1 - \mu_2 \geq 0 \\ H_1: \ \mu_1 - \mu_2 < 0 \end{cases}$$

Rechazamos la hipótesis nula $H_0 : \mu_1 \geq \mu_2$ frente a $\mu_1 < \mu_2$ si

$$\frac{\bar{D}}{S_D/\sqrt{n}} \leq -t_{\alpha}$$

Ejercicio 2: Se quiere estudiar los efectos del abandono de la bebida sobre la presión sistólica en individuos alcohólicos. Para ello se mide la presión sistólica en 10 individuos alcohólicos antes y después de 2 meses de haber dejado al bebida.

Sujeto	1	2	3	4	5	6	7	8	9	10
X_1 presión antes	140	165	160	160	175	190	170	175	155	160
X_2 presión después	145	150	150	160	170	175	160	165	145	170

¿Existen diferencias significativas en la presión sistólica media antes y después de abandonar la bebida?

4 Contrastes sobre una proporción

Queremos contrastar hipótesis como las propuestas en la sección anterior pero sobre una proporción p. Para ello utilizaremos como estadístico de referencia la proporción muestral $\hat{p}$.

Contraste bilateral. Si queremos determinar si la proporción p es significativamente distinta de cierto valor dado p_0, entonces el contraste planteado sería un contraste bilateral

$$\begin{cases} H_0: \ p = p_0 \\ H_1: \ p \neq p_0 \end{cases}$$

Si la hipótesis nula $H_0 : p = p_0$ es cierta, entonces (para muestras grandes)

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1).$$

El sentido común nos aconseja rechazar la hipótesis nula de que la proporción es p_0 cuando la proporción muestral $\hat{p}$ sea muy distinta de p_0. Para respetar además un nivel de significación α prefijado, rechazamos la hipótesis nula $H_0 : p = p_0$ frente a $H_1 : p \neq p_0$ si

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_{\alpha/2}$$

Aquí $z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo $Z \in N(0,1)$.

Contraste unilateral por la derecha. Supongamos ahora que queremos determinar si la proporción p es significativamente mayor que un valor dado p_0. Entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ p \leq p_0 \\ H_1: \ p > p_0 \end{cases}$$

Dados los valores de una muestra, parece claro que deberíamos rechazar H_0 si $\hat{p}$ es considerablemente mayor que p_0. Rechazamos la hipótesis nula $H_0 : p \leq p_0$ frente a $H_1 : p > p_0$ si

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geq z_{\alpha}$$

Contraste unilateral por la izquierda. Supongamos ahora que queremos determinar si la proporción p es significativamente menor que un valor dado p_0. Entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ p \geq p_0 \\ H_1: \ p < p_0 \end{cases}$$

Dados los valores de una muestra, parece claro que deberíamos rechazar H_0 si $\hat{p}$ es considerablemente menor que p_0. Rechazamos la hipótesis nula $H_0 : p \geq p_0$ frente a $H_1 : p < p_0$ si

$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq -z_\alpha.$$

Ejercicio 3: Una empresa farmacéutica quiere comercializar un medicamento que cura cierta dolencia. Se sabe que el 40% de los pacientes se curan sin tomar este medicamento. La empresa debe probar que su medicamento es eficaz y para ello administra el medicamento a 100 pacientes, de los cuales se curan 50. ¿Es el medicamento significativamente eficaz? Calcula e interpreta el p-valor del contraste.

5 Contrastes sobre dos proporciones

En algunas ocasiones estamos interesados en contrastes sobre las proporciones p_1 y p_2 de dos poblaciones. Tenemos en ese caso dos muestras:

- Una muestra formada por n_1 variables independientes de la población 1.
- Una muestra formada por n_2 variables independientes de la población 2.

Suponemos que las muestras son independientes.

Contraste bilateral. Si nos preguntamos si podemos asumir que la proporción es la misma en ambas poblaciones, entonces el contraste planteado sería un contraste bilateral

$$\begin{cases} H_0: \ p_1 = p_2 \\ H_1: \ p_1 \neq p_2 \end{cases} \quad \text{o equivalentemente} \quad \begin{cases} H_0: \ p_1 - p_2 = 0 \\ H_1: \ p_1 - p_2 \neq 0 \end{cases}$$

Si la hipótesis nula $H_0 : p_1 = p_2$ es cierta, entonces (para tamaños muestrales grandes)

$$\frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}} \sim N(0,1)$$

Siguiendo el mismo razonamiento que en casos anteriores, fijado un nivel de significación α, rechazamos la hipótesis nula $H_0 : p_1 = p_2$ frente a $H_1 : p_1 \neq p_2$ si

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \geq z_{\alpha/2}$$

Aquí $z_{\alpha/2}$ denota el punto tal que $P(Z > z_{\alpha/2}) = \alpha/2$ siendo $Z \in N(0,1)$.

Contraste unilateral por la derecha. Supongamos ahora que queremos determinar si la proporción p_1 es significativamente mayor que p_2. Entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ p_1 \leq p_2 \\ H_1: \ p_1 > p_2 \end{cases} \text{ o equivalentemente } \begin{cases} H_0: \ p_1 - p_2 \leq 0 \\ H_1: \ p_1 - p_2 > 0 \end{cases}$$

Rechazamos la hipótesis nula $H_0 : p_1 \leq p_2$ frente a $H_1 : p_1 > p_2$ si

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \geq z_\alpha$$

Contraste unilateral por la izquierda. Supongamos ahora que queremos determinar si la proporción p_1 es significativamente menor que p_2. Entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ p_1 \geq p_2 \\ H_1: \ p_1 < p_2 \end{cases} \text{ o equivalentemente } \begin{cases} H_0: \ p_1 - p_2 \geq 0 \\ H_1: \ p_1 - p_2 < 0 \end{cases}$$

Rechazamos la hipótesis nula $H_0 : p_1 \geq p_2$ frente a $p_1 < p_2$ si

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \leq -z_\alpha.$$

Ejemplo 4: La exostosis auditiva externa (EAE) es una anomalía ósea del canal auditivo externo. Esta lesión está asociada a una prolongada inmersión en agua fría y aparece con frecuencia en individuos que participan en actividades acuáticas como el surf. Se cree además que la temperatura del agua es un factor que influye en la prevalencia de EAE. Supongamos que en un estudio se examinan a 307 surfistas que surfean fundamentalmente en aguas frías (por debajo de 12°C). De los 307 surfistas examinados, 230 fueron diagnosticados de EAE. En otro estudio realizado a 75 surfistas de aguas templadas, 30 fueron diagnosticados de EAE. Para una significación del 5%, ¿se puede concluir que la prevalencia de EAE es significativamente mayor en los surfistas de aguas frías?

Sea p_1 la prevalencia de EAE en surfistas de agua fría y p_2 la prevalencia de EAE en surfistas de agua templada. Entonces, el contraste se plantea como un contraste unilateral de la forma

$$\begin{cases} H_0: \ p_1 \leq p_2 \\ H_1: \ p_1 > p_2 \end{cases} \quad \text{o equivalentemente} \quad \begin{cases} H_0: \ p_1 - p_2 \leq 0 \\ H_1: \ p_1 - p_2 > 0 \end{cases}$$

ya que queremos determinar si existe evidencia de que p_1 es mayor que p_2. Según los datos del estudio $\hat{p}_1 = 0.749$ y $\hat{p}_2 = 0.4$. El estadístico del contraste será en este caso

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = 5.65.$$

Para un nivel de significación $\alpha = 0.05$, rechazamos la hipótesis nula ya que el valor del estadístico es mayor que $z_\alpha = 1.64$ (obtenemos z_α buscando en la tabla de la N(0,1) el valor que deja a su izquierda una probabilidad 0.95). En resumen, se puede concluir que la prevalencia de EAE es significativamente mayor en surfistas de agua fría.

Bioestadística. Curso 2014-2015
Capítulo 9

Carmen Mª Cadarso, Mª del Carmen Carollo, Xosé Luis Otero, Beatriz Pateiro

Contents

1 Introducción

En el capítulo anterior hemos estudiado los métodos básicos para el contraste de hipótesis sobre parámetros de variables continuas. Los datos con los que trabajábamos consistían en una o dos muestras (dependiendo de si el contraste era sobre una o dos poblaciones) y asumíamos que dichas muestras procedían de una distribución normal. Por ejemplo, nos preguntábamos si el nivel medio de amilasa en sangre es significativamente mayor que un valor dado o si la talla media de los niños de tres años es significativamente superior que la de las niñas de la misma edad. En ambos casos estamos suponiendo que las variables de interés (nivel de amilasa en sangre, talla de niños y niñas) se distribuyen como una variable normal.
Sin embargo hay ocasiones en que la variable de estudio no es continua, sino que sus valores son de tipo categórico. Por ejemplo, supongamos que se ha llevado a cabo un estudio en niños y niñas de 3 años consistente en determinar su talla. En vez de registrar el valor numérico de la estatura, el resultado observado se clasificó en tres categorías: bajo, normal, alto. Podemos estar interesados en determinar si existe una relación estadísticamente significativa entre la talla y el sexo del niño pero ahora la variable estatura es categórica y los métodos de inferencia que debemos usar serán distintos a los vistos en el capítulo anterior. En este tema trataremos el estudio de datos categóricos y los procedimientos de inferencia adecuados en este caso.

2 Tablas de contingencia para datos categóricos

Los datos categóricos son datos que provienen de experimentos cuyos resultados son de tipo categórico, es decir, se presentan en diferentes categorías que pueden o no estar ordenadas.

En Bioestadística manejamos muchas variables con dos posibles valores o categorías: presencia o ausencia de una enfermedad o síntoma, hombre o mujer, mejoría o no mejoría tras un tratamiento,...

Ejemplo 1: Se hizo un estudio consistente en experimentar la efectividad de dos tratamientos analgésicos para la reducción del dolor en 165 pacientes con cefalea. Se registró el tipo de dolor (ausente, leve, moderado o intenso) que manifestaron sufrir los pacientes sometidos a cada tratamiento. De los 83 pacientes sometidos al tratamiento A, 12 manifestaron no sufrir dolor de cabeza, 24 dolor leve, 31 dolor moderado y 16 dolor intenso. De los 82 pacientes sometidos al tratamiento B, 20 manifestaron no sufrir dolor de cabeza, 18 dolor leve, 30 dolor moderado y 14 dolor intenso.

La forma de organizar datos de dos variables categóricas es mediante una tabla de doble entrada, llamada **tabla de contingencia**. Las tablas de contingencia están compuestas por filas (horizontales), para la información de una variable y columnas (verticales) para la información de otra variable. En cada casilla de la tabla se muestra el número de casos o individuos que poseen un nivel de una de las variables y otro nivel de la otra variable (frecuencias observadas).

Ejemplo 1: La tabla de contingencia 2×4 (2 filas y 4 columnas) asociada al Ejemplo 1 es:

Las frecuencias representadas en cada casilla de una tabla de contingencia se denominan frecuencias observadas

	Dolor			
Tratamiento	Ausente	Leve	Moderado	Intenso
A	12	24	31	16
B	20	18	30	14

En la tabla de contingencia también se suelen representar:

- Los totales de cada fila, que se llaman marginales de las filas.
- Los totales de cada columna, que se llaman marginales de las columnas.
- El número total de individuos.

Ejemplo 1: Podemos completar la tabla de contingencia del Ejemplo 1 con los totales.

	Dolor				
Tratamiento	Ausente	Leve	Moderado	Intenso	Total
A	12	24	31	16	83
B	20	18	30	14	82
Total	32	42	61	30	165

2.1 Tablas 2×2

Una tabla de contingencia 2×2 está formada por dos filas y dos columnas. Se utiliza para representar datos de dos variables, cada una de las cuales presenta dos únicos valores o categorías. En esta situación la tabla de contingencia se reduce a una tabla 2×2 como la que se muestra a continuación:

	Variable 1	
Variable 2	Valor 1	Valor 2
Valor 1	a	b
Valor 2	c	d

Ejemplo 2: Se ha planteado la hipótesis de que el cáncer de mama en mujeres está causado en parte por eventos que ocurren entre la edad de la primera menstruación y la edad al nacer el primer hijo. En particular, se cree que el riesgo de cáncer de mama aumenta cuanto mayor es este intervalo de tiempo. Esto significaría que la edad a la que las mujeres tienen su primer hijo es un factor de riesgo importante en la incidencia de esta enfermedad. Se ha llevado a cabo un estudio a nivel internacional para contrastar esta hipótesis. En él participaron 3220 mujeres con cáncer de mama (casos) y 10245 mujeres sin cáncer de mama (controles). La edad a la que las mujeres del estudio tuvieron su primer hijo fue categorizada en ≥ 30 años y ≤ 29 años. Los datos del estudio se resumen en la siguiente tabla 2×2.

	Tipo	
Edad al tener el primer hijo	Caso	Control
≥ 30	683	1498
≤ 29	2537	8747

Las frecuencias observadas son a = 683, b = 1498, c = 2537 y d = 8747.

Ejemplo tomado del libro Fundamentals of Biostatistics. Rosner, B. (2000)

Estudio de casos y controles: Este tipo de estudio identifica a personas con una enfermedad (casos) y los compara con un grupo control apropiado que no tenga la enfermedad. Una vez seleccionados los individuos en cada grupo, se investiga si estuvieron expuestos o no a una característica de interés y se compara la proporción de expuestos en el grupo de casos frente a la del grupo de controles.

Si además representamos los totales, tendremos:

	Variable 1		
Variable 2	Valor 1	Valor 2	Total
Valor 1	a	b	a + b
Valor 2	c	d	c + d
Total	a + c	b + d	a + b + c + d

Ejemplo 2: Volviendo al estudio sobre el cáncer de mama obtenemos:

	Tipo		
Edad al tener el primer hijo	Caso	Control	Total
≥ 30	683	1498	2181
≤ 29	2537	8747	11284
Total	3220	10245	13465

Así, el número de casos es $a + c = 3220$ mujeres. El número de controles es $b + d = 10245$. El número de mujeres del estudio que han tenido su primer hijo con más de 30 años es $a + b = 2181$ mujeres. El número de mujeres del estudio que han tenido su primer hijo con menos de 29 años es $c + d = 11284$ mujeres. El total de observaciones es $n = a + b + c + d = 13465$ mujeres.

Ante una tabla de contingencia como las anteriores se pueden plantear distintas cuestiones. Por ejemplo, podemos estar interesados en determinar si existe una relación estadísticamente significativa entre las variables estudiadas. Para responder a esta cuestión utilizaremos la metodología de análisis de las tablas de contingencia. Existen diferentes procedimientos como el test Chi-cuadrado que veremos a continuación. También nos puede interesar cuantificar la relación entre las variables de interés y estudiar su relevancia clínica. Esta última cuestión podrá resolverse mediante las denominadas medidas de asociación o de efecto como el riesgo relativo (RR) y odds-ratio (OR).
Tanto las medidas de efecto como las pruebas estadísticas a utilizar dependerán del diseño del estudio del que proceden los datos. Veremos diferentes tipos de estudios que se pueden llevar a cabo.

3 Pruebas Chi-cuadrado

Las pruebas Chi-cuadrado, o pruebas χ^2 de Pearson, son un grupo de contrastes de hipótesis que se aplican en dos situaciones básicas:

Las pruebas que tiene por objetivo determinar si los datos se ajustan a una determinada distribución se denominan pruebas de bondad de ajuste

- Para comprobar afirmaciones acerca de las funciones de probabilidad (o densidad) de una variable aleatoria. Por ejemplo, si queremos contrastar si una determinada variable sigue una distribución normal.
- Para determinar si dos variables son independientes estadísticamente. En este caso la prueba que aplicaremos será el test χ^2 de independencia.

3.1 Test Chi-cuadrado de independencia en tablas 2×2

El test χ^2 de independencia nos permite determinar si dos variables cualitativas X e Y están o no asociadas. Si concluimos que las variables no están relacionadas podremos decir con un determinado

nivel de confianza, previamente fijado, que ambas son independientes. El contraste se plantea como:

$$\begin{cases} H_0: \ X \text{ e } Y \text{ son independientes} \\ H_1: \ X \text{ e } Y \text{ no son independientes} \end{cases}$$

Veremos como se lleva a cabo el test χ^2 de independencia en el caso particular de una tabla de contingencia 2×2. El test se podrá generalizar para contrastar la independencia de variables que presenten más de dos posibles valores o categorías.

Ejemplo 2: Volvemos al estudio sobre el cáncer de mama. El objetivo es determinar si existe una relación estadísticamente significativa entre el desarrollo de la enfermedad y la edad a la que la mujer tiene el primer hijo. Es decir, si llamamos:

- X = Cáncer de mama (sí o no)
- Y = Edad a la que la mujer tiene el primer hijo (≤ 29 ó ≥ 30)

entonces, el contraste planteado sería:

$$\begin{cases} H_0: \ X \text{ e } Y \text{ son independientes} \\ H_1: \ X \text{ e } Y \text{ no son independientes.} \end{cases}$$

Si la hipótesis nula fuese cierta, la proporción de mujeres con cáncer de mama que tuvieron su primer hijo con menos de 29 años debería ser la misma que la proporción de mujeres con cáncer de mama que tuvieron su primer hijo con más de 30 años. Entonces, si H_0 fuese cierta, de las 3220 mujeres con cáncer de mama ¿cuántas esperaríamos que hubiesen tenido su primer hijo con más de 30 años? ¿y con menos de 29?
El número esperado de casos con más de 30 años de edad al tener el primer hijo sería:

$$E_{11} = 2181 \frac{3220}{13465} = 521.561.$$

El número esperado de casos con menos de 29 años al tener el primer hijo sería:

$$E_{21} = 11284 \frac{3220}{13465} = 2698.439.$$

Del mismo modo, si la hipótesis nula fuese cierta, la proporción de mujeres sin cáncer de mama que tuvieron su primer hijo con menos de 29 años debería ser la misma que la de mujeres sin cáncer de mama que tuvieron su primer hijo con más de 30 años. Entonces, si H_0 fuese cierta, de las 10245 mujeres sin cáncer de mama ¿cuántas esperaríamos que hubiesen tenido su primer hijo con más de 30 años? ¿y con menos de 29? El número esperado de controles con más de 30 años al tener el primer hijo sería:

$$E_{12} = 2181 \frac{10245}{13465} = 1659.439.$$

El número esperado de controles con menos de 29 años al tener el primer hijo sería:

$$E_{22} = 11284 \frac{10245}{13465} = 8585.561.$$

La tabla de valores observados y esperados bajo la hipótesis nula es entonces:

	Tipo		
Edad al tener el primer hijo	Caso	Control	Total
≥ 30	683 (521.561)	1498 (1659.439)	2181
≤ 29	2537 (2698.439)	8747 (8585.561)	11284
Total	3220	10245	13465

Comparamos ahora los datos observados con los datos esperados (entre paréntesis). Si dichos valores son considerablemente distintos, deberíamos rechazar la hipótesis nula de independencia.

Los valores esperados en una tabla de contingencia se calculan a través del producto de los totales marginales dividido por el número total de individuos. En el caso particular de una tabla 2×2 se tiene:

$$E_{11} = \frac{(a+c)\cdot(a+b)}{a+b+c+d} \qquad E_{12} = \frac{(b+d)\cdot(a+b)}{a+b+c+d}$$

$$E_{21} = \frac{(a+c)\cdot(c+d)}{a+b+c+d} \qquad E_{22} = \frac{(b+d)\cdot(c+d)}{a+b+c+d}$$

Si denotamos por O_{ij} los valores observados en la realidad, podemos representar los valores observados y esperados en la misma tabla como se muestra a continuación.

	Variable 1	
Variable 2	Valor 1	Valor 2
Valor 1	O_{11} (E_{11})	O_{12} (E_{12})
Valor 2	O_{21} (E_{21})	O_{22} (E_{22})

El test χ^2 de independencia mide la diferencia entre los valores E_{ij} que deberíamos haber obtenido si las dos variables fuesen independientes y los que se han observado en la realidad O_{ij}. El estadístico del contraste es:

$$\chi^2 = \sum_{\text{todas las celdas}} \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}}.$$

Es decir,

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Cuanto mayor sea la diferencia entre los valores esperados y observados, mayor será el valor de este estadístico. Por lo tanto, deberemos rechazar H_0 cuando el valor de χ^2 sea "grande".
Bajo la hipótesis nula de independencia, se sabe que los valores del estadístico se distribuyen aproximadamente según una distribución Chi-cuadrado. Para el caso de una tabla de contingencia de r filas y s columnas, los grados de libertad son $(r-1)(s-1)$. Por lo tanto, para el caso particular de una tabla 2×2, el estadístico sigue aproximadamente una distribución Chi-cuadrado con 1 grado de libertad bajo H_0.
En resumen, para tablas de contingencia 2×2, rechazaremos la hipótesis nula de independencia para una significación α si

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \geq \chi_\alpha$$

donde χ_α es el punto que deja a su derecha una probabilidad α en una distribución Chi-cuadrado con 1 grado de libertad, ver Figura 1.

Para que la aproximación por la distribución Chi-cuadrado sea buena, es conveniente que las frecuencias esperadas sean grandes. Como criterio en tablas 2×2 se pide que todos los valores esperados E_{ij} sean mayores que 5.

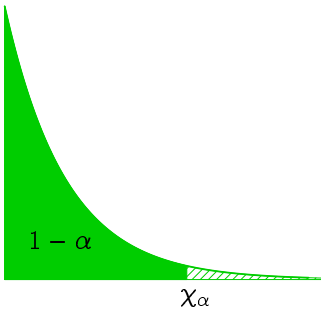


Figure 1: Densidad de una χ^2 con 1 grado de libertad. Regiones de aceptación y rechazo del estadístico χ^2 de Pearson para tablas de contingencia 2×2.

Ejemplo 2: Calculamos el valor del estadístico χ^2 para el ejemplo del estudio sobre el cáncer de mama. Se tiene:

$$\chi^2 = \frac{(683 - 521.561)^2}{521.561} + \frac{(1498 - 1659.439)^2}{1659.439} + \frac{(2537 - 2698.439)^2}{2698.439} + \frac{(8747 - 8585.561)^2}{8585.561} = 78.369.$$

Consultamos la tabla de la distribución Chi-cuadrado (con 1 grado de libertad) y concluimos que, para un nivel de significación $\alpha = 0.05$, rechazamos la hipótesis nula de que el desarrollo de la enfermedad es independiente de la edad a la que la mujer tiene el primer hijo ya que el valor del estadístico es mayor que $\chi_\alpha = 3.84$.
De hecho, incluso para una significación $\alpha = 0.005$, rechazaríamos la hipótesis nula de independencia ya que también en este caso el valor del estadístico es mayor que $\chi_\alpha = 7.88$. Por lo tanto, podemos concluir que el cáncer de mama está significativamente asociado con la edad a la que la mujer tiene el primer hijo.

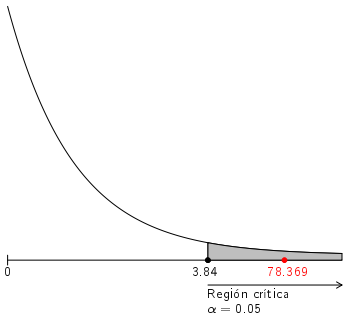


Figure 2: Región crítica del contraste de independencia del Ejemplo 2. El estadístico del contraste pertenece a la región crítica y, por lo tanto, se rechaza la hipótesis nula H_0.

Además, para el caso de una tabla 2×2, la expresión del estadístico χ^2 puede simplificarse y obtenerse

como:

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Ejemplo 2: Comprobamos que el estadístico χ^2 para los estudio sobre el cáncer de mama se calcula también como:

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} = \frac{13465 \cdot (683 \cdot 8747 - 1498 \cdot 2537)^2}{2181 \cdot 11284 \cdot 3220 \cdot 10245} = 78.369.$$

Ejemplo 3: El problema planteado en el Ejemplo 2 se puede enfocar desde la perspectiva de los contrastes sobre dos proporciones vistos en el capítulo anterior. Recuerda que el problema planteaba la hipótesis de que la edad a la que las mujeres tienen su primer hijo podría ser un factor de riesgo importante en la incidencia del cáncer de mama. En el estudio participaron 13465 mujeres (3220 casos y 10245 controles).

De entre los casos, 683 mujeres tuvieron su primer hijo con más de 30 años. De entre los controles, 1498 mujeres tuvieron su primer hijo con más de 30 años. En base a esos datos, ¿hay evidencia significativa de que retrasar la edad a la que se tiene el primer hijo afecta a la incidencia de cáncer de mama? Si llamamos p_1 a la proporción de mujeres con cáncer de mama que han tenido su primer hijo con más de 30 años y p_2 a la proporción de mujeres sin cáncer de mama que han tenido su primer hijo con más de 30 años, el contraste se puede plantear como

$$\begin{cases} H_0: \ p_1 = p_2 \\ H_1: \ p_1 \neq p_2 \end{cases} \quad \text{o equivalentemente} \quad \begin{cases} H_0: \ p_1 - p_2 = 0 \\ H_1: \ p_1 - p_2 \neq 0 \end{cases}$$

En este caso $\hat{p}_1 = 683/3220 = 0.212$ y $\hat{p}_2 = 1498/10245 = 0.146$. El estadístico del contraste será:

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} = 8.231.$$

Para un nivel de significación $\alpha = 0.05$, rechazamos la hipótesis nula ya que el valor del estadístico es mayor que $z_{\alpha/2} = 1.95$. Se concluye entonces que hay evidencia significativa de que la proporción de mujeres con cáncer de pecho que han tenido su primer hijo con más de 30 años es significativamente distinta que la de mujeres sin cáncer de pecho que han tenido su primer hijo con más de 30 años. Además, puedes comprobar que si se plantea un contraste unilateral del tipo

$$\begin{cases} H_0: \ p_1 \leq p_2 \\ H_1: \ p_1 > p_2 \end{cases} \quad \text{o equivalentemente} \quad \begin{cases} H_0: \ p_1 - p_2 \leq 0 \\ H_1: \ p_1 - p_2 > 0 \end{cases}$$

también se rechaza la hipótesis nula. Es decir, la proporción de mujeres con cáncer de pecho que han tenido su primer hijo con más de 30 años es significativamente mayor que la de mujeres sin cáncer de pecho que han tenido su primer hijo con más de 30 años.

3.1.1 Corrección por continuidad

Ya hemos comentado que, para que la aproximación por la distribución Chi-cuadrado sea buena, es conveniente que las frecuencias esperadas sean grandes. Como criterio en tablas 2×2 se pide que todos los valores esperados E_{ij} sean mayores que 5. Aun así, en tablas 2×2 la aproximación a la Chi-cuadrado puede no ser buena y, por eso, se suele aplicar la llamada **corrección por continuidad de Yates**. Esta corrección consiste en restar 0.5 a cada una de las diferencias (sin signo) entre valores observados y esperados, es decir:

La notación $|\cdot|$ se utiliza para representar valor absoluto. El valor absoluto de un número es su valor numérico sin tener en cuenta su signo

$$\chi^2_{\text{corregido}} = \sum_{i\,j} \frac{(|O_{ij} - E_{ij}| - 0.5)^2}{E_{ij}}.$$

Ejemplo 2: Calculamos el valor del estadístico χ^2 corregido para el ejemplo del estudio sobre el cáncer de mama. Observamos que las diferencias entre valores observados y esperados son todas 161.438 o -161.438. Entonces:

$$\begin{aligned} \chi^2_{\text{corregido}} &= \frac{(161.438-0.5)^2}{521.561} + \frac{(161.438-0.5)^2}{2698.439} + \frac{(161.438-0.5)^2}{1659.439} + \frac{(161.438-0.5)^2}{8585.561} \\ &= 77.885. \end{aligned}$$

Consultamos la tabla de la distribución Chi-cuadrado (con 1 grado de libertad) y concluimos que, para un nivel de significación $\alpha = 0.05$, rechazamos la hipótesis nula de que el desarrollo de la enfermedad es independiente de la edad a la que la mujer tiene el primer hijo ya que el valor del estadístico corregido es mayor que $\chi_\alpha = 3.84$.

3.2 Test Chi-cuadrado de independencia en tablas $r \times s$

Veremos ahora como se lleva a cabo el test χ^2 de independencia en el caso general de una tabla de contingencia $r \times s$ (r filas, s columnas).

Ejemplo 3: Se ha llevado a cabo una encuesta sobre salud en un determinado país. En la siguiente tabla se muestran los resultados de dos de las preguntas incluidas en el cuestionario. La primera pregunta era: "En general, ¿definiría su estado de salud como excelente, bueno, normal o deficiente?". La segunda pregunta era: "¿Puede hace frente al pago de los servicios sanitarios que necesita?" Las posibles respuestas eran *casi nunca*, *normalmente no*, *normalmente sí* o *siempre*.

	Pago servicios sanitarios				
Estado de Salud	Casi nunca	Normalmente no	Normalmente sí	Siempre	Total
Excelente	4	20	21	99	144
Bueno	12	43	59	195	309
Normal	11	21	15	58	105
Deficiente	8	9	8	17	42
Total	35	93	103	369	600

Recordamos que estamos interesados en determinar si dos variables cualitativas X e Y están o no asociadas. Ahora X o Y pueden presentar más de dos posibles valores o categorías. El contraste se

plantea como:

$$\begin{cases} H_0: \ X \text{ e } Y \text{ son independientes} \\ H_1: \ X \text{ e } Y \text{ no son independientes} \end{cases}$$

Igual que antes, el test χ^2 de independencia mide la diferencia entre los valores esperados E_{ij} que deberíamos haber obtenido si las dos variables fuesen independientes y los que se han observado en la realidad O_{ij}. El estadístico del contraste es:

$$\chi^2 = \sum_{\text{todas las celdas}} \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}}.$$

Los valores esperados se calculan usando el mismo método que para tablas 2×2. Para cada celda, se multiplican los totales marginales de la fila y columna correspondiente y se divide el resultado entre el número total de individuos. Es decir

$$E_{ij} = \frac{\text{Total marginal de la fila } i \cdot \text{Total marginal de la columna } j}{\text{Total de individuos}}$$

Ejemplo 3: Nos preguntamos si el estado de salud está relacionado con la capacidad que tienen los pacientes de hacer frente al pago de los servicios sanitarios. Calculamos la tabla de valores observados y esperados (entre paréntesis) para la tabla del Ejemplo 3.

	Pago servicios sanitarios				
Estado de Salud	Casi nunca	Normalmente no	Normalmente sí	Siempre	Total
Excelente	4(8.40)	20(22.32)	21(24.72)	99(88.56)	144
Bueno	12(18.02)	43(47.90)	59(53.04)	195(190.04)	309
Normal	11(6.13)	21(16.27)	15(18.02)	58(64.57)	105
Deficiente	8(2.45)	9(6.51)	8(7.21)	17(25.83)	42
Total	35	93	103	369	600

Por ejemplo, si suponemos que el estado de salud es independiente de la capacidad para hacer frente al pago de los servicios sanitarios, el número esperado de pacientes con un estado de salud bueno y que normalmente pueden hacer frente al pago de los servicios sanitarios sería E_{23} (fila 2, columna 3)

$$E_{23} = \frac{309 \cdot 103}{600} = 53.04.$$

Una vez calculada la tabla de valores observados y esperados, podemos calcular el valor del estadístico Chi-cuadrado,

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Cuanto mayor sea la diferencia entre los valores esperados y observados, mayor será el valor de este estadístico. Por lo tanto, deberemos rechazar H_0 cuando el valor de χ^2 sea "grande". Bajo la hipótesis nula de independencia, se sabe que los valores del estadístico se distribuyen aproximadamente según una distribución Chi-cuadrado. Para el caso de una tabla de contingencia de r filas y s columnas, los grados de libertad son $(r-1)(s-1)$. Es decir, rechazaremos la hipótesis nula de independencia para

una significación α si

$$\chi^2 = \sum_{i\,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \geq \chi_\alpha$$

donde χ_α es el punto que deja a su derecha una probabilidad α en una distribución Chi-cuadrado con $(r-1)(s-1)$ grados de libertad, ver Figura 3.

Para que la aproximación por la distribución Chi-cuadrado sea buena, es conveniente que las frecuencias esperadas sean grandes. Como criterio en tablas $r \times s$ se pide que no más del 20% de los valores esperados E_{ij} sean inferiores a 5.

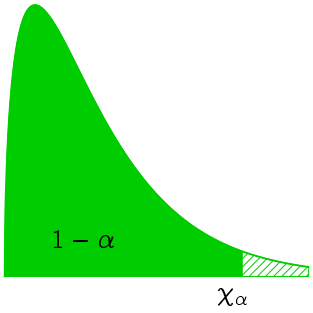


Figure 3: Densidad de una χ^2 con $(r-1)(s-1)$ grado de libertad. Regiones de aceptación y rechazo del estadístico χ^2 de Pearson para tablas de contingencia $r \times s$.

Ejemplo 3: Calculamos el valor del estadístico χ^2 para el Ejemplo 3. Se tiene:

$$\chi^2 = \frac{(4-8.40)^2}{8.40} + \frac{(20-22.32)^2}{22.32} + \ldots + \frac{(17-25.83)^2}{25.83} = 30.7078.$$

Consultamos la tabla de la Chi-cuadrado con $(r-1)(s-1) = (4-1)(4-1) = 9$ grados de libertad y concluimos que, para un nivel de significación $\alpha = 0.05$, rechazamos la hipótesis nula de que el estado de salud es independiente de la capacidad para hacer frente al pago de los servicios sanitarios ya que el valor del estadístico es mayor que $\chi_\alpha = 16.9$.

4 Tipos de estudios

Los estudios epidemiológicos son los estudios en los que se basa la investigación médica y permiten establecer la relación entre las causas de una enfermedad y la influencia de éstas en el desarrollo (o no) de la enfermedad. Existen numerosas clasificaciones de los estudios epidemiológicos dependiendo de si atendemos a su finalidad, a su secuencia temporal, al control que se tenga sobre los factores del estudio,...
Clasificaremos aquí los estudios epidemiológicos según el tipo de intervención que exista en el estudio. Dependiendo de si existe o no intervención, los estudios se clasifican en:

- Estudios observacionales: Son aquellos en los que el factor de estudio no es controlado por el investigador. El investigador se limita a observar y medir. Son ejemplos de estudios observacionales el estudio caso–control, estudio de cohortes y el estudio de prevalencia o transversal.
  - Estudios caso–control: En los estudios de casos y controles los sujetos incluidos proceden típicamente de dos grupos, según sean casos (con la enfermedad o daño en estudio) o

controles (sin el daño en cuestión). La idea es comparar los antecedentes de los enfermos de una población con los de los sanos de la misma población. Se trata de poner de manifiesto posibles diferencias en las exposiciones que expliquen, al menos parcialmente, la razón por la que unos enfermaron y otros no.

  - Estudios de cohortes: Los estudios de cohorte se basan en el seguimiento en el tiempo de dos o más grupos de individuos que han sido divididos según el grado de exposición a un determinado factor (corrientemente en 2 grupos: expuestos y no expuestos). Al inicio, ninguno de los individuos incluidos en ambos grupos tiene la enfermedad o daño en estudio. Los individuos son seguidos durante un período de tiempo para observar la frecuencia de aparición del fenómeno que nos interesa. Si al finalizar el período de observación la incidencia de la enfermedad es mayor en el grupo de expuestos, podremos concluir que existe una asociación estadística entre la exposición a la variable y la incidencia de la enfermedad.
  - Estudio de prevalencia o transversal: Los estudios transversales examinan las relaciones entre las enfermedades y otras variables de interés en una población y momento determinados. La presencia o ausencia de la enfermedad y de las otras variables se determinan en cada miembro de la población estudiada o en una muestra representativa en un momento dado. La secuencia temporal de causa a efecto no queda necesariamente determinada en un estudio de este tipo.

- Estudios experimentales: El investigador asigna un factor de estudio y lo controla a lo largo de la investigación. Este tipo de estudios se utilizan para evaluar la eficacia de diferentes terapias, de actividades preventivas o para la evaluación de actividades de planificación y programación sanitarias. Son ejemplos de estudios experimentales los ensayos clínicos.

  - Ensayos clínicos: Los ensayos clínicos son experimentos planificados sobre pacientes cuyo objetivo es evaluar la eficacia de tratamientos e intervenciones médicas o quirúrgicas.

Ejemplo 4: El estudio sobre el cancer de mama descrito en el Ejemplo 2 es un estudio caso–control.

Ejemplo 5: El Estudio del Corazón de Framingham. El Estudio de Framingham es un conocido estudio de cohorte que se inició en 1948 bajo la dirección del Instituto Nacional Cardíaco, Pulmonar y Sanguíneo de EEUU. El objetivo del mismo era la identificación de los factores o características comunes que contribuían a las enfermedades cardiovasculares, mediante el seguimiento a largo plazo de un gran número de individuos que en el momento de su incorporación al estudio todavía no habían manifestado evidencia clínica de la enfermedad. Inicialmente se reclutaron 5.209 varones y mujeres con edades comprendidas entre los 30 y 62 años, residentes en Framingham, Massachussets. Así comenzó la primera serie de exámenes médicos, clínicos, bioquímicos y de estilos de vida que constituirían las bases para el análisis de los patrones comunes relacionados con el desarrollo de las enfermedades cardiovasculares.

http://www.framinghamheartstudy.org/index.html

Ejemplo 6: Un estudio transversal para conocer la prevalencia de osteoporosis y su relación con algunos factores de riesgo potenciales incluyó a 400 mujeres con edades entre 50 y 54 años. A cada una se le realizó una densitometría de columna y en cada caso se completó un cuestionario de antecedentes.

Ejemplo 7: El hospital Northwestern Medicine de Chicago participa en el primer ensayo clínico con células madre embrionarias humanas. El ensayo pretende probar la seguridad y tolerancia, y eventualmente la eficacia, del tratamiento en parapléjicos recientes, para reparar los daños sufridos en la médula espinal.

Noticia de El País (30/09/2010).

5 Medidas de efecto: riesgo relativo y odds-ratio

La relación entre las variables se puede cuantificar mediante el cálculo de medidas de asociación como el riesgo relativo (RR) y la odds-ratio (OR).

5.1 Riesgo relativo

El riesgo relativo (RR) es una razón que relaciona la incidencia en dos grupos de población que difieren por el grado de exposición a un factor determinado. Es decir:

$$\text{RR} = \frac{\text{Incidencia en el grupo 1}}{\text{Incidencia en el grupo 2}}$$

Generalmente, el grupo 2 se encuentra en condiciones normales (no expuestos a cierto factor de riesgo) mientras que el grupo 1 se encuentra expuesto al factor de riesgo. De esta forma, un RR mayor que 1 indicaría efectos nocivos del factor de riesgo, es decir, la presencia del factor de riesgo se asocia a una mayor incidencia. Un RR menor que 1 indicaría que la presencia del factor de riesgo se asocia a una menor incidencia (factor de protección). Un RR igual a 1 indicaría que no hay asociación entre la presencia del factor de riesgo y la incidencia de la enfermedad.

Si consideramos la tabla de contingencia 2×2

	Enfermedad	
Factor de riesgo o exposición	Sí	No
Presente	a	b
Ausente	c	d

se tendría:

$$\text{RR} = \frac{a/(a+b)}{c/(c+d)}.$$

5.2 Odds-ratio

En muchas ocasiones el número de sujetos clasificados como enfermos es pequeño comparado con el número de sujetos clasificados como no enfermos, es decir:

$$a + b \approx b$$
$$c + d \approx d$$

En ese caso el riesgo relativo se aproxima por:

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}.$$

A esta medida se le denomina odds-ratio o razón de ventajas.

Bioestadística. Curso 2014-2015
Capítulo 10

Carmen Mª Cadarso, Mª del Carmen Carollo, Xosé Luis Otero, Beatriz Pateiro

Contents

1 Introducción

En el primer capítulo nos hemos ocupado de la descripción de variables estadísticas unidimensionales, es decir, cada individuo de la muestra era descrito de acuerdo a una única característica. Sin embargo, lo habitual es que tendamos a considerar un conjunto amplio de características para describir a cada uno de los individuos de la población, y que estas características puedan presentar relación entre ellas. Así, si para un mismo individuo observamos simultáneamente k características obtenemos como resultado una variable estadística k-dimensional. Nos centraremos en el estudio de variables estadísticas bidimensionales, es decir, tendremos dos características por cada individuo. Representaremos por (X,Y) la variable bidimensional estudiada, donde X e Y son las variables unidimensionales correspondientes a las primera y segunda características, respectivamente, medidas para cada individuo. En el estudio de variables bidimensionales tiene mucho interés buscar posibles relaciones entre las variables X e Y. Por ejemplo, ¿existe relación entre la altura en el peso?, ¿cómo se relaciona la cantidad de dinero que se ha invertido un laboratorio para anunciar un nuevo fármaco con las cifras de ventas durante el primer mes?, ¿está relacionada la altura de un padre con la de su hijo?. El tipo de relación más sencilla que se establece entre un par de variables es la relación lineal. Estudiaremos en este capítulo este tipo de relaciones.

2 Conceptos generales

Estudiaremos las características (X,Y) de una población a partir de la información recogida en una muestra de tamaño n de la forma $(x_1, y_1), \ldots, (x_n, y_n)$.

Ejemplo 1: EL Volumen Expiratorio Forzado (VEF) es una medida de la función pulmonar. Se cree que el VEF está relacionado con la estatura. Nos interesa estudiar la variable bidimensional (X,Y) siendo X la estatura de niños de 10 a 15 años de edad e Y el VEF. A continuación se muestra la estatura (en cm.) y el VEF (en l.) de 12 niños en ese rango de edad:

Estatura	134	138	142	146	150	154	158	162	166	170	174	178
VEF	1.7	1.9	2.0	2.1	2.2	2.5	2.7	3.0	3.1	3.4	3.8	3.9

Es decir, contamos con la información recogida en una muestra de tamaño $n = 12$ de la forma $(134, 1.7), (138, 1.9), \ldots, (178, 3.9)$.

2.1 El diagrama de dispersión

La representación gráfica más útil de dos variables continuas es el **diagrama de dispersión**. Consiste en representar en un eje de coordenadas los pares de observaciones (x_i, y_i). La nube así dibujada (a este gráfico también se le llama nube de puntos) refleja la posible relación entre las variables. A mayor relación entre las variables más estrecha y alargada será la nube. En la Figura 1 se muestran ejemplos de diferentes diagramas de dispersión.

Ejercicio 1: ¿Te parece que existe relación lineal entre las variables X e Y representadas en los gráficos de dispersión de la Figura 1? ¿Qué tipo de relación crees que existe en cada uno de los ejemplos representados?

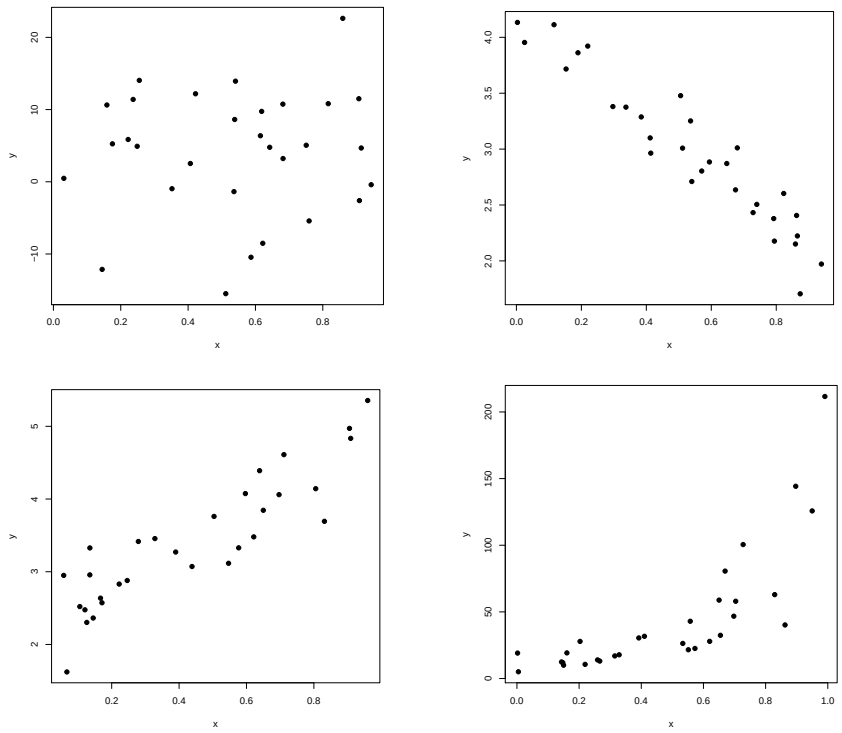


Figure 1: Diferentes diagramas de dispersión

Ejemplo 1: Para los datos del Ejemplo 1, se obtiene el diagrama de dispersión de la Figura 2. A partir de la gráfica se observa que parece existir una clara relación lineal entre ambas variables, de manera que a medida que aumenta la estatura, también aumenta el VEF y además lo hace de forma lineal.

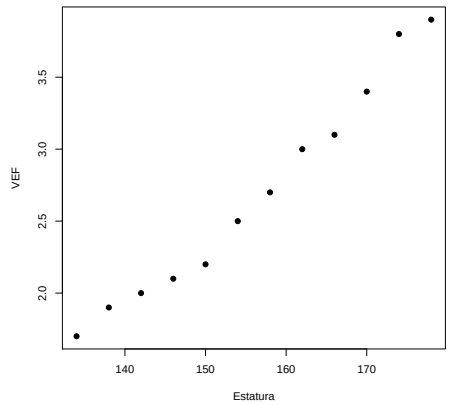


Figure 2: Diagrama de dispersión para los datos del Ejemplo 1

2.2 Covarianza

La mayoría de las medidas características estudiadas en el caso unidimensional (como por ejemplo la media) pueden extenderse al caso bidimensional. Además, en el contexto bidimensional surgen nuevas medidas que nos permiten cuantificar la dispersión conjunta de dos variables estadísticas. Consideremos una muestra de n observaciones de una variable bidimensional cuantitativa (X, Y).

Se define la **covarianza** entre X e Y (que se denota por s_{xy}) como:

$$\text{Cov}(X, Y) = s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

La covarianza puede interpretarse como una medida de relación lineal entre las variables X e Y.

Propiedades:

1. La covarianza de (X, Y) es igual a la de (Y, X), es decir, $s_{xy} = s_{yx}$
2. La covarianza de (X, X) es igual a la varianza de X, es decir $s_{xx} = s_x^2$

Ejemplo 1: Para los datos del Ejemplo 1 se obtiene que la estatura media es $\bar{x} = 156$ centímetros y el VEF medio es $\bar{y} = 2.691$ litros. La covarianza entre X e Y se calcula como

$$s_{xy} = \frac{(134 - 156) \cdot (1.7 - 2.691) + \ldots + (178 - 156) \cdot (3.9 - 2.691)}{11} = 10.672$$

El signo de la covarianza nos indica que hay una relación positiva, es decir, a medida que aumenta la estatura aumenta el VEF.

2.3 Coeficiente de correlación lineal

La covarianza cambia si modificamos las unidades de medida de las variables. Esto es un inconveniente porque no nos permite comparar la relación entre distintos pares de variables medidas en diferentes unidades. La solución es utilizar el **coeficiente de correlación lineal**, que consiste en tipificar la covarianza dividiéndola por las desviaciones típicas de ambas variables, y se calcula mediante,

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

La correlación lineal toma valores entre -1 y 1 y sirve para investigar la relación lineal entre las variables. Así, si toma valores cercanos a -1 diremos que tenemos una relación inversa entre X e Y (esto es, cuando una variable toma valores altos la otra toma valores bajos). Si toma valores cercanos a $+1$ diremos que tenemos una relación directa (valores altos de una variable en un individuo, asegura valores altos de la otra variable). Si toma valores cercanos a cero diremos que no existe relación lineal entre las variables. Cuando el valor de la correlación lineal sea exactamente 1 o -1 diremos que existe una dependencia exacta entre las variables mientras que si toma el valor cero diremos que son incorreladas.

Ejemplo 1: Para los datos del Ejemplo 1 se obtiene que la desviación típica de la estatura es $s_x = 14.422$ centímetros y la desviación típica del VEF es $s_y = 0.748$ litros. Por lo tanto, el coeficiente de correlación lineal será

$$r_{xy} = \frac{10.672}{14.422 \cdot 0.7488} = 0.9881$$

La correlación es próxima a 1 y por lo tanto la relación entre ambas variables es directa.

3 El modelo de regresión lineal

En el estudio de variables bidimensionales tiene mucho interés buscar posibles relaciones entre las variables. La más sencilla de estas relaciones es la dependencia lineal donde se supone que la relación entre dos variables X e Y viene dada por la ecuación $Y = \beta_0 + \beta_1 X$. Sin embargo, este modelo supone que una vez determinados los valores de los parámetros β_0 y β_1 es posible predecir exactamente la respuesta Y dado cualquier valor de la variable de entrada X. En la práctica tal precisión casi nunca es alcanzable, de modo que lo máximo que se puede esperar es que la ecuación anterior sea válida sujeta a un error aleatorio, es decir, la relación entre la **variable dependiente** (Y) y la **variable independiente** (X) se articula mediante una **recta de regresión**:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

En un modelo de regresión lineal $Y = \beta_0 + \beta_1 X + \epsilon$, la variable Y recibe el nombre de variable dependiente, respuesta o explicada. La variable X recibe el nombre de variable independiente, regresora o explicativa

3.1 El método de mínimos cuadrados

Dada una muestra $(x_1,y_1),\ldots,(x_n,y_n)$, el objetivo es determinar los valores de los parámetros desconocidos β_0 y β_1 (mediante estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$) de manera que la recta definida ajuste de la mejor forma posible a los datos. Aunque existen muchos métodos, el más clásico es el conocido como método de mínimos cuadrados que consiste en encontrar los valores de los parámetros que, dada la muestra de partida, minimizan la suma de los errores al cuadrado. Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ se determinan minimizando las distancias verticales entre los puntos observados, y_i, y las ordenadas previstas por la recta para dichos puntos $\hat{y}_i$. Es decir, el criterio será minimizar

$$M(\beta_0,\beta_1) = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2.$$

Los valores de los parámetros se obtienen, por tanto, derivando e igualando a cero. Se tiene:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

y

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

que serán llamados **coeficientes de la regresión**. De esta manera obtendremos la ecuación de la recta de regresión:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x = \bar{y} + \hat{\beta}_1 (x - \bar{x}) = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

que llamaremos **recta de regresión de Y sobre X** para resaltar que se ha obtenido suponiendo que Y es la variable respuesta y que X es la variable explicativa.

Intercambiando los papeles de X e Y obtendremos una recta de regresión llamada recta de regresión de X sobre Y, que representada en el mismo eje de coordenadas será en general distinta a la recta de regresión de Y sobre X. Solamente coincidirán en el caso de que la relación entre X e Y sea exacta.

Ejemplo 1: Volvamos al Ejemplo 1, donde se recogían datos de la estatura (X) y el VEF (Y). Los coeficientes de la recta de regresión de Y sobre X son:

$$\hat{\beta}_1 = \frac{10.672}{14.422^2} = 0.0513, \qquad \hat{\beta}_0 = 2.691 - 156 \cdot 0.0513 = -5.312$$

En la Figura 3 se muestra la recta de regresión de ecuación:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = -5.312 + 0.0513x$$

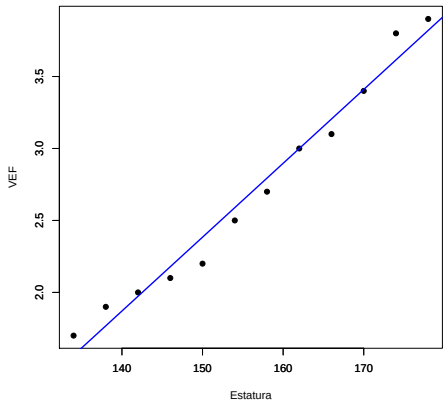


Figure 3: Recta de regresión $y = -5.312 + 0.0513x$ para los datos del Ejemplo 1

3.2 Descomposición de la variabilidad

Los métodos de regresión pretenden darnos una explicación de cómo la variable respuesta, Y, se comporta de distinta manera en función del valor que tome la variable explicativa, X. En consecuencia, parte de la variabilidad de Y quedaría justificada por la influencia de la variable X, mientras que otra parte sería fruto del error del modelo.

La variabilidad de toda la muestra la denominamos variabilidad total (VT) o suma total de cuadrados y se calcula como

$$\text{VT} = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

La variabilidad total se descompone en dos sumandos:

- El primero de ellos representa las desviaciones de las predicciones $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ respecto a la media global. Por tanto, sirve como medición de la variabilidad que podemos explicar en base al modelo de regresión. Se denomina variabilidad explicada (VE).

$$\text{VE} = \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y}\right)^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

- El segundo representa las desviaciones de los valores observados y_i respecto de las predicciones, y en consecuencia refleja la variabilidad no explicada (VNE) por la regresión, sino debida al error. Por ello se interpreta como variabilidad residual, se calcula mediante la suma de los residuos al cuadrado, denominada también como suma residual de cuadrados:

$$\text{VNE} = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Se tiene así la siguiente descomposición de la variabilidad del modelo de regresión:

$$\text{VT} = \text{VE} + \text{VNE}.$$

3.2.1 Coeficiente de determinación

Una vez resuelto el problema de estimar los parámetros surge la pregunta de si la recta estimada es o no representativa para los datos. Esto se resuelve mediante el **coeficiente de determinación** (R^2), que se define como la proporción de variabilidad de la variable dependiente que es explicada por la regresión. Se calcula como:

$$R^2 = \frac{\text{VE}}{\text{VT}} = 1 - \frac{\text{VNE}}{\text{VT}}.$$

En el modelo de regresión lineal simple, el coeficiente de determinación coincide con el cuadrado del coeficiente de correlación entre la variable explicativa y la variable respuesta, es decir

$$R^2 = r_{xy}^2$$

Ejemplo 1: Para los datos del Ejemplo 1 se puede observar que la recta de regresión no pasa por todos los puntos observados (ver Figura 3). Sin embargo, están muy próximos a ella, el grado de ajuste viene determinado por el coeficiente de determinación

$$R^2 = 0.9881^2 = 0.976$$

que se calcula como el cuadrado del coeficiente de correlación. Es decir, con el modelo de regresión lineal simple hallado, la variable X es capaz de explicar el 97.6% de la variación de Y.