Bioestadística. Curso 2012-2013
Grado en Medicina
Capítulo 1. Estadística descriptiva

Beatriz Pateiro López

Introducción

estadística.
(Del al. Statistik).
1. f. Estudio de los datos cuantitativos de la población, de los recursos naturales e industriales, del tráfico o de cualquier otra manifestación de las sociedades humanas.
2. f. Conjunto de estos datos.
3. f. Rama de la matemática que utiliza grandes conjuntos de datos numéricos para obtener inferencias basadas en el cálculo de probabilidades.

Diccionario de la lengua española. Real Academia Española

Introducción

La estadística es una ciencia con base matemática referente a la recolección, análisis e interpretación de datos, que busca explicar condiciones regulares en fenómenos de tipo aleatorio.
Es transversal a una amplia variedad de disciplinas, desde la física hasta las ciencias sociales, desde las ciencias de la salud hasta el control de calidad, y es usada para la toma de decisiones en áreas de negocios e instituciones gubernamentales.

Wikipedia

Introducción

- Se puede 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 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.

Introducción

Un ejemplo

- 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?

Conceptos básicos

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

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

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

Estadística

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.

Tipos de Variables

Variables cualitativas: No aparecen en forma numérica, sino como categorías o atributos.

- el sexo
- color de ojos
-
-

Variables cuantitativas: Toman valores numéricos porque son frecuentemente el resultado de una medición.

- el peso (kg.) de una persona
- número de llamadas diarias a un servicio de urgencias
-
-

Tipos de Variables. Variables cualitativas

Se clasifican a su vez en:

- **Cualitativas nominales:** Miden características que no toman valores numéricos. A estas características se les llama modalidades.
  - el sexo (hombre o mujer)
  - color de ojos (azul, verde, marrón,...)
  -
  -
- **Cualitativas ordinales:** Miden características que no toman valores numéricos pero sí presentan entre sus posibles valores una relación de orden.
  - 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.
  -
  -

Tipos de Variables. Variables cuantitativas

Se clasifican a su vez en:

- **Cuantitativas discretas:** Toman un número discreto de valores (en el conjunto de números naturales).
  - el número de hijos de una familia
  - número de cigarrillos fumados por día
  -
  -
- **Cuantitativas continuas:** Toman valores numéricos dentro de un intervalo real.
  - el peso
  - concentración de un elemento
  -
  -

Ejemplo

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

Ejemplo

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

¿Cómo resumimos la información contenida en los datos de la variable Dolor?

Descripción de variables cualitativas y cuantitativas discretas

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,

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

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

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

Descripción de variables cualitativas y cuantitativas discretas

Las frecuencias se pueden escribir ordenadamente mediante una **tabla de frecuencias**, que adopta esta forma:

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

Descripción de variables cualitativas y cuantitativas discretas

Las frecuencias se pueden escribir ordenadamente mediante una **tabla de frecuencias**, que adopta esta forma:

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	$0 \leq F_i \leq 1$	$F_m = 1$

Ejemplo

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

¿Cómo resumimos la información contenida en los datos de la variable Visitas?

Ejemplo

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

¿Cómo resumimos la información contenida en los datos de la variable Peso?

Descripción de variables cuantitativas continuas

- Para construir las frecuencias 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.

Descripción de variables cuantitativas continuas

Algunas consideraciones a tener en cuenta:

- *Número de intervalos a considerar:*
  - Cuantos menos intevalos tomemos, menos información se recoge.
  - Cuantos más intervalos tomemos, más difícil es manejar las frecuencias.

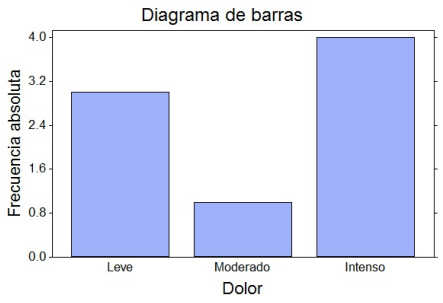
  Se suele tomar como número de intervalos el entero más próximo a $\sqrt{n}$.
- *Amplitud de cada intervalo:* Lo más común, salvo justificación en su contra, es tomar todos los intervalos de igual longitud.
- *Posición de los intervalos:* Los intervalos deben situarse allí donde se encuentran las observaciones y de forma contigua.

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.

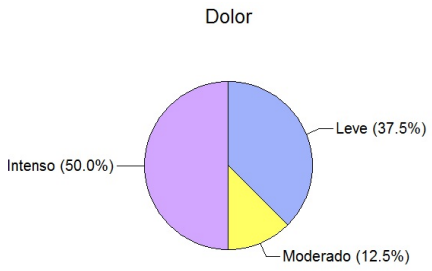
Representaciones gráficas de variables cualitativas

- Diagrama de barras: Representa frecuencias absolutas o relativas



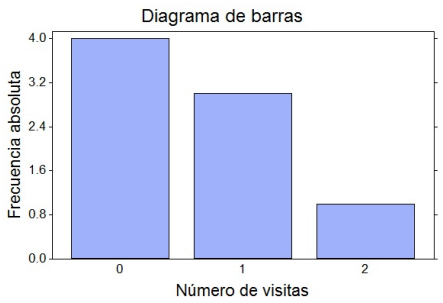
Representaciones gráficas de variables cualitativas

- 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.



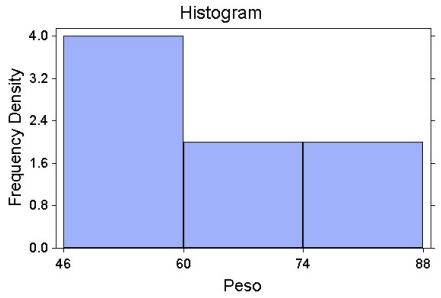
Representaciones gráficas de variables cuantitativas discretas

- Diagrama de barras: Representa frecuencias absolutas o relativas
- Diagrama de frecuencias acumuladas o diagrama escalonado: Representa frecuencias acumuladas absolutas o relativas



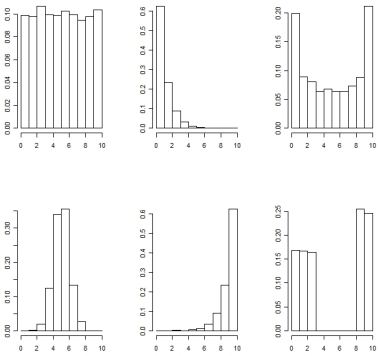
Representaciones gráficas de variables cuantitativas continuas

- 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.**



Interpretación del histograma

Una determinada operación de vesícula se puede realizar siguiendo seis técnicas distintas. Para cada técnica, hemos registrado el tiempo de postoperatorio de 100 pacientes sometidos a dicha operación. Los resultados aparecen resumidos en los siguientes histogramas.



Medidas características: Medidas de posición, de dispersión y de forma

Por **medida** entendemos un número que se calcula sobre la muestra y que refleja cierta cualidad de la misma. Parece claro que el cálculo de estas medidas requiere la posibilidad de efectuar operaciones con los valores que toma la variable. Por este motivo, en lo que resta del tema tratamos sólo con variables cuantitativas.

Medidas características: Medidas de posición, de dispersión y de forma

Por **medida** entendemos un número que se calcula sobre la muestra y que refleja cierta cualidad de la misma. Parece claro que el cálculo de estas medidas requiere la posibilidad de efectuar operaciones con los valores que toma la variable. Por este motivo, en lo que resta del tema tratamos sólo con variables cuantitativas.

- Medidas de posición: son medidas que nos indican la posición que ocupa la muestra
- Medidas de dispersión: se utilizan para describir la variabilidad o esparcimiento de los datos de la muestra respecto a la posición central
- Medidas de forma: tratan de medir el grado de simetría y apuntamiento en los datos

Medidas de posición

- Media aritmética
- Mediana
- Moda
- Cuantiles

Medidas de posición. 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:

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

Medidas de posición. 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.

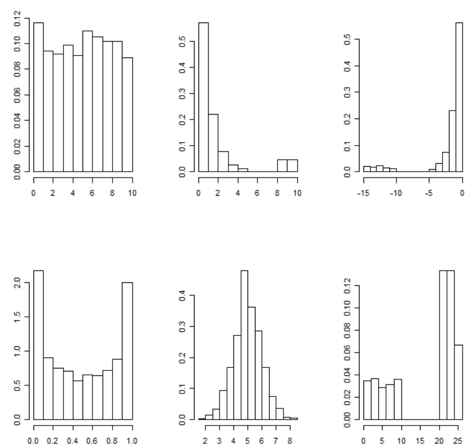
Medidas de posición. Moda

- 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.**

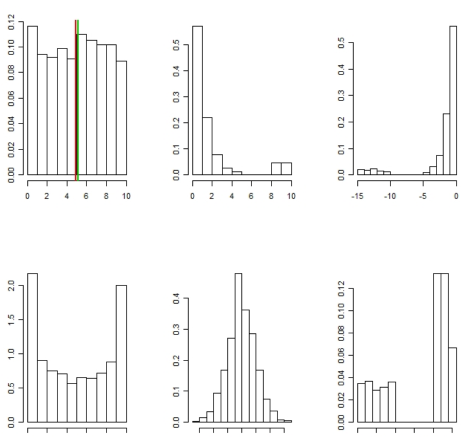
Medidas de posición. Cuantiles

- 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.
- Sea $p \in (0, 1)$. Se define el cuantil p como el número que deja a su izquierda una frecuencia relativa p. 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.
- 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.

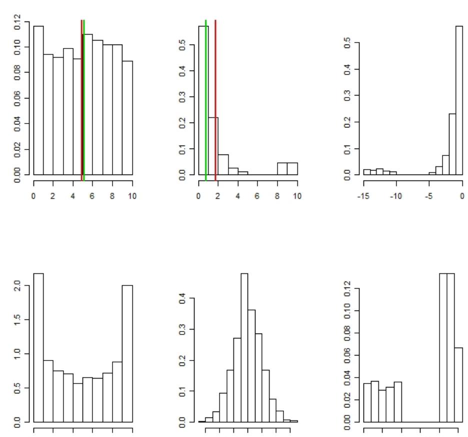
Medidas de posición.
¿Serías capaz de deducir cuál es aproximadamente la media y mediana de los conjuntos de datos con los siguientes histogramas?



Medidas de posición.
¿Serías capaz de deducir cuál es aproximadamente la media y mediana de los conjuntos de datos con los siguientes histogramas?

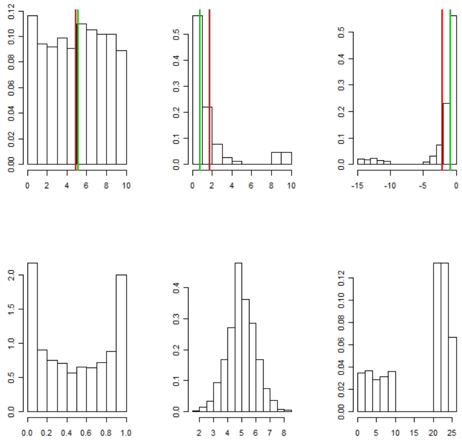


Medidas de posición.
¿Serías capaz de deducir cuál es aproximadamente la media y mediana de los conjuntos de datos con los siguientes histogramas?



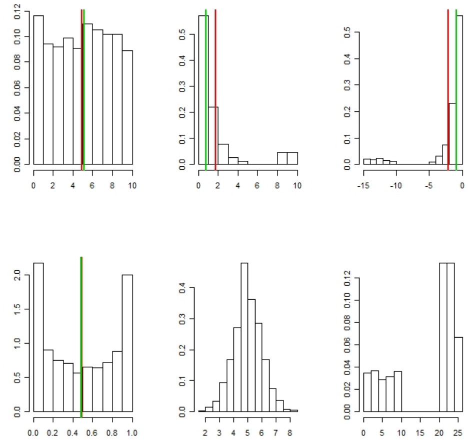
Medidas de posición.

¿Serías capaz de deducir cuál es aproximadamente la media y mediana de los conjuntos de datos con los siguientes histogramas?



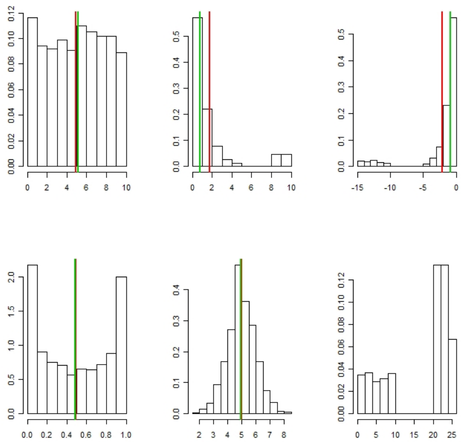
Medidas de posición.

¿Serías capaz de deducir cuál es aproximadamente la media y mediana de los conjuntos de datos con los siguientes histogramas?



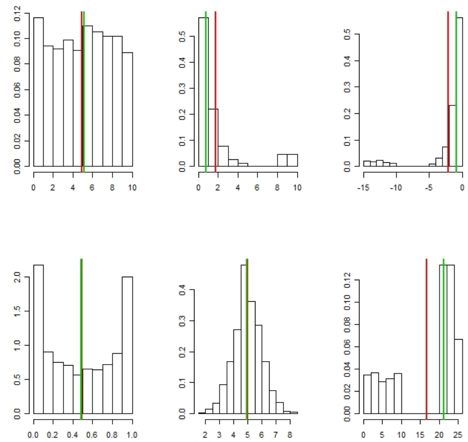
Medidas de posición.

¿Serías capaz de deducir cuál es aproximadamente la media y mediana de los conjuntos de datos con los siguientes histogramas?



Medidas de posición.

¿Serías capaz de deducir cuál es aproximadamente la media y mediana de los conjuntos de datos con los siguientes histogramas?



Medidas de dispersión

- Recorrido o rango
- Recorrido intercuartílico
- Varianza
- Desviación típica
- Coeficiente de variación

Medidas de dispersión. Recorrido o rango

- $R = \text{máx}\, x_i - \text{mín}\, x_i$.

Medidas de dispersión. Recorrido intercuártilico o rango intercuartílico

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

Medidas de dispersión. Varianza

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

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

Medidas de dispersión. Desviación típica

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

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

Coeficiente de variación

Hay situaciones en las que tenemos que comparar poblaciones en las que

- las unidades de medida son distintas

  Ejemplo:

  Peso de hormigas en gramos: ($s = 2{,}41$ gramos)

8.180881	10.503650	8.210198	13.096271	9.259044
15.540982	7.854185	12.010111	8.725924	11.712810

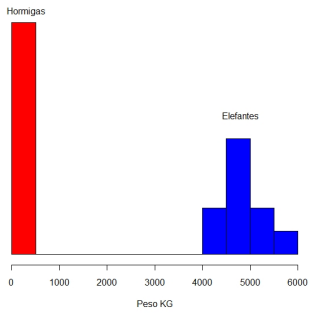
  Peso de elefantes en kg: ($s = 320{,}0495$ kilos)

5100.636	4987.702	5035.441	5321.591	5502.833
4737.402	4537.105	4731.434	4742.981	4444.282

Medidas de dispersión. Coeficiente de variación

Hay situaciones en las que tenemos que comparar poblaciones en las que

- o que aún teniendo la misma unidad de medida difieren en sus magnitudes.



Medidas de dispersión. Coeficiente de variación

- Hay situaciones en las que tenemos que comparar poblaciones en las que las unidades de medida son distintas, o que aún teniendo la misma unidad de medida difieren en sus magnitudes. Para estos casos necesitamos una medida de la dispersión en la que no influyan las unidades, sería conveniente tener una medida adimensional.
- 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.

Medidas de dispersión. Coeficiente de variación

Ejemplo:

Peso de hormigas en gramos: ($CV = 0{,}229$)

8.180881	10.503650	8.210198	13.096271	9.259044
15.540982	7.854185	12.010111	8.725924	11.712810

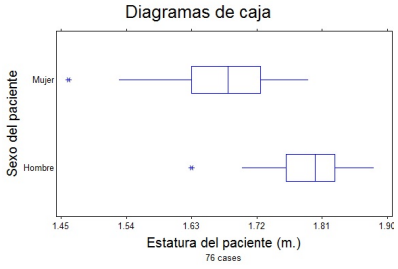
Peso de elefantes en kg: ($CV = 0{,}065$)

5100.636	4987.702	5035.441	5321.591	5502.833
4737.402	4537.105	4731.434	4742.981	4444.282

Diagramas de caja

Los diagramas de caja (boxplots) nos dan información visual sobre como están distribuidos los datos. El diagrama de caja consta de:

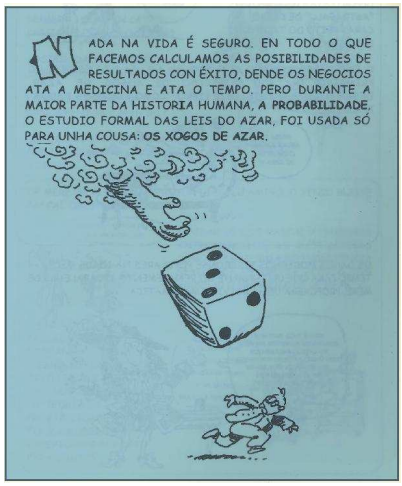
- una caja central delimitada por los cuartiles Q_1 y Q_3.
- Dentro de esa caja se dibuja la línea que representa la mediana (cuartil Q_2).
- De los extremos de la caja salen los bigotes que se extienden hasta los puntos $LI = \text{máx}\{\text{mín}(x_i), Q_1 - 1{,}5RI\}$ y $LS = \text{mín}\{\text{máx}(x_i), Q_3 + 1{,}5RI\}$
- Los datos que caen fuera de los bigotes se representan individualmente mediante "$*$" (datos atípicos moderados) y "o" (datos atípicos extremos).



Bioestadística. Curso 2012-2013
Grado en Medicina
Capítulo 2. Probabilidad

Beatriz Pateiro López

Introducción



A Estatística en caricaturas. Larry Gonick, Woollcott Smith

Introducción

Vinculada inicialmente a los juegos de azar, la probabilidad aparece siempre que queremos saber si algo va a ocurrir o no:

- ¿Cuál es la probabilidad de que salga un seis en una tirada de dado?

Introducción

Vinculada inicialmente a los juegos de azar, la probabilidad aparece siempre que queremos saber si algo va a ocurrir o no:

- ¿Cuál es la probabilidad de que salga un seis en una tirada de dado?
- ¿Cuál es la probabilidad de acertar los seis números de la lotería primitiva?

Introducción

Vinculada inicialmente a los juegos de azar, la probabilidad aparece siempre que queremos saber si algo va a ocurrir o no:

- ¿Cuál es la probabilidad de que salga un seis en una tirada de dado?
- ¿Cuál es la probabilidad de acertar los seis números de la lotería primitiva?
- ¿Cuál es la probabilidad de que me caiga en el examen un tema de los que tengo preparados?

Introducción

Vinculada inicialmente a los juegos de azar, la probabilidad aparece siempre que queremos saber si algo va a ocurrir o no:

- ¿Cuál es la probabilidad de que salga un seis en una tirada de dado?
- ¿Cuál es la probabilidad de acertar los seis números de la lotería primitiva?
- ¿Cuál es la probabilidad de que me caiga en el examen un tema de los que tengo preparados?
- ¿Cuál es la probabilidad de que un paciente sobreviva a una determinada operación de trasplante?

Introducción

Vinculada inicialmente a los juegos de azar, la probabilidad aparece siempre que queremos saber si algo va a ocurrir o no:

- ¿Cuál es la probabilidad de que salga un seis en una tirada de dado?
- ¿Cuál es la probabilidad de acertar los seis números de la lotería primitiva?
- ¿Cuál es la probabilidad de que me caiga en el examen un tema de los que tengo preparados?
- ¿Cuál es la probabilidad de que un paciente sobreviva a una determinada operación de trasplante?
- Y si el paciente sobrevive a la operación, ¿cuál es la probabilidad de que su cuerpo rechace el trasplante en menos de un mes?

Introducción

La mayoría de la gente tiene una noción de lo que significa la probabilidad de que algo ocurra:

Introducción

La mayoría de la gente tiene una noción de lo que significa la probabilidad de que algo ocurra:

- Las probabilidades son números comprendidos entre 0 y 1 que reflejan las expectativas de que un suceso ocurra.
- Probabilidades próximas a 1 indican que cabe esperar que ocurran los sucesos en cuestión.
- Probabilidades próximas a 0 indican que no cabe esperar que ocurran los sucesos en cuestión.
- Probabilidades próximas a 0.5 indican que es tan verosímil que ocurra el suceso como que no.

Conceptos básicos

- Experimento aleatorio
- Espacio muestral
- Suceso

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**.
- 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.

Experimento aleatorio

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.

Experimento aleatorio

Ejemplos de experimentos aleatorios son:

- $\mathcal{E}_1$ =*Lanzar una moneda al aire*
- $\mathcal{E}_2$ =*Lanzar dos veces una moneda*
- $\mathcal{E}_3$ =*Determinar la temperatura corporal*

Espacio muestral

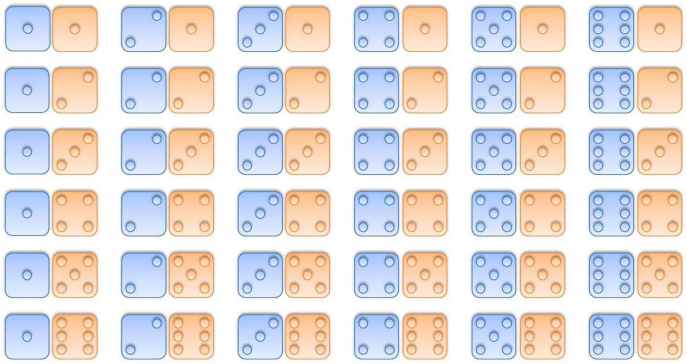
- Llamamos **espacio muestral** al conjunto formado por todos los resultados posibles del experimento aleatorio. Lo denotamos por Ω.

Sucesos elementales

- **Suceso elemental:** Un suceso elemental es cada uno de los posibles resultados $\omega \in \Omega$ del experimento aleatorio.

Sucesos elementales

Consideremos ahora el experimento $\mathcal{E}$ =*Lanzar un par de dados*
Este espacio muestral tiene 36 (6×6) sucesos elementales.



Sucesos

- **Suceso:** Cualquier subconjunto del espacio muestral.

Sucesos

- 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**.

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. Escrito en notación conjuntista, resulta $A \cap B = \emptyset$.
- **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

La intervención quirúrgica de colocación de prótesis de rodilla se realiza mediante anestesia general o epidural. Durante la intervención se realiza una incisión en la rodilla para cortar y extraer parcialmente uno de los huesos (fémur, tibia o peroné) en la zona próxima a la rodilla, y a continuación se sustituye por la prótesis, que puede ser de metal o resina.

Intervención	Posibilidades
Anestesia	General o epidural
Hueso	Fémur, tibia o peroné
Prótesis	Metal o resina

- Indica el espacio muestral de posibles condiciones (anestesia, hueso y prótesis) en las que se realizan las intervenciones de colocación de prótesis.
- Si A es el suceso consistente en que la intervención se realiza con prótesis de metal, lista los elementos de A.
- Si B es el suceso consistente en que la intervención se realiza con anestesia general, lista los elementos de B.
- ¿Cuáles son los elementos de $A \cap B$?
- Si C es el suceso consistente en que la intervención se realiza con anestesia epidural, lista los elementos de $B \cup C$.
- ¿Cuáles son los elementos de $B \cap C$?
- Si D es el suceso consistente en que la intervención se realiza con extracción parcial del fémur, y E es el suceso consistente en que la intervención se realiza con extracción parcial del peroné, lista los elementos de $C \cap (D \cup E)$.

Definición de probabilidad

Una vez definido un experimento aleatorio, se trata de asignar un **peso numérico** o **probabilidad** a cada suceso que mida su grado de ocurrencia.

Definición clásica o de Laplace

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**

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

La Teoría de la Probabilidad no es, en el fondo, más que sentido común reducido a cálculo.
(Laplace, Théorie Analytique des Probabilités)

Un ejemplo

1. Una clase de primaria está formada por 60 niñas y 40 niños. Se observa que 26 niñas y 14 niños usan gafas. Si un estudiante es elegido al azar, ¿cuál es la probabilidad de que use gafas?

Definición axiomática de Kolmogorov

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.

Definición axiomática de Kolmogorov

A partir de la definición anterior se pueden sacar una serie de consecuencias:

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
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Un ejemplo

Una tabla de contingencia clasica es la presentada por Sir Ronald Fisher en 1940, que presenta la clasificación de 5387 escolares escoceses según su color de pelo y color de ojos.

$X \backslash Y$	rubio	pelirrojo	castaño	oscuro	negro	
claros	688	116	584	188	4	1580
azules	326	38	241	110	3	718
castaños	343	84	909	412	26	1774
oscuros	98	48	403	681	85	1315
total	1455	286	2137	1391	118	5387

Cuadro: Color de ojos y el color del pelo (Fisher, 1940)

Se elige una persona de la clase al azar

1. ¿Cuál es la probabilidad de que la persona elegida tenga ojos castaños?
2. ¿Cuál es la probabilidad de que la persona elegida tenga pelo rubio?
3. ¿Cuál es la probabilidad de que la persona elegida tenga ojos castaños o pelo rubio?
4. ¿Cuál es la probabilidad de que la persona elegida tenga ojos castaños y pelo rubio?
5. ¿Cuál es la probabilidad de que la persona elegida tenga pelo castaño o pelo rubio?

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? ¿Cuál es la probabilidad de sacar un 1 al lanzar un dado si sabemos que el resultado ha sido un número impar?

Probabilidad condicionada

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

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

Probabilidad condicionada

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

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

Un ejemplo

Volvemos al ejemplo de Fisher de clasificación de 5387 escolares escoceses según su color de pelo y color de ojos.

$X \backslash Y$	rubio	pelirrojo	castaño	oscuro	negro	
claros	688	116	584	188	4	1580
azules	326	38	241	110	3	718
castaños	343	84	909	412	26	1774
oscuros	98	48	403	681	85	1315
total	1455	286	2137	1391	118	5387

Cuadro: Color de ojos y el color del pelo (Fisher, 1940)

Se elige una persona de la clase al azar

1. ¿Cual es la probabilidad de que una persona con ojos castaños tenga pelo rubio?
2. ¿Cuál es la probabilidad de que una persona con ojos oscuros tenga pelo rubio?

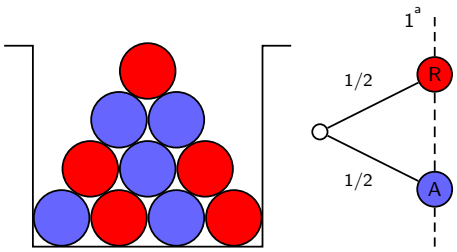
Resultados importantes en Teoría de la Probabilidad

- Regla del producto.
- Ley de las probabilidades totales
- Regla de Bayes

La regla del producto

La regla del producto es muy útil en experimentos aleatorios que tienen varias etapas. Las diversas etapas y alternativas se suelen representar en un diagrama de árbol tal como se muestra en el siguiente ejemplo.

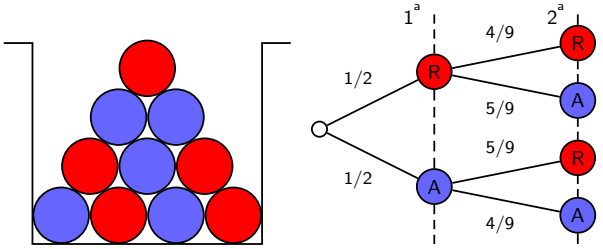
Ejemplo: En la urna de la figura se extraen (sin reemplazamiento) dos bolas. Calcula la probabilidad de que las dos sean rojas



La regla del producto

La regla del producto es muy útil en experimentos aleatorios que tienen varias etapas. Las diversas etapas y alternativas se suelen representar en un diagrama de árbol tal como se muestra en el siguiente ejemplo.

Ejemplo: En la urna de la figura se extraen (sin reemplazamiento) dos bolas. Calcula la probabilidad de que las dos sean rojas



La regla del producto

La 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

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

Un ejemplo en medicina de la regla del producto

1. La probabilidad de sobrevivir a cierta operación de trasplante es 0.55. Si un paciente sobrevive a la operación, la probabilidad de que su cuerpo rechace el trasplante en menos de un mes es 0.2. ¿Cuál es la probabilidad de que sobreviva a estas etapas críticas?

Independencia de sucesos

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

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

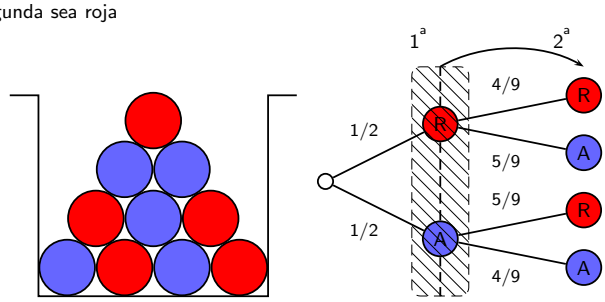
Comentarios:

- 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**

La ley de las probabilidades totales

La ley de las probabilidades totales considera todas las ramas que llegan al resultado final observado.

Ejemplo: Calcula la probabilidad de al extraer dos bolas (sin reemplazamiento) la segunda sea roja



Ley de las probabilidades totales

A menudo, la probabilidad de ocurrencia de un suceso B se calcula más facilmente en términos de probabilidades condicionadas. La idea es encontrar una sucesion de sucesos mutuamente excluyentes como se indica a continuación.

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)
- 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:

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

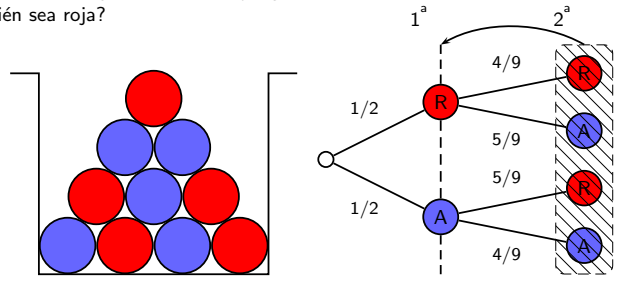
Un ejemplo en medicina de la ley de probabilidades totales

3. La probabilidad de que una unidad de sangre proceda de un donante remunerado es 0.67. Si el donante es remunerado, la probabilidad de que la unidad contenga el suero de la hepatitis es 0.0144. Si el donante es desinteresado, esta probabilidad es 0.0012. Un paciente recibe una unidad de sangre. ¿Cuál es la probabilidad de que contraiga hepatitis como consecuencia de ello?

Teorema de Bayes

Los resultados de un experimento dan información sobre lo que ocurrió en las etapas intermedias.

Ejemplo: Si la segunda bola es roja, ¿cuál es la probabilidad de que la primera también sea roja?



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, $\quad i = 1, \ldots, n$.

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, $\quad i = 1, \ldots, n$.

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

Teorema de Bayes. En las condiciones anteriores,

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

Teorema de Bayes

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:

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

Este teorema resulta de aplicar en el numerador la regla del producto y en el denominador la ley de probabilidades totales.

Un ejemplo en medicina del Teorema de Bayes

1. Volvemos al ejemplo de la transfusión de sangre. Un paciente recibe una unidad de sangre y contrae hepatitis. ¿Cuál es la probabilidad de que la unidad de sangre utilizada en la transfusión proceda de un paciente remunerado?

Pruebas diagnósticas: Sensibilidad y especificidad. Prevalencia e incidencia.

- 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.

Prevalencia e incidencia

Prevalencia: La prevalencia es la proporción de individuos de la población que presentan la enfermedad. Se calcula dividiendo el número de personas que sufren la enfermedad objeto de estudio entre el número total de individuos examinados.

Prevalencia e incidencia

Prevalencia: La prevalencia es la proporción de individuos de la población que presentan la enfermedad. Se calcula dividiendo el número de personas que sufren la enfermedad objeto de estudio entre el número total de individuos examinados.

- Por ejemplo, en un estudio sobre incontinencia se examinó a un total de 6139 individuos de los cuales 519 sufrían incontinencia. La prevalencia de la enfermedad en ese momento es:

$$P(E) = \frac{519}{6139} = 0.085$$

- Según datos de 2008, la prevalencia del VIH en adultos en Europa occidental y central es del 0.3 %
- Según datos de 2008, la prevalencia del VIH en adultos en África subsahariana es del 5.2 %

Prevalencia e incidencia

Incidencia: La incidencia es una 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.

Prevalencia e incidencia

Incidencia: La incidencia es una 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.

- La incidencia (incidencia acumulada) se calcula como el número de nuevos casos de la enfermedad objeto de estudio en un período específico de tiempo dividido entre el tamaño de la población que inicialmente estaba sana. Por ejemplo, durante un período de 1 año se siguió a 525 mujeres sanas, con colesterol y tensión arterial normal, para detectar la presencia de cardiopatía isquémica, registrándose al final del período 15 casos de cardiopatía isquémica. La incidencia acumulada en este caso sería:

$$\text{IA} = \frac{15}{525} = 0.028 \text{ en un año.}$$

Pruebas diagnósticas: Sensibilidad y especificidad. Prevalencia e incidencia.

- 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.
- 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 el paciente no está enfermo.

Un **falso negativo** resulta cuando una prueba indica que el estado es negativo, cuando en realidad el paciente está enfermo.

Pruebas diagnósticas: Sensibilidad y especificidad. Prevalencia e incidencia.

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?

Pruebas diagnósticas: Sensibilidad y especificidad. Prevalencia e incidencia.

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.

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

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.

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

Sensibilidad de una prueba diagnóstica

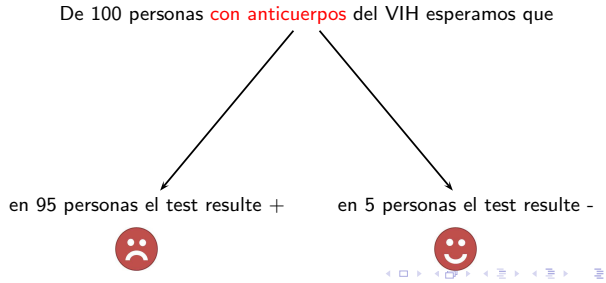
La sensibilidad de un determinado test de anticuerpos del VIH es del 95 %.

$$P(+/E) = 0.95$$

Sensibilidad de una prueba diagnóstica

La sensibilidad de un determinado test de anticuerpos del VIH es del 95 %.

$$P(+/E) = 0.95$$



Dado que un individuo no tiene la enfermedad, ¿qué probabilidad existe de que la prueba resulte negativa?

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.

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

Especificidad de una prueba diagnóstica

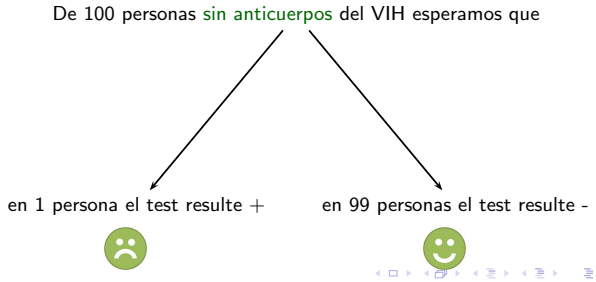
La especificidad de un determinado test de anticuerpos del VIH es del 99 %.

$$P(-/S) = 0.99$$

Especificidad de una prueba diagnóstica

La especificidad de un determinado test de anticuerpos del VIH es del 99 %.

$$P(-/S) = 0.99$$



Dado un resultado positivo de una prueba de detección, ¿qué probabilidad existe de que el individuo tenga la enfermedad?

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.

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

Dado un resultado positivo de una prueba de detección, ¿qué probabilidad existe de que el individuo tenga la enfermedad?

Teniendo en cuenta que la prevalencia del VIH en adultos en África subsahariana es del 5.2 %, ¿cuál es el valor predictivo positivo en dicha población de un determinado test de anticuerpos del VIH cuya sensibilidad es del 95 % y cuya especificidad es del 99 %?

Dado un resultado negativo de una prueba de detección, ¿qué probabilidad existe de que el individuo no tenga la enfermedad?

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.

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

Dado un resultado negativo de una prueba de detección, ¿qué probabilidad existe de que el individuo no tenga la enfermedad?

Teniendo en cuenta que la prevalencia del VIH en adultos en África subsahariana es del 5.2 %, ¿cuál es el valor predictivo negativo en dicha población de un determinado test de anticuerpos del VIH cuya sensibilidad es del 95 % y cuya especificidad es del 99 %?

Algunas cuestiones importantes

- Hemos visto que los valores de sensibilidad y especificidad definen la validez de la prueba diagnóstica. Sin embargo no proporcionan información relevante a la hora de tomar una decisión sobre el estado de salud del paciente.
- La sensibilidad y especificidad son propiedades intrínsecas a la prueba diagnóstica (independientes de la prevalencia de la enfermedad).
- Los valores predictivos (positivo y negativo) dependen de la prevalencia.

Algunas cuestiones importantes

Teniendo en cuenta que la prevalencia del VIH en adultos en Europa es del 0.3 %, ¿cuáles son los valores predictivos positivo y negativo en dicha población de un determinado test de anticuerpos del VIH cuya sensibilidad es del 95 % y cuya especificidad es del 99 %?

Bioestadística. Curso 2012-2013
Grado en Medicina
Capítulo 3. Variables aleatorias discretas

Beatriz Pateiro López

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.

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.

Variable aleatoria

Al realizar un experimento aleatorio generalmente estamos interesados en alguna función del resultado más que en el resultado en sí mismo. Por ejemplo, al arrojar un dado dos veces podríamos estar interesados sólo en la suma de los puntos obtenidos y no en el par de valores que dio origen a ese valor de la suma. De manera informal, esa cantidad de interés se denomina variable aleatoria.

- Variable porque toma distintos valores
- aleatoria porque el valor observado no puede ser predicho antes de la realización del experimento, aunque sí se sabe cuáles son sus posibles valores.

Dado que el valor de una variable aleatoria (v.a.) es determinado por el resultado de un experimento, podremos asignar probabilidades a los posibles valores o conjuntos de valores de la variable.

Variable aleatoria

Definición

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.

Variables aleatorias

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, será discreta si dichos valores se encuentran separados entre sí. Por tanto será representable por conjuntos discretos, como $\mathbb{Z}$ o $\mathbb{N}$. Para dichas variables veremos:
  - Función de probabilidad o función de masa
  - Función de distribución
- Será continua cuando el conjunto de valores que puede tomar es un intervalo. Para dichas variables veremos:
  - Función de densidad
  - Función de distribución

Variables aleatorias discretas. Función de probabilidad

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. Esta función se denomina **función de probabilidad o función de masa**. La función de probabilidad se puede representar análogamente al diagrama de barras.

Variables aleatorias discretas. Función de probabilidad

Ejemplo: Los servicios médicos de un equipo de fútbol establecen un período de entre 7 y 9 días de baja para un futbolista que ha sufrido una fuerte contusión en el tríceps sural. Además se estima que

- La probabilidad de que el período de baja sea de 7 días es 0.4.
- La probabilidad de que el período de baja sea de 8 días es 0.5.
- La probabilidad de que de que el período de baja sea de 9 día es 0.1.

Comprueba que se trata efectivamente de una distribución de probabilidad y a represéntala.

Variables aleatorias discretas. Función de distribución

Definición

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

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

Variables aleatorias discretas. Función de distribución

Ejemplo: Los servicios médicos de un equipo de fútbol establecen un período de entre 7 y 9 días de baja para un futbolista que ha sufrido una fuerte contusión en el tríceps sural. Además se estima que

- La probabilidad de que el período de baja sea de 7 días es 0.4.
- La probabilidad de que el período de baja sea de 8 días es 0.5.
- La probabilidad de que de que el período de baja sea de 9 día es 0.1.

Calcula y representa la función de distribución. Interpreta los resultados.

Variables aleatorias discretas. Función de distribución

Suponiendo que la variable X toma los valores $x_1 \leq x_2 \leq ... \leq x_n$, la función de distribución viene definida por:

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

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

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

Medidas características de una variable aleatoria.

- 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.

Media y varianza de variables aleatorias.

Para distinguir entre las propiedades de los conjuntos de datos y las de las distribuciones de probabilidad, usaremos cierta terminología y ciertos símbolos que describimos a continuación.

- Las propiedades de los datos se llaman **propiedades muestrales**. Por ejemplo, hablamos en el tema 1 de la media muestral $\bar{x}$ o de la desviación típica muestral s.
- Las propiedades de las distribuciones de probabilidad se llaman **propiedades poblacionales**.
  - Usaremos la letra griega μ para denotar la **media poblacional**.
  - Usaremos la letra griega σ para denotar la **desviación típica poblacional**.

Media y Varianza poblacional de una variable aleatoria discreta.

- Consideremos el ejemplo del futbolista que ha sufrido una fuerte contusión en el tríceps sural. Estamos interesados en el número de días de baja del jugador.

x_i	p_i
7	0.4
8	0.5
9	0.1

Media y Varianza poblacional de una variable aleatoria discreta.

- Consideremos el ejemplo del futbolista que ha sufrido una fuerte contusión en el tríceps sural. Estamos interesados en el número de días de baja del jugador.

x_i	p_i
7	0.4
8	0.5
9	0.1

- ¿Cómo definirías el número medio (o número esperado) de días que el jugador pasará de baja?

$$\mathbb{E}(X) = \mu = \sum_i x_i p_i = 7 \cdot 0{,}4 + 8 \cdot 0{,}5 + 9 \cdot 0{,}1 = 7{,}7$$

- ¿Cómo definirías la varianza de la variable X?

$$\text{Var}(X) = \sigma^2 = \sum_i (x_i - \mu)^2 p_i = (7-7{,}7)^2 \cdot 0{,}5 + (8-7{,}7)^2 \cdot 0{,}5 + (9-7{,}7)^2 \cdot 0{,}1 = 0{,}41$$

Propiedades de la media y varianza de una variable aleatoria discreta.

Propiedades

Sea X una variable aleatoria discreta con valores x_i. Entonces:

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

Propiedades de la media y varianza de una variable aleatoria discreta.

- Consideremos el ejemplo del futbolista que ha sufrido una fuerte contusión en el tríceps sural. Por cada lesión que sufre el jugador el seguro le debe pagar 5000 euros, además de 1000 euros por cada día de baja. ¿Cuánto dinero espera recibir el jugador del seguro?

Principales modelos de distribuciones discretas

- Estudiaremos distribuciones de variables aleatorias **discretas** que han adquirido una especial relevancia por ser adecuadas para modelizar una gran cantidad de situaciones.
- Caracterizaremos estas distribuciones mediante la función de masa y función de distribución.
- Calcularemos también los momentos (media y varianza) y destacaremos las propiedades de mayor utilidad.

Principales modelos de distribuciones discretas: Variable Bernoulli

Variable 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, recuperarse o no recuperarse de una enfermedad...)
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 $q = 1 - p$. Un experimento de este tipo se llama **experimento de Bernoulli** $Be(p)$.

Principales modelos de distribuciones discretas: Variable Bernoulli

Variable 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, recuperarse o no recuperarse de una enfermedad...)
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 $q = 1 - p$. Un experimento de este tipo se llama **experimento de Bernoulli** $Be(p)$.

- Calcula la función de masa y la función de distribución de una $Be(p)$.
- Si $X \in Be(p)$, entonces:
  - $\mu = p$
  - $\sigma^2 = p(1-p)$

Principales modelos de distribuciones discretas: Variable Binomial

Ejemplo: Una pareja descubre que la probabilidad de que un hijo de la pareja sufra una determinada enfermedad genética es 0.6. Si la pareja se plantea tener tres hijos, ¿cuál es la probabilidad de que exactamente uno de ellos sufra la enfermedad genética?
Cada hijo es independiente de los demás y podemos considerarlo como un ensayo de Bernoulli, donde el éxito es estar sano ($p = 0{,}4$). Lo que hacemos es repetir el experimento 3 veces y queremos calcular la probabilidad de que el número de éxitos sea igual a 2 (es decir, 2 hijos sanos y 1 enfermo)

Principales modelos de distribuciones discretas: Variable Binomial

Variable 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.
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

Principales modelos de distribuciones discretas: Variable Binomial

Variable Binomial

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

$$X = \text{Número de éxitos en las } n \text{ pruebas}$$

Denotaremos esta variable como $Bin(n, p)$.

Principales modelos de distribuciones discretas: Variable Binomial

Variable Binomial

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

$$X = \text{Número de éxitos en las } n \text{ pruebas}$$

Denotaremos esta variable como $Bin(n, p)$.

- ¿Qué valores toma una $Bin(n, p)$?
- ¿Cuál es su función de masa?

Principales modelos de distribuciones discretas: Variable Binomial

Variable Binomial

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

$$X = \text{Número de éxitos en las } n \text{ pruebas}$$

La probabilidad de obtener k éxitos en n pruebas es

$$P(X = k) = \left(\begin{array}{c} n \\ k \end{array} \right) \cdot p^k \cdot (1-p)^{n-k}$$

El coeficiente binomial

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

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).

Coeficientes binomiales

El coeficiente binomial

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

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).

Por ejemplo, si para un partido de dobles de la Copa Davis tenemos a tres jugadores ({Robredo, Feliciano López, Verdasco}), el entrenador tendrá

$$\left(\begin{array}{c} 3 \\ 2 \end{array} \right) = \frac{3!}{2!1!} = 3$$

posibles formas de elegir a los jugadores del partido ({Robredo, Feliciano López}, {Robredo, Verdasco}, {Feliciano López, Verdasco}).

Principales modelos de distribuciones discretas: Poisson

- En muchas circunstancias (llamadas a una centralita telefónica de un hospital, número de leucocitos en una gota de sangre, ...) 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).

Coeficientes binomiales

El coeficiente binomial

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

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).

Principales modelos de distribuciones discretas: Variable Binomial

Variable Binomial

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

$$X = \text{Número de éxitos en las } n \text{ pruebas}$$

- La media y la varianza de una $Bin(n, p)$ son:
  - $\mu = n \cdot p$
  - $\sigma^2 = n \cdot p \cdot (1-p)$

Principales modelos de distribuciones discretas: Poisson

Variable Poisson

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

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

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

- $\mu = \lambda$
- $\sigma^2 = \lambda$

Principales modelos de distribuciones discretas: Poisson

- Utilizaremos la distribución de Poisson como aproximación de la distribución binomial cuando n sea grande y p pequeño, en base al límite que hemos visto.
- Como criterio podremos aproximar cuando $n > 50$ y $p < 0{,}1$.

Principales modelos de distribuciones discretas: Poisson

Ejemplo
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?

Principales modelos de distribuciones discretas: Poisson

Ejemplo
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

$$P(X = 18) = \begin{pmatrix} 3000 \\ 18 \end{pmatrix} \cdot 0{,}005^{18} \cdot 0{,}995^{2982} = 0{,}07071.$$

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:

$$P(X = 18) \approx P(Poisson(15) = 18) = e^{-15}\frac{15^{18}}{18!} = 0{,}07061.$$

Principales modelos de distribuciones discretas: Poisson

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.

Principales modelos de distribuciones discretas: Poisson

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:

- 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.

Principales modelos de distribuciones discretas: Poisson

Ejemplo
El número de nacimientos en un hospital constituye un proceso de Poisson con intensidad de 10 nacimientos por semana. ¿Cuál es la probabilidad de que se produzcan al menos tres nacimientos en una semana?

Ejemplo

El número de nacimientos en un hospital constituye un proceso de Poisson con intensidad de 10 nacimientos por semana. ¿Cuál es la probabilidad de que se produzcan al menos tres nacimientos en una semana?

$$P(X \geq 3) = 1 - P(X < 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$

Principales modelos de distribuciones discretas: Poisson

Ejemplo

El número de nacimientos en un hospital constituye un proceso de Poisson con intensidad de 10 nacimientos por semana. ¿Cuál es la probabilidad de que se produzcan al menos tres nacimientos en una semana?

$$\begin{aligned} P(X \geq 3) &= 1 - P(X < 3) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)] \\ &= 1 - \left[e^{-10} \frac{10^0}{0!} + e^{-10} \frac{10^1}{1!} + e^{-10} \frac{10^2}{2!} \right] \end{aligned}$$

¿Cuál es la probabilidad de que se produzcan 5 nacimientos un día?

Bioestadística. Curso 2012-2013
Grado en Medicina
Capítulo 4. Variables aleatorias continuas

Beatriz Pateiro López

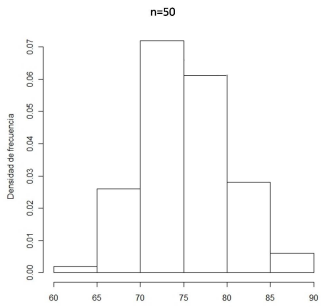
Variables aleatorias continuas

- Una variable aleatoria es **continua** cuando puede tomar cualquier valor en un intervalo.
  - el peso de una persona
  - el contenido de paracetamol en un lote de pastillas
  - el tiempo de recuperación de una operación,...
- 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.

Variables aleatorias continuas

Ejemplo
En un estudio sobre atención a la tercera edad se desea evaluar la edad a la que las personas mayores deciden ingresar en un centro geriátrico.

- Se registra la edad a la que ingresaron los 50 residentes de un determinado centro gerontológico y se construye el histograma correspondiente.

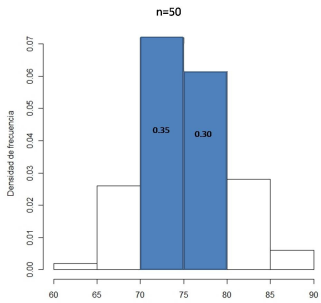


Variables aleatorias continuas

Ejemplo
En un estudio sobre atención a la tercera edad se desea evaluar la edad a la que las personas mayores deciden ingresar en un centro geriátrico.

- Se registra la edad a la que ingresaron los 50 residentes de un determinado centro gerontológico y se construye el histograma correspondiente.

Sea A el suceso "El residente ingresa con edad entre 70 y 80 años".

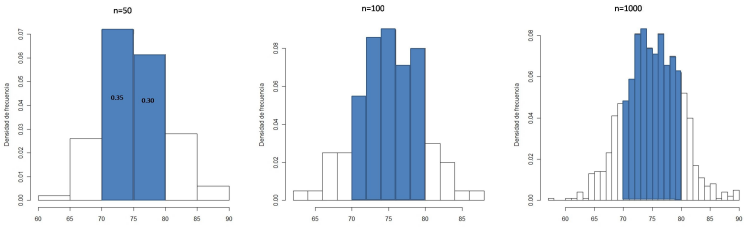


Variables aleatorias continuas

Ejemplo
En un estudio sobre atención a la tercera edad se desea evaluar la edad a la que las personas mayores deciden ingresar en un centro geriátrico.

- Se registra la edad a la que ingresaron los 50 residentes de un determinado centro gerontológico y se construye el histograma correspondiente.
- Se registra la edad a la que ingresaron los 100 residentes de un determinado centro gerontológico y se construye el histograma correspondiente.
- Se registra la edad a la que ingresaron los 1000 residentes de un determinado centro gerontológico y se construye el histograma correspondiente.

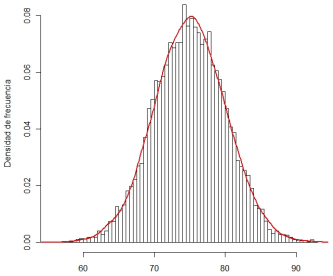
Sea A el suceso "El residente ingresa con edad entre 70 y 80 años".



Variables aleatorias continuas

Ejemplo
En un estudio sobre atención a la tercera edad se desea evaluar la edad a la que las personas mayores deciden ingresar en un centro geriátrico.

- Idealmente, se registra la edad de todos los residentes de centros gerontológicos y se construye el histograma correspondiente.

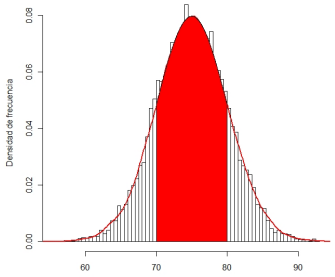


Variables aleatorias continuas

Ejemplo
En un estudio sobre atención a la tercera edad se desea evaluar la edad a la que las personas mayores deciden ingresar en un centro geriátrico.

- Idealmente, se registra la edad de todos los residentes de centros gerontológicos y se construye el histograma correspondiente.

Sea A el suceso "El residente ingresa con edad entre 70 y 80 años".

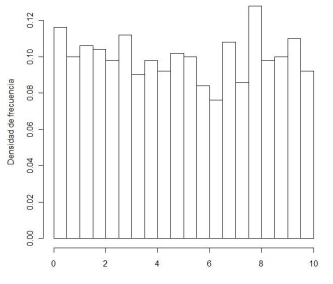


Variables aleatorias continuas

- 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.
- 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.
- La función de densidad cumple dos propiedades básicas: es no negativa y el área total que contiene es uno.

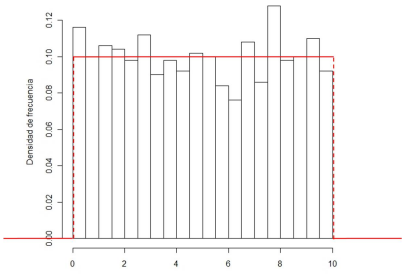
Variables aleatorias continuas. Función de densidad

Ejemplo
Un estudiante va todos los días a la facultad en la línea 1 del autobús urbano. Llega a la parada a las 3 de la tarde y cuenta el tiempo (en minutos) que tiene que esperar hasta que llega el autobús. A continuación se muestra el histograma correspondiente al tiempo de espera de los últimos 1000 días. A la vista del histograma, ¿cómo modelizarías el tiempo de espera?



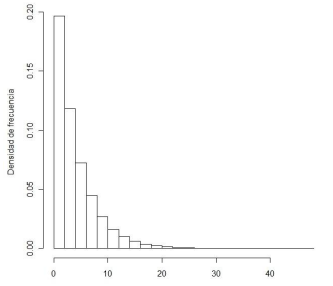
Variables aleatorias continuas. Función de densidad

Ejemplo
Un estudiante va todos los días a la facultad en la línea 1 del autobús urbano. Llega a la parada a las 3 de la tarde y cuenta el tiempo (en minutos) que tiene que esperar hasta que llega el autobús. A continuación se muestra el histograma correspondiente al tiempo de espera de los últimos 1000 días. A la vista del histograma, ¿cómo modelizarías el tiempo de espera?



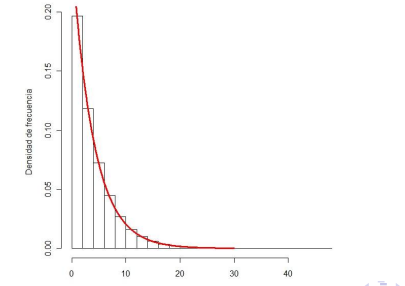
Variables aleatorias continuas. Función de densidad

Ejemplo
Un estudiante va todos los días a la facultad en la línea 6 del autobús urbano. Llega a la parada a las 3 de la tarde y cuenta el tiempo (en minutos) que tiene que esperar hasta que llega el autobús. A continuación se muestra el histograma correspondiente al tiempo de espera de los últimos 1000 días. A la vista del histograma, ¿cómo modelizarías el tiempo de espera?



Variables aleatorias continuas. Función de densidad

Ejemplo
Un estudiante va todos los días a la facultad en la línea 6 del autobús urbano. Llega a la parada a las 3 de la tarde y cuenta el tiempo (en minutos) que tiene que esperar hasta que llega el autobús. A continuación se muestra el histograma correspondiente al tiempo de espera de los últimos 1000 días. A la vista del histograma, ¿cómo modelizarías el tiempo de espera?



Variables aleatorias continuas: Función de densidad

Una función $f(x)$, definida sobre el conjunto de todos los números reales $\mathbb{R}$, se denomina **función de densidad** si

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

Variables aleatorias continuas. Función de distribución

Definición

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

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

Variables aleatorias continuas: Función de densidad

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,

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

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

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

Momentos poblacionales de una variable aleatoria continua.

Propiedades

Sea X una variable aleatoria continua con función de densidad $f(x)$. Entonces:

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

Principales modelos de distribuciones continuas

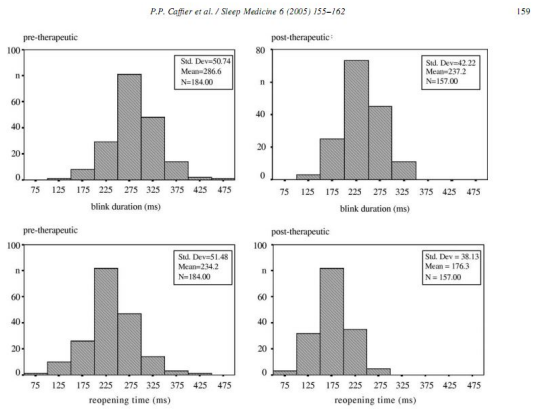


Fig. 2. Example of individual histograms of blink duration and reopening time of pre- and post-therapeutic measurement of one male OSA patient with EDS.

The spontaneous eye-blink as sleepiness indicator in patients with obstructive sleep apnoea syndrome-a pilot study. Sleep Medicine 6 (2005) 155-162.

Principales modelos de distribuciones continuas

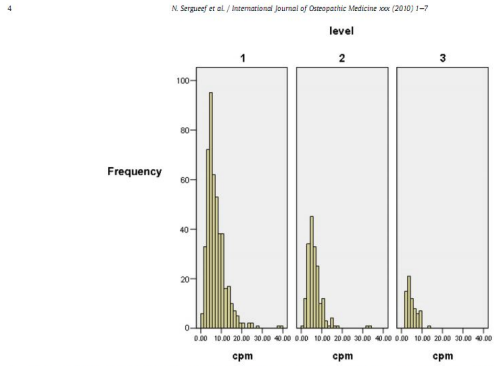


Fig. 2. Frequency histograms of cranial rhythmic impulse (CRI) rates partitioned by level of training. Level 1, 1 yr ($N = 463$); Level 2, 2 yr ($N = 190$); Level 3, 3–25 yr ($N = 74$). Each bar represents a CRI range of 1.43 counts per minute (cpm); frequency is the number of the total 727 participants in a given range for each training level.

The palpated cranial rhythmic impulse (CRI): Its normative rate and examiner experience. International Journal of Osteopathic Medicine (2010)

Principales modelos de distribuciones continuas

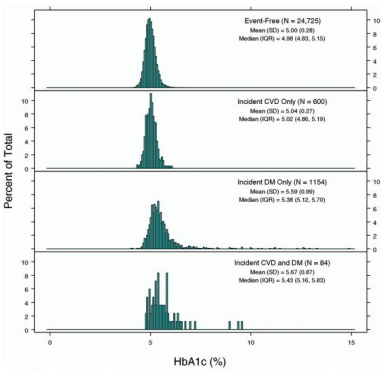


Figure 1 Histograms of HbA1c distribution according to 4 main groups: individuals remaining disease-free (N = 24,725), developing incident cardiovascular disease only (N = 600), developing incident diabetes mellitus only (N = 1154), or developing both cardiovascular disease and diabetes mellitus (N = 84). SD = standard deviation; IQR = interquartile range; CVD = cardiovascular disease; DM = diabetes mellitus; HbA1c = hemoglobin A1c.

Hemoglobin A1c Predicts Diabetes but Not Cardiovascular Disease in Nondiabetic Women.

The American Journal of Medicine (2007)

Principales modelos de distribuciones continuas

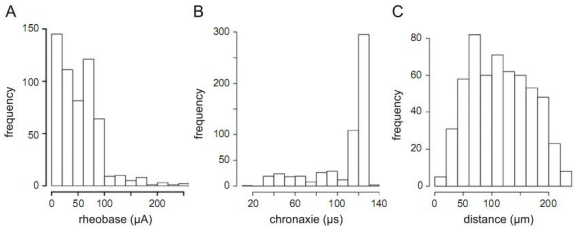


Fig. 5. Strength-duration characteristics of generated axons. (A) Histogram of rheobase amplitudes. (B) Histogram of chronaxie times. (C) Histogram of distance between the site of action potential initiation and the center of the stimulating electrode.

Modeling potential generation during single and dual electrode stimulation of CA3 axons in hippocampal slice.

Computers in Biology and Medicine (2010)

Principales modelos de distribuciones continuas

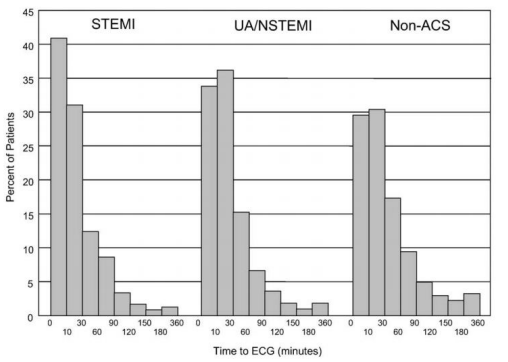


Fig. 1 Histogram showing the time to ECG for each of the patient groups.

Door-to-ECG time in patients with chest pain presenting to the ED.

American Journal of Emergency Medicine (2006)

Principales modelos de distribuciones continuas

for the between-run precision. For the recovery tests with three spiked pool-serums, values between 100.9% and 104.7% and a maximal RSD of 1.6% were obtained. The limit of detection and the limit of quantification were determined at 0.5 and 1.1 μg/L, respectively, by measuring the diluent as a blank solution [15].

In order to check the performance of the analysis method during the whole measuring period, an aliquot of the pool-serum sample was quantified in each measuring series. A mean concentration ± standard deviation (S.D.) of 104.4 ± 3.3 μg/L ($n = 93$) was thereby obtained. By plotting these values in a Quality Control Chart, no drift effect was observed visually.

Statistical method

The statistical evaluation was carried out with the software Systat version 10. Influences of the *age*, *gender* and *region/area* variables on the selenium concentrations of the blood donors and the subjects from the medical practice were evaluated by the analysis of variance (one-way ANOVA). The Scheffé post hoc test

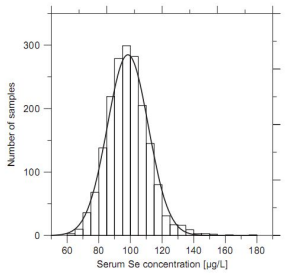


Fig. 1. Histogram of the serum selenium concentrations. The bar width is 5 μg/L; the line is a normal profile adjusted to the measured values.

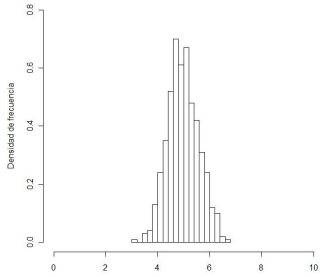
Selenium status of the Swiss population: Assessment and change over a decade.

Journal of Trace Elements in Medicine and Biology (2008)

Principales modelos de distribuciones continuas: Variable Normal

Ejemplo

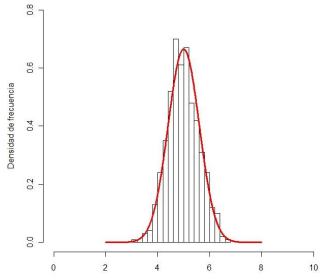
Un centro hospitalario dispone de 3 máquinas de electrocardiograma (máquina de ECG). A continuación se muestra el histograma correspondiente al tiempo (medido en minutos) de 500 registros de la actividad eléctrica del corazón producidos con la **primera máquina**. A la vista del histograma. ¿cómo modelizarías el tiempo de registro de la primera máquina?



Principales modelos de distribuciones continuas: Variable Normal

Ejemplo

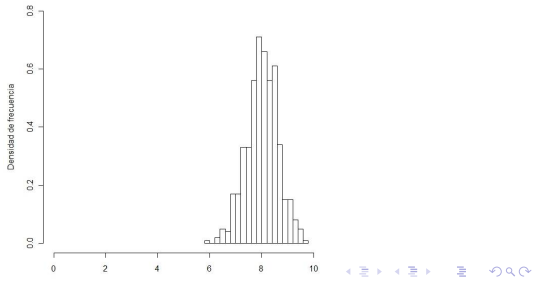
Un centro hospitalario dispone de 3 máquinas de electrocardiograma (máquina de ECG). A continuación se muestra el histograma correspondiente al tiempo (medido en minutos) de 500 registros de la actividad eléctrica del corazón producidos con la **primera máquina**. A la vista del histograma. ¿cómo modelizarías el tiempo de registro de la primera máquina?



Principales modelos de distribuciones continuas: Variable Normal

Ejemplo

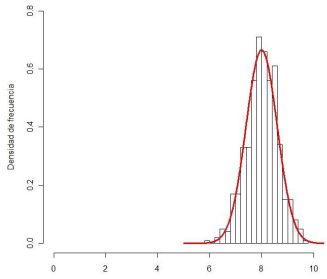
Un centro hospitalario dispone de 3 máquinas de electrocardiograma (máquina de ECG). A continuación se muestra el histograma correspondiente al tiempo (medido en minutos) de 500 registros de la actividad eléctrica del corazón producidos con la **segunda máquina**. A la vista del histograma. ¿cómo modelizarías el tiempo de registro de la segunda máquina?



Principales modelos de distribuciones continuas: Variable Normal

Ejemplo

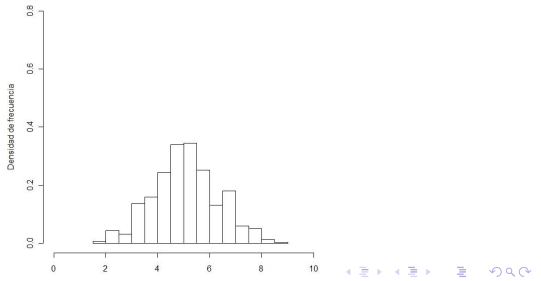
Un centro hospitalario dispone de 3 máquinas de electrocardiograma (máquina de ECG). A continuación se muestra el histograma correspondiente al tiempo (medido en minutos) de 500 registros de la actividad eléctrica del corazón producidos con la **segunda máquina**. A la vista del histograma. ¿cómo modelizarías el tiempo de registro de la segunda máquina?



Principales modelos de distribuciones continuas: Variable Normal

Ejemplo

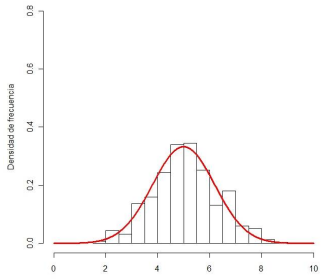
Un centro hospitalario dispone de 3 máquinas de electrocardiograma (máquina de ECG). A continuación se muestra el histograma correspondiente al tiempo (medido en minutos) de 500 registros de la actividad eléctrica del corazón producidos con la **tercera máquina**. A la vista del histograma. ¿cómo modelizarías el tiempo de registro de la tercera máquina?



Principales modelos de distribuciones continuas: Variable Normal

Ejemplo

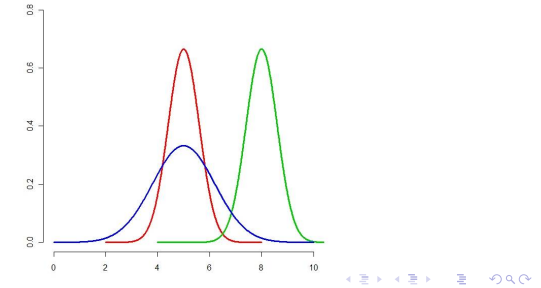
Un centro hospitalario dispone de 3 máquinas de electrocardiograma (máquina de ECG). A continuación se muestra el histograma correspondiente al tiempo (medido en minutos) de 500 registros de la actividad eléctrica del corazón producidos con la **tercera máquina**. A la vista del histograma. ¿cómo modelizarías el tiempo de registro de la tercera máquina?



Principales modelos de distribuciones continuas: Variable Normal

Ejemplo

Un centro hospitalario dispone de 3 máquinas de electrocardiograma (máquina de ECG). Supongamos que modelizamos el tiempo de registro de la tres máquinas mediante las siguientes curvas. ¿Qué tienen en común dichas curvas? ¿Qué las diferencia?



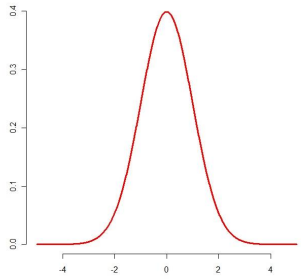
Principales modelos de distribuciones continuas: Variable Normal

- 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 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.

Variable Normal Estándar

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

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



Principales modelos de distribuciones continuas: Variable Normal

Variable Normal Estándar

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

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

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

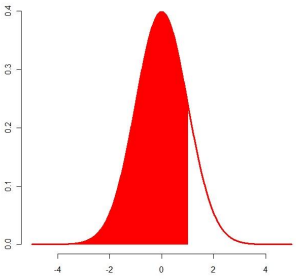
Principales modelos de distribuciones continuas: Variable Normal

Supongamos entonces que $Z \in N(0,1)$. ¿Cómo calcularías $P(Z \leq 1)$?

Principales modelos de distribuciones continuas: Variable Normal

Supongamos entonces que $Z \in N(0,1)$. ¿Cómo calcularías $P(Z \leq 1)$?

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



Principales modelos de distribuciones continuas: Variable Normal

Supongamos entonces que $Z \in N(0,1)$. ¿Cómo calcularías $P(Z \leq 1)$?

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

- 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, $\Phi(z_0) = P(Z \leq z_0)$, de la normal estándar para valores positivos de z, donde z está aproximado hasta el segundo decimal.

Principales modelos de distribuciones continuas: Variable Normal

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 $P(Z > 4{,}2)$?

Variable Normal

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.

Si $Z \in N(0,1)$ entonces

$$X = \mu + \sigma Z$$

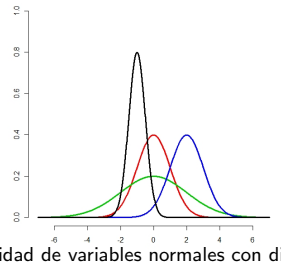
tiene **distribución normal de media μ y desviación típica σ**.
Denotaremos $X \in N(\mu, \sigma)$.

- Si $X \in N(\mu, \sigma)$ entonces la media de X es μ y su varianza es σ^2.

Variable Normal

Sea $X \in N(\mu, \sigma)$. La función de densidad de una $N(\mu, \sigma)$ es

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



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

Supongamos entonces que $X \in N(\mu, \sigma)$. ¿Cómo calcularías $P(X \leq 1)$?

Supongamos entonces que $X \in N(\mu, \sigma)$. ¿Cómo calcularías $P(X \leq 1)$?

$$P(X \leq 1) = \int_{-\infty}^{1} f(x)dx = \int_{-\infty}^{1} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

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.

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

Ejemplo

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

Ejemplo

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

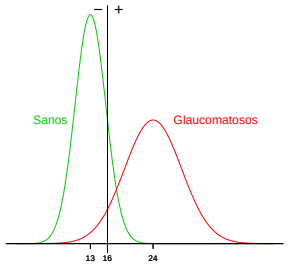
$$P(X \leq 1) = P\left(\frac{X-5}{2} \leq \frac{1-5}{2}\right) = P(Z \leq -2)$$

donde $Z = \frac{X-5}{2} \in N(0,1)$.

Puntos de corte para el diagnóstico de enfermedades

Ejemplo: Queremos estudiar la capacidad diagnóstica de la tonometría ocular en el diagnóstico del glaucoma.

- Se establece como criterio diagnóstico una cifra de tensión ocular de 16mmHg.
- Los estudios determinan que la tensión ocular en pacientes sanos se distribuye como una normal de media 13mmHg y desviación típica 2.7mmHg.
- También se establece que la tensión ocular en pacientes glaucomatosos se distribuye como una normal de media 24mmHg y desviación típica 5mmHg.

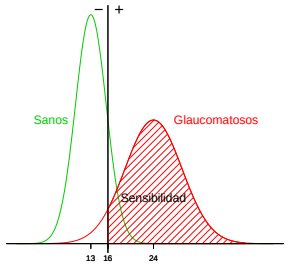


1. ¿Cuál es la sensibilidad y la especificidad de la prueba si el punto de corte es 16mmHg?
2. ¿Cuál es la probabilidad de falso positivo? ¿Y la de falso negativo?

Puntos de corte para el diagnóstico de enfermedades

Ejemplo: Queremos estudiar la capacidad diagnóstica de la tonometría ocular en el diagnóstico del glaucoma.

- Se establece como criterio diagnóstico una cifra de tensión ocular de 16mmHg.
- Los estudios determinan que la tensión ocular en pacientes sanos se distribuye como una normal de media 13mmHg y desviación típica 2.7mmHg.
- También se establece que la tensión ocular en pacientes glaucomatosos se distribuye como una normal de media 24mmHg y desviación típica 5mmHg.



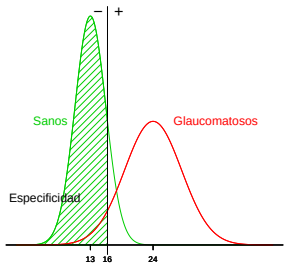
1. ¿Cuál es la sensibilidad y la especificidad de la prueba si el punto de corte es 16mmHg?
2. ¿Cuál es la probabilidad de falso positivo? ¿Y la de falso negativo?

Bioestadística. Grado en Medicina. Beatriz Pateiro López
Capítulo 4. Variables aleatorias continuas

Puntos de corte para el diagnóstico de enfermedades

Ejemplo: Queremos estudiar la capacidad diagnóstica de la tonometría ocular en el diagnóstico del glaucoma.

- Se establece como criterio diagnóstico una cifra de tensión ocular de 16mmHg.
- Los estudios determinan que la tensión ocular en pacientes sanos se distribuye como una normal de media 13mmHg y desviación típica 2.7mmHg.
- También se establece que la tensión ocular en pacientes glaucomatosos se distribuye como una normal de media 24mmHg y desviación típica 5mmHg.



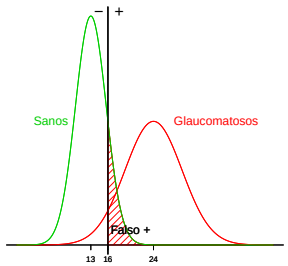
1. ¿Cuál es la sensibilidad y la especificidad de la prueba si el punto de corte es 16mmHg?
2. ¿Cuál es la probabilidad de falso positivo? ¿Y la de falso negativo?

Bioestadística. Grado en Medicina. Beatriz Pateiro López
Capítulo 4. Variables aleatorias continuas

Puntos de corte para el diagnóstico de enfermedades

Ejemplo: Queremos estudiar la capacidad diagnóstica de la tonometría ocular en el diagnóstico del glaucoma.

- Se establece como criterio diagnóstico una cifra de tensión ocular de 16mmHg.
- Los estudios determinan que la tensión ocular en pacientes sanos se distribuye como una normal de media 13mmHg y desviación típica 2.7mmHg.
- También se establece que la tensión ocular en pacientes glaucomatosos se distribuye como una normal de media 24mmHg y desviación típica 5mmHg.



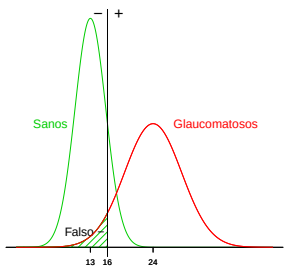
1. ¿Cuál es la sensibilidad y la especificidad de la prueba si el punto de corte es 16mmHg?
2. ¿Cuál es la probabilidad de falso positivo? ¿Y la de falso negativo?

Bioestadística. Grado en Medicina. Beatriz Pateiro López
Capítulo 4. Variables aleatorias continuas

Puntos de corte para el diagnóstico de enfermedades

Ejemplo: Queremos estudiar la capacidad diagnóstica de la tonometría ocular en el diagnóstico del glaucoma.

- Se establece como criterio diagnóstico una cifra de tensión ocular de 16mmHg.
- Los estudios determinan que la tensión ocular en pacientes sanos se distribuye como una normal de media 13mmHg y desviación típica 2.7mmHg.
- También se establece que la tensión ocular en pacientes glaucomatosos se distribuye como una normal de media 24mmHg y desviación típica 5mmHg.



1. ¿Cuál es la sensibilidad y la especificidad de la prueba si el punto de corte es 16mmHg?
2. ¿Cuál es la probabilidad de falso positivo? ¿Y la de falso negativo?

Bioestadística. Grado en Medicina. Beatriz Pateiro López
Capítulo 4. Variables aleatorias continuas

Bioestadística. Curso 2012-2013
Grado en Medicina
Capítulo 5. Inferencia estadística

Beatriz Pateiro López

Introducción

- Nuestro objetivo es el estudio de una población y sus características.
- Llamaremos **parámetro** a una característica numérica que nos interese conocer de la población.

  Ejemplos:
  - la presión sistólica **media** de una población,
  - nivel de colesterol **medio**,
  - **proporción** de pacientes que responden satisfactoriamente a un medicamento para la diabetes,...
- En la práctica contaremos con una muestra representativa de la población.

Introducción

- Capítulo 1: conceptos básicos de Estadística Descriptiva, que nos proporcionaban herramientas para resumir, ordenar y extraer los aspectos más relevantes de la información de la muestra.
- Capítulo 2: bases para trabajar con incertidumbres o probabilidades.
- Capítulos 3 y 4: principales modelos de variables aleatorias.

Introducción

- Capítulo 1: conceptos básicos de Estadística Descriptiva, que nos proporcionaban herramientas para resumir, ordenar y extraer los aspectos más relevantes de la información de la muestra.
- Capítulo 2: bases para trabajar con incertidumbres o probabilidades.
- Capítulos 3 y 4: principales modelos de variables aleatorias.

INFERENCIA ESTADÍSTICA

Ahora podremos empezar a hacer inferencia sobre la población de interés basándonos en lo que observamos en una muestra

Introducción

Dependiendo de los objetivos, podremos clasificar las labores de inferencia en dos grandes categorías:

1ª) en la que el interés se centra en **estimar o aproximar el valor de un parámetro**

Ejemplo: la proporción de pacientes que responden a un determinado medicamento para la diabetes

2ª) en la que el interés se centra en **contrastar posibles valores de un parámetro**

Ejemplo: determinar si el nivel de colesterol medio en hombres es superior al nivel de colesterol medio en mujeres

Introducción

Los sondeos no dan un ganador claro en las elecciones en Reino Unido

A tres días para los comicios los conservadores se mantienen en cabeza en las encuestas con un 33% de los votos, seguidos por liberaldemócratas y laboristas con el 28%

Internacional | 03/05/2010 - 12:43h | 05/05/2010 - 18:21h

1 | 1 visita | Compartir | Notificar error | Tengo más información | A A

LONDRES. (EUROPA PRESS).- A falta de tres días para las elecciones generales en **Reino Unido**, los últimos sondeos publicados este lunes no ofrecen un ganador claro, si bien los conservadores se mantienen en cabeza, seguidos por liberaldemócratas y laboristas.

Según el sondeo de **ICM** para **'The Guardian'**, que publica hoy el diario, los conservadores obtendrían el 33% de los votos, igual que hace una semana, mientras que los laboristas se mantienen en el 28% y los liberaldemócratas caen dos puntos y empatan con el partido gobernante.

Suspense total a 15 días de las elecciones legislativas británicas

Compartir 0 Twittear

Escrito por Denis Hiault LONDRES (AFP)

Viernes, 23 abril 2010 13:00

Recomendar Sé el primero de tus amigos en recomendar esto.

El suspense parece total a menos de dos semanas de unas elecciones generales británicas que inicialmente se anunciaban como un trámite para los conservadores, debido al avance de los liberal demócratas y a la volatilidad de los votantes que favorecen un parlamento sin mayoría absoluta.

Los analistas destacaban unánimemente este viernes la multitud de escenarios posibles cuando los tres principales partidos se mueven en una horquilla que va del 27% al 32% de intención de voto: corta victoria de los 'Tories' de David Cameron pese a una campaña considerada trabajosa hasta en su propio campo; mantenimiento del impopular Gordon Brown gracias a un sistema electoral muy favorable a los laboristas, tanto que un tercer lugar podría garantizarle el mayor número de diputados; formación de alianzas o de un gobierno de coalición en el que los "Lib Dems" tendrían la llave. O incluso la ausencia de una solución viable, lo que obligaría a convocar nuevas elecciones.

La Voz de Galicia.es

PORTADA GALICIA DEPORTES SOCIEDAD DINERO ESPAÑA MUNDO OPINIÓN BLOGS OCIO Y CULTURA SERVICIOS TIE

A Coruña A Mariña Arousa Barbanza Carballo Deza Ferrol Lemos Lugo Ourense Pontevedra Santiago

El consello destaca que el 99% de los viticultores cumplieron la ley

18/3/2011

Valoración (0 votos) Menéame Twittear 0 Me gusta

Rías Baixas quiso dejar ayer muy claro que el número de viticultores que incumple la normativa y vendimia más de lo permitido es anecdótico. De hecho, la denominación de origen afirma que más del 99% de los agricultores cumplieron con la normativa durante la vendimia del 2009. Además, el consello argumentó que la pertenencia a la denominación de origen es una cuestión voluntaria.

Público.es

Iniciar sesi

Regist

Portada Opinión Internacional España Catalunya Dinero Ciencias Culturas Deportes TV

La ONU afirma que menos del 5% de la población mundial es adicta a drogas ilegales

EFE | Viena | 10/03/2008 16:14 | Recomendar

1. **Estimación Puntual**. Consiste en aventurar un valor, calculado a partir de la muestra, que esté lo más próximo posible al verdadero parámetro.
2. **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.
3. **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.

¿En qué problema de inferencia enmarcarías las siguientes noticias?

1. El insomnio, que es la falta de sueño a la hora de dormir, afecta entre un 10 y 20 % de la población general, pero se dispara hasta 32 % en los mayores de 65 años.
2. El resultado del síndrome de piernas inquietas es una interrupción del sueño que puede dar lugar a insomnio y somnolencia diurna. La prevalencia de este trastorno aumenta con la edad, estimándose que lo padecen entre un 10 y un 20 % de los mayores de 65 años.
3. Según un estudio el 25 % de la población sufre problemas mentales por la situación económica. El mismo estudio afirma que el 40 % de la población utiliza el alcohol para evadirse de la situación económica. Sin embargo, hay otros análisis que dudan de la veracidad de dichas conclusiones.

- 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.
- Al valor del estadístico obtenido con una realización muestral concreta se le llama **estimación**.
- Un estadístico es también una variable aleatoria y por ello tendrá una cierta distribución, que se denomina **distribución del estadístico en el muestreo**.

Teorema Central del Límite

El siguiente resultado nos permitirá calcular la distribución en el muestreo de muchos estadísticos de interés.

Teorema Central del Límite

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.

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

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

- la distribución χ^2
- la distribución t de Student.

La distribución χ^2

La χ^2_n **con** n **grados de libertad** es otro modelo de variable aleatoria continua

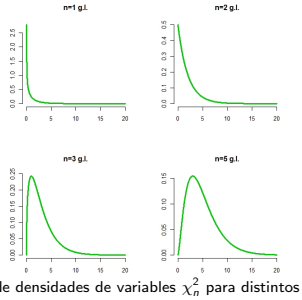


Figura : En verde densidades de variables χ^2_n para distintos valores de n.

Propiedades.

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

La distribución t de Student

La t **de Student con** k **grados de libertad** es otro modelo de variable aleatoria continua

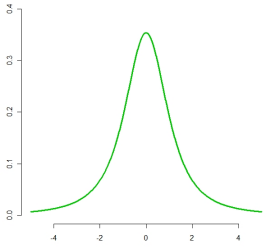


Figura : En verde densidad de una t de Student con 2 grados de libertad

La distribución t de Student

La t **de Student con** k **grados de libertad** es otro modelo de variable aleatoria continua como los vistos en el tema anterior.

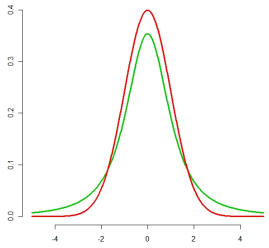


Figura : En verde densidad de una t de Student con 2 grados de libertad y en rojo densidad de una N(0,1)

La distribución t de Student

La t **de Student con** k **grados de libertad** es otro modelo de variable aleatoria continua como los vistos en el tema anterior.

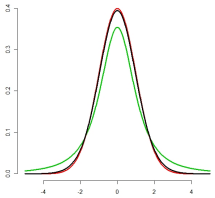


Figura : En verde densidad de una t de Student con 2 grados de libertad, en rojo N(0,1) y en negro densidad de una t de Student con 20 grados de libertad

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 $k \to \infty$.

Bioestadística. Curso 2012-2013
Grado en Medicina
Capítulo 6. Estimación puntual e Intervalos de confianza

Beatriz Pateiro López

Introducción

- **Estimación Puntual**. Consiste en aventurar un valor, calculado a partir de la muestra, que esté lo más próximo posible al verdadero parámetro.
- **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.
- **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.

Introducción

- **Estimación Puntual**. Consiste en aventurar un valor, calculado a partir de la muestra, que esté lo más próximo posible al verdadero parámetro.

Estimación puntual (de una proporción)

Sea $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$ una muestra aleatoria simple donde

$$\mathrm{X_i} = \begin{cases} 1 & \text{, con probabilidad } p \\ 0 & \text{, con probabilidad } 1-p \end{cases}$$

Estimación puntual de una proporción p

$$\hat{p} = \frac{\mathrm{X}_1 + \mathrm{X}_2 + \ldots + \mathrm{X}_n}{n}$$

Para n grande, por el Teorema Central de Límite:

Distribución de p̂

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

Estimación puntual (de una media)

Sea $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$ una muestra aleatoria simple con $X_i \sim N(\mu, \sigma)$.

Estimación puntual de la media μ

$$\bar{X} = \frac{\mathrm{X}_1 + \mathrm{X}_2 + \ldots + \mathrm{X}_n}{n}$$

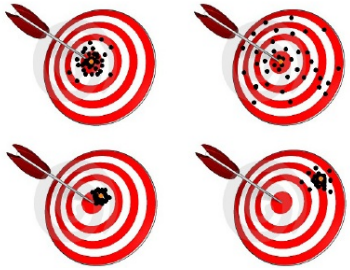
Entonces,

Distribución de $\bar{X}$

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

Propiedades de un estimador

Supongamos que queremos estimar un parámetro desconocido θ y lo hacemos mediante el estadístico $\hat{\theta}$



- $\hat{\theta}$ es **insesgado** si $\mathbb{E}(\hat{\theta}) = \theta$
- Si además $\mathrm{Var}(\hat{\theta}) \to 0$ cuando $n \to \infty$, el estimador es **consistente**

Intervalo de confianza

- **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.
- 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**.

Intervalo de confianza

- Sea θ el parámetro desconocido y $\alpha \in [0, 1]$.
- Se dice que el intervalo $[L_1, L_2]$ tiene un nivel de confianza $1 - \alpha$ si

$$P(L_1 \leq \theta \leq L_2) \geq 1 - \alpha$$

Intervalo de confianza

- Sea θ el parámetro desconocido y $\alpha \in [0, 1]$.
- Se dice que el intervalo $[L_1, L_2]$ tiene un nivel de confianza $1 - \alpha$ si

$$P(L_1 \leq \theta \leq L_2) \geq 1 - \alpha$$

- Los valores de L_1 y L_2 **dependerán de la muestra!!!!**.

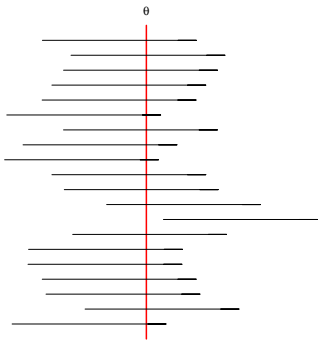
Intervalo de confianza

- Sea θ el parámetro desconocido y $\alpha \in [0, 1]$.
- Se dice que el intervalo $[L_1, L_2]$ tiene un nivel de confianza $1 - \alpha$ si

$$P(L_1 \leq \theta \leq L_2) \geq 1 - \alpha$$

- Los valores de L_1 y L_2 **dependerán de la muestra!!!!**.
- 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 con una probabilidad de 0,95.

Interpretación del nivel de confianza $1 - \alpha$



- Dada una realización muestral, el intervalo construido puede contener o no al parámetro desconocido
- Esperamos que el $100(1 - \alpha)$ % de los intervalos contengan al parámetro desconocido

Intervalo de confianza para la media μ de una población normal (σ^2 conocida)

Sea $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$

- Recordamos que es este caso

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

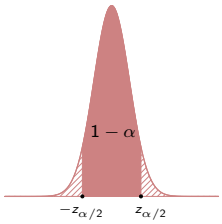
- Este estadístico (pivote) nos servirá para construir un intervalo de confianza con nivel de confianza $1 - \alpha$ para la media μ cuando la **varianza σ^2 es conocida**.

Intervalo de confianza para la media μ de una población normal (σ^2 conocida)

Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$. Supongamos que σ^2 es conocida.

- Sea $z_{\alpha/2}$ el valor tal que $P(Z > z_{\alpha/2}) = \alpha/2$, siendo $Z \in N(0,1)$. Entonces:

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



Intervalo de confianza para la media μ de una población normal (σ^2 conocida)

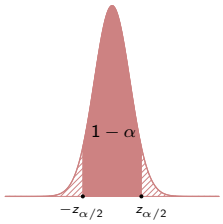
Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$. Supongamos que σ^2 es conocida.

- Sea $z_{\alpha/2}$ el valor tal que $P(Z > z_{\alpha/2}) = \alpha/2$, siendo $Z \in N(0,1)$. Entonces:

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

- Equivalentemente,

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



Intervalo de confianza para la media μ de una población normal (σ^2 conocida)

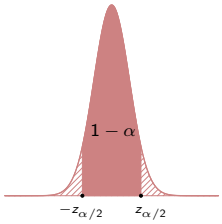
Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$. Supongamos que σ^2 es conocida.

- Sea $z_{\alpha/2}$ el valor tal que $P(Z > z_{\alpha/2}) = \alpha/2$, siendo $Z \in N(0,1)$. Entonces:

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

- Equivalentemente,

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



Intervalo de confianza de nivel $1-\alpha$ para la media μ cuando σ^2 es conocida

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

Intervalo de confianza para la media μ de una población normal (σ^2 conocida)

Intervalo de confianza de nivel $1-\alpha$ para la media μ cuando σ^2 es conocida

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

Ejemplo: Un investigador está interesado en determinar el nivel medio de determinada proteína en el cuerpo humano. Para ello toma una muestra de 10 individuos y obtiene el nivel de proteína de cada uno de ellos. Los resultados son los siguientes:

$$22, 20, 24, 18, 23, 25, 26, 20, 19, 23$$

- ¿Cómo estimarías el nivel medio de proteína a partir de esta muestra?
- Nuevas investigaciones determinan que la variable de interés es aproximadamente normal con varianza igual a 45. Construye un intervalo de confianza para el nivel medio de proteína en el cuerpo humano con nivel de confianza del 95 %.
- ¿Cuál sería el intervalo de confianza para un nivel de confianza del 90 %?

Intervalo de confianza para la media μ de una población normal (σ^2 desconocida)

Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$.

- En la práctica no es habitual conocer la varianza de la variable de interés.

Intervalo de confianza para la media μ de una población normal (σ^2 desconocida)

Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$.

- En la práctica no es habitual conocer la varianza de la variable de interés.
- Cuando la varianza σ^2 es desconocida, usaremos como estadístico (pivote) para construir un intervalo de confianza para la media μ

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

- Recuerda que:

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

Intervalo de confianza para la media μ de una población normal (σ^2 desconocida)

Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$.

- En la práctica no es habitual conocer la varianza de la variable de interés.
- Cuando la varianza σ^2 es desconocida, usaremos como estadístico (pivote) para construir un intervalo de confianza para la media μ

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

- Recuerda que:

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

- En este caso:

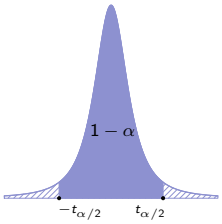
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \in t_{n-1}$$

Intervalo de confianza para la media μ de una población normal (σ^2 desconocida)

Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$. Supongamos que 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. Entonces:

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



Intervalo de confianza para la media μ de una población normal (σ^2 desconocida)

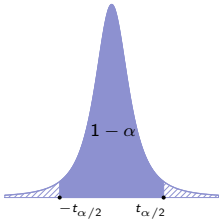
Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$. Supongamos que σ^2 es desconocida.

- 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. Entonces:

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

- Equivalentemente,

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



Intervalo de confianza para la media μ de una población normal (σ^2 desconocida)

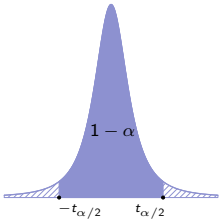
Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$. Supongamos que σ^2 es desconocida.

- 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. Entonces:

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

- Equivalentemente,

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



Intervalo de confianza de nivel $1-\alpha$ para la media μ cuando σ^2 es desconocida

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

t de Student con $n-1$ g.l.

Intervalo de confianza para la media μ de una población normal (σ^2 desconocida)

Intervalo de confianza de nivel $1-\alpha$ para la media μ cuando σ^2 es desconocida

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

t de Student con $n-1$ g.l.

Ejemplo: Considera las siguientes medidas, correspondientes al Volumen Espiratorio Forzado[1] (litros) de 10 sujetos de un estudio que examina la respuesta al ozono entre adolescentes que sufren asma.

3,50, 2,60, 2,75, 2,82, 4,05, 2,25, 2,68, 3,00, 4,02, 2,85

- ¿Cómo estimarías el Volumen Espiratorio Forzado medio?
- Construye un intervalo de confianza para el Volumen Espiratorio Forzado medio con nivel de confianza del 95 %.
- ¿Cuál sería el intervalo de confianza para un nivel de confianza del 90 %?

[1] El Volumen Espiratorio Forzado es la cantidad de aire expulsado durante el primer segundo de la espiración máxima, realizada tras una inspiración máxima

Intervalo de confianza para la diferencia de medias de poblaciones normales

- En algunas ocasiones estamos interesados en estimar la diferencia de medias $\mu_1 - \mu_2$ de dos poblaciones.
- Tenemos 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** (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 que las **varianzas σ_1^2 y σ_2^2 son conocidas**.

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

Intervalo de confianza para la diferencia de medias de poblaciones normales

Intervalo de confianza de nivel $1-\alpha$ para la diferencia de medias $\mu_1 - \mu_2$ de poblaciones normales. Muestras independientes y varianzas conocidas

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

Ejemplo: Un equipo de investigación está interesado en la diferencia en el nivel de ácido úrico en pacientes con y sin un determinado síndrome. Se recogieron en un hospital especializado en dicha enfermedad, los niveles de ácido úrico de 12 individuos con el síndrome. Se obtuvo una media muestral de 4.5 unidades. En otro hospital general se recogieron los niveles de ácido úrico de 15 individuos sin el síndrome. En ese caso la media muestral obtenida fue 3.4 unidades. Asumimos que ambas poblaciones se distribuyen según una normal con varianzas 1 y 1.5, respectivamente. Calcula el intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ al 95 %.

Intervalo de confianza para la diferencia de medias de poblaciones normales

- En algunas ocasiones estamos interesados en estimar la diferencia de medias $\mu_1 - \mu_2$ de dos poblaciones.
- Tenemos 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** (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 que las **varianzas σ_1^2 y σ_2^2 son desconocidas pero iguales**. Sea:

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

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

Intervalo de confianza para la diferencia de medias de poblaciones normales

Intervalo de confianza de nivel $1-\alpha$ para la diferencia de medias $\mu_1 - \mu_2$ de poblaciones normales. Muestras independientes y varianzas desconocidas pero iguales

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

Ejemplo: Un equipo de investigación está interesado en determinar la diferencia en el número medio de días de tratamiento necesario en pacientes con dos tipos de desórdenes mentales. Por un lado se determinó el nº de días de tratamiento en 18 pacientes con esquizofrenia. El número medio de días fue 4.7 con una desviación típica muestral de 9.3 días. Por otro lado se determinó el nº de días de tratamiento en 10 pacientes con trastorno bipolar. El número medio de días fue 8.8 con una desviación típica muestral de 11.5 días. Calcula el intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ al 95 %. Se supone que el número de días de tratamiento es aproximadamente normal y las varianzas son iguales en ambos desórdenes.

Intervalo de confianza para la diferencia de medias de poblaciones normales

- 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.
- Se supone $X_1 \in N(\mu_1, \sigma_1)$ y $X_2 \in N(\mu_2, \sigma_2)$ pero X_1 y X_2 **no son independientes**.
- Consideraremos la variable $D = X_1 - X_2$

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

Ejemplo: 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

Intervalo de confianza para la diferencia de medias de poblaciones normales

Intervalo de confianza de nivel $1-\alpha$ para la para la diferencia de medias $\mu_1 - \mu_2$. Muestras apareadas

$$\left(\bar{D} - t_{\alpha/2}\frac{S_D}{\sqrt{n}} \ , \bar{D} + t_{\alpha/2}\frac{S_D}{\sqrt{n}} \right)_{t \text{ con } n-1 \text{ g.l.}}$$

Ejemplo: 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. Calcula el IC para $\mu_1 - \mu_2$ al 95 %.

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

Intervalo de confianza para la diferencia de medias de poblaciones normales

Intervalo de confianza de nivel $1-\alpha$ para la para la diferencia de medias $\mu_1 - \mu_2$. Muestras apareadas

$$\left(\bar{D} - t_{\alpha/2}\frac{S_D}{\sqrt{n}} \ , \bar{D} + t_{\alpha/2}\frac{S_D}{\sqrt{n}} \right)_{t \text{ con } n-1 \text{ g.l.}}$$

Ejemplo: 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. Calcula el IC para $\mu_1 - \mu_2$ al 95 %.

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

Intervalo de confianza para la diferencia de medias de poblaciones normales

Intervalo de confianza de nivel 1 − α para la para la diferencia de medias $\mu_1 - \mu_2$. Muestras apareadas

$$\left(\bar{D} - t_{\alpha/2}\frac{S_D}{\sqrt{n}}\,,\bar{D} + t_{\alpha/2}\frac{S_D}{\sqrt{n}}\right)_{\ t \text{ con } n-1 \text{ g.l.}}$$

Ejemplo: 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. Calcula el IC para $\mu_1 - \mu_2$ al 95 %.

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

$$\bar{D} = \frac{-5+15+\ldots+10-10}{10} = 6, \qquad S_D^2 = \frac{(-5-6)^2+\ldots+(-10-6)^2}{9} = 71{,}111.$$

Intervalo de confianza para la diferencia de medias de poblaciones normales

Intervalo de confianza de nivel 1 − α para la para la diferencia de medias $\mu_1 - \mu_2$. Muestras apareadas

$$\left(\bar{D} - t_{\alpha/2}\frac{S_D}{\sqrt{n}}\,,\bar{D} + t_{\alpha/2}\frac{S_D}{\sqrt{n}}\right)_{\ t \text{ con } n-1 \text{ g.l.}}$$

Ejemplo: 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. Calcula el IC para $\mu_1 - \mu_2$ al 95 %.

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

$$\bar{D} = \frac{-5+15+\ldots+10-10}{10} = 6, \qquad S_D^2 = \frac{(-5-6)^2+\ldots+(-10-6)^2}{9} = 71{,}111.$$

$$\left(\bar{D} - t_{\alpha/2}\frac{S_D}{\sqrt{n}}\,,\bar{D} + t_{\alpha/2}\frac{S_D}{\sqrt{n}}\right) = \left(6 - 2{,}26\frac{8{,}4327}{\sqrt{10}}\,,6 + 2{,}26\frac{8{,}4327}{\sqrt{10}}\right) = (-0{,}0266, 12{,}0266).$$

Intervalo de confianza para una proporción p

Intervalo de confianza de nivel 1 − α para la proporción p

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

Ejemplo: Una encuesta del proyecto "Pew Internet and American Life Project"[2] llevada a cabo en 2010 determina que el 16 % de los usuarios de internet utilizan la red para consultar información sobre resultados de pruebas médicas. La encuesta, que forma parte de un estudio sobre el uso de internet en América, se basa en entrevistas telefónicas a un total de 3001 adultos. Asumimos que los encuestados fueron elegidos de manera aleatoria. Contruye un intervalo de confianza al 95 % para la proporción de usuarios de internet que consultan información sobre resultados de pruebas médicas en América.

[2] http://www.pewinternet.org/

Intervalo de confianza para la diferencia de proporciones $p_1 - p_2$

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

Intervalo de confianza de nivel 1 − α para la diferencia de proporciones $p_1 - p_2$

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\,,\ (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$

Intervalo de confianza para la diferencia de proporciones $p_1 - p_2$

Intervalo de confianza de nivel 1 − α para la diferencia de proporciones $p_1 - p_2$

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\,,\ (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$

Ejemplo: En un centro educativo se llevó a cabo un estudio para conocer la prevalencia del tabaquismo entre los jóvenes y estudiar las diferencias en el porcentaje de fumadores entre hombres y mujeres. Para ello se seleccionaron dos muestras independientes en cada una de estas poblaciones: 220 alumnos, entre los que había 50 fumadores y 280 alumnas, de las cuales fumaban 90. Calcula el intervalo de confianza para la diferencia de proporciones de fumadores en ambos sexos al 95 %.

Bioestadística. Curso 2012-2013
Grado en Medicina
Capítulo 7. Contrastes de hipótesis

Beatriz Pateiro López

Contraste de hipótesis

- Los procedimientos de inferencia que hemos realizado hasta ahora son:
  - La estimación puntual
  - Los intervalos de confianza
- En este tema vamos a ver otro procedimiento de inferencia basado en **contrastes de hipótesis** en el que el objetivo de la experimentación está orientado a corroborar una hipótesis inicial sobre la población de estudio.

Contraste de hipótesis

- Cuando un investigador trata de entender o explicar algo, generalmente formula su problema de investigación por medio de una hipótesis
- Ejemplo: No sé si la edad media que tienen las mujeres gallegas cuando deciden tener su primer hijo es igual que en el resto de España (29.3 años)

Hipótesis nula
$H_0 : \mu = 29.3$

- Tomo una muestra de 6 mujeres gallegas embarazadas primerizas

- $\bar{X} = 30.5$ años
- ¿Existe suficiente evidencia en los datos para rechazar H_0?
- ¿O la diferencia entre $\bar{X}$ y el valor hipotético de μ puede ser debido al azar?

Contraste de hipótesis

- Cuando un investigador trata de entender o explicar algo, generalmente formula su problema de investigación por medio de una hipótesis
- Ejemplo: No sé si la edad media que tienen las mujeres gallegas cuando deciden tener su primer hijo es igual que en el resto de España (29.3 años)

Hipótesis nula
$H_0 : \mu = 29.3$

- Tomo una muestra de 36 mujeres gallegas embarazadas primerizas

- $\bar{X} = 30.5$ años
- ¿Existe suficiente evidencia en los datos para rechazar H_0?
- ¿O la diferencia entre $\bar{X}$ y el valor hipotético de μ puede ser debido al azar?

Contraste de hipótesis

- 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.
- Llamaremos **hipótesis alternativa**, y la denotamos por H_1 (o H_a) a lo que sucede cuando no es cierta la hipótesis nula.
- Por gozar la hipótesis nula de presunción de inocencia, sobre la hipótesis alternativa recae la carga de la prueba. Por tanto, cuando rechazamos H_0 en favor de H_1 es porque hemos encontrado pruebas significativas a partir de la muestra.

Contraste de hipótesis

Representamos este problema de decisión mediante el siguiente gráfico:

		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.

- Error de tipo I: cuando rechazamos la hipótesis nula, siendo cierta.
- Error de tipo II: cuando aceptamos la hipótesis nula, siendo falsa.

Contraste de hipótesis. Analogía con un juicio

Supongamos un juicio en el que se trata de decidir la culpabilidad o inocencia de un acusado.

- Hipótesis nula: el acusado es inocente (todo acusado es inocente hasta que se demuestre lo contrario).
- Hipótesis alternativa: el acusado es culpable.
- Juicio: es el procedimiento en el cual se trata de probar la culpabilidad del acusado y la evidencia debe ser muy fuerte para que se rechace la inocencia (H_0) en favor de la culpabilidad (H_a).
- Decisión: el veredicto.
- Error de tipo I: condenar a un inocente.
- Error de tipo II: absolver a un culpable.

Contraste de hipótesis

- La probabilidad del error de tipo I se denota por α y se denomina **nivel de significación**.

Nivel de significación

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

- La probabilidad del error de tipo II se denota por β

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

- **Potencia:** Es la probabilidad de detectar que una hipótesis es falsa.

Potencia

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

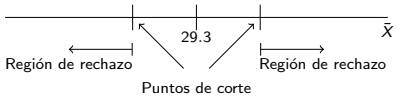
Región crítica. Contrastes bilaterales y unilaterales

- Debemos establecer una regla de decisión para determinar cuando rechazamos o no la hipótesis nula H_0
- Ejemplo: ¿Difiere la edad media de las madres primerizas en Galicia de la edad media de las madres primerizas en el resto de España (29.3 años)?

Contraste bilateral

$$H_0 : \mu = 29.3$$
$$H_1 : \mu \neq 29.3$$

- Si estamos interesados en determinar si μ **difiere significativamente** de 29.3, deberíamos rechazar H_0 si $\bar{X}$ está "lejos" de 29.3 **en ambas direcciones**.



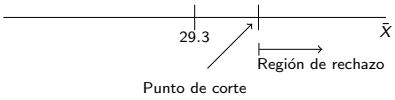
Región crítica. Contrastes bilaterales y unilaterales

- Debemos establecer una regla de decisión para determinar cuando rechazamos o no la hipótesis nula H_0
- Ejemplo: ¿Es la edad media de las madres primerizas en Galicia mayor que la edad media de las madres primerizas en el resto de España (29.3 años)?

Contraste unilateral

$$H_0 : \mu \leq 29.3$$
$$H_1 : \mu > 29.3$$

- Si estamos interesados en determinar si μ **es significativamente mayor** que 29.3, deberíamos rechazar H_0 si $\bar{X}$ está "lejos" de 29.3 **en una sola dirección**.



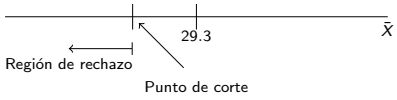
Región crítica. Contrastes bilaterales y unilaterales

- Debemos establecer una regla de decisión para determinar cuando rechazamos o no la hipótesis nula H_0
- Ejemplo: ¿Es la edad media de las madres primerizas en Galicia menor que la edad media de las madres primerizas en el resto de España (29.3 años)?

Contraste unilateral

$$H_0 : \mu \geq 29.3$$
$$H_1 : \mu < 29.3$$

- Si estamos interesados en determinar si μ **es significativamente menor** que 29.3, deberíamos rechazar H_0 si $\bar{X}$ está "lejos" de 29.3 **en una sola dirección**.



Contraste de hipótesis

Las etapas en la resolución de un contraste de hipótesis son:

- Especificar las hipótesis nula H_0 y alternativa H_1.
- Elegir un estadístico de contraste apropiado, para medir la discrepancia entre la hipótesis y la muestra.
- Fijar el nivel de significación α en base a cómo de importante se considere rechazar H_0 cuando realmente es cierta.
- 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 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.
- 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.

Bioestadística. Curso 2012-2013
Grado en Medicina
Capítulo 8. Contrastes de hipótesis II

Beatriz Pateiro López

Contraste de hipótesis

Las etapas en la resolución de un contraste de hipótesis son:

- Especificar las hipótesis nula H_0 y alternativa H_1.
- Elegir un estadístico de contraste apropiado, para medir la discrepancia entre la hipótesis y la muestra.
- Fijar el nivel de significación α en base a cómo de importante se considere rechazar H_0 cuando realmente es cierta.
- 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).
- 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.
- 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.

Contraste sobre la media de una población normal con varianza conocida

Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$.

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

Contraste bilateral (hipótesis nula simple)
$H_0 : \mu = \mu_0$ $H_1 : \mu \neq \mu_0$

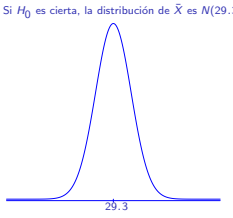
- El sentido común nos aconseja rechazar la hipótesis nula de que la media poblacional es μ_0 cuando la media muestral $\bar{X}$ sea muy distinta de μ_0.

Contraste sobre la media de una población normal con varianza conocida

- Se sabe que la edad de las madres primerizas en Galicia sigue una distribución normal con una desviación típica $\sigma = 2$ años.
- Tomamos una muestra de 36 madres primerizas gallegas. Queremos contrastar si la edad media de las madres primerizas en Galicia difiere de la edad media de las madres primerizas en el resto de España (29.3 años).

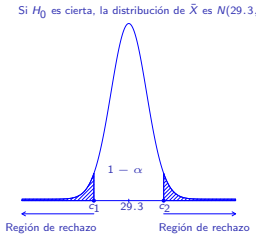
Contraste sobre la media de una población normal con varianza conocida

- Se sabe que la edad de las madres primerizas en Galicia sigue una distribución normal con una desviación típica $\sigma = 2$ años.
- Tomamos una muestra de 36 madres primerizas gallegas. Queremos contrastar si la edad media de las madres primerizas en Galicia difiere de la edad media de las madres primerizas en el resto de España (29.3 años).



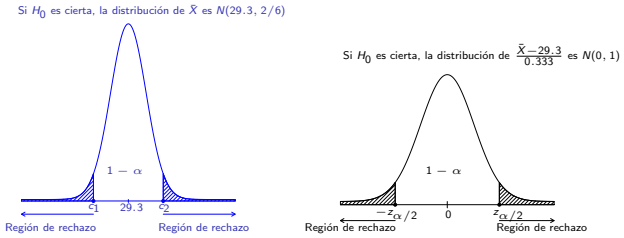
Contraste sobre la media de una población normal con varianza conocida

- Se sabe que la edad de las madres primerizas en Galicia sigue una distribución normal con una desviación típica $\sigma = 2$ años.
- Tomamos una muestra de 36 madres primerizas gallegas. Queremos contrastar si la edad media de las madres primerizas en Galicia difiere de la edad media de las madres primerizas en el resto de España (29.3 años).



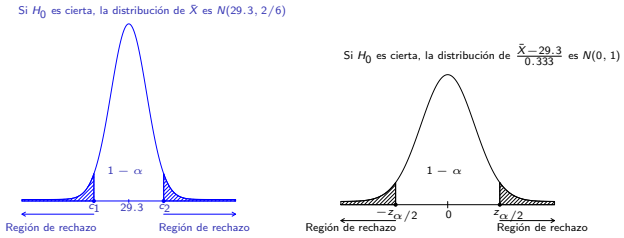
Contraste sobre la media de una población normal con varianza conocida

- Se sabe que la edad de las madres primerizas en Galicia sigue una distribución normal con una desviación típica $\sigma = 2$ años.
- Tomamos una muestra de 36 madres primerizas gallegas. Queremos contrastar si la edad media de las madres primerizas en Galicia difiere de la edad media de las madres primerizas en el resto de España (29.3 años).



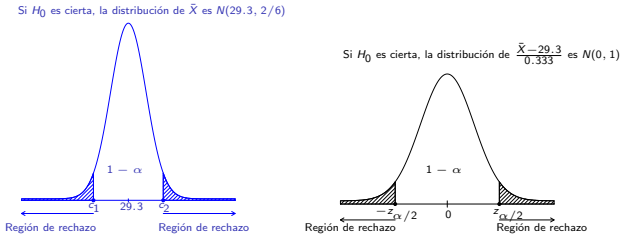
Contraste sobre la media de una población normal con varianza conocida

- Se sabe que la edad de las madres primerizas en Galicia sigue una distribución normal con una desviación típica $\sigma = 2$ años.
- Tomamos una muestra de 36 madres primerizas gallegas. Queremos contrastar si la edad media de las madres primerizas en Galicia difiere de la edad media de las madres primerizas en el resto de España (29.3 años).
- Observamos que $\bar{X} = 30.5$ años. En base a la muestra, ¿podrías concluir que la edad media de las madres primerizas en Galicia difiere de la edad media de las madres primerizas en el resto de España?



Contraste sobre la media de una población normal con varianza conocida

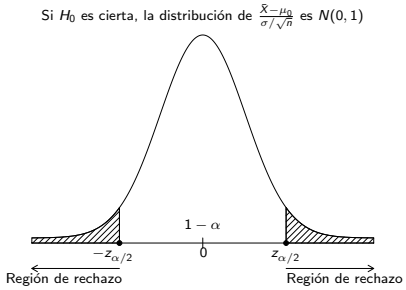
- Se sabe que la edad de las madres primerizas en Galicia sigue una distribución normal con una desviación típica $\sigma = 2$ años.
- Tomamos una muestra de 36 madres primerizas gallegas. Queremos contrastar si la edad media de las madres primerizas en Galicia difiere de la edad media de las madres primerizas en el resto de España (29.3 años).
- Observamos que $\bar{X} = 30.5$ años. En base a la muestra, ¿podrías concluir que la edad media de las madres primerizas en Galicia difiere de la edad media de las madres primerizas en el resto de España?



Rechazamos la hipótesis nula $H_0 : \mu = 29.3$ frente a $H_1 : \mu \neq 29.3$ si

$$\frac{30.5 - 29.3}{0.333} \leq -z_{\alpha/2} \quad \text{ó} \quad \frac{30.5 - 29.3}{0.333} \geq z_{\alpha/2}$$

Contraste sobre la media de una población normal con varianza conocida

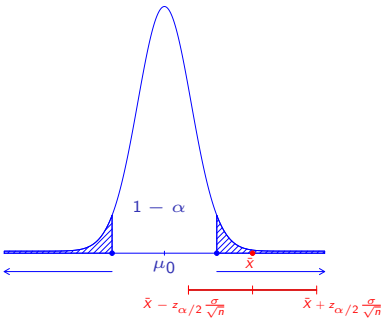


Rechazamos la hipótesis nula $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$ si

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

Relación entre el contraste bilateral y los Intervalos de confianza

- $H_0 : \mu = \mu_0$
- Si H_0 es cierta, la distribución de $\bar{X}$ es $N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right)$



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

Contraste sobre la media de una población normal con varianza conocida

Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$.

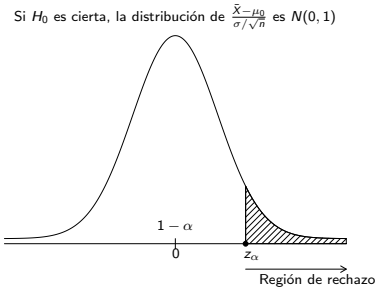
- Supongamos que la varianza σ^2 es conocida
- Se desea contrastar una hipótesis relativa a la media, μ.

Contraste unilateral (hipótesis nula compuesta)

$$H_0 : \mu \leq \mu_0$$
$$H_1 : \mu > \mu_0$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la media poblacional es μ_0 cuando la media muestral $\bar{X}$ sea "considerablemente mayor" que μ_0.

Contraste sobre la media de una población normal con varianza conocida



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 sobre la media de una población normal con varianza conocida

Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$.

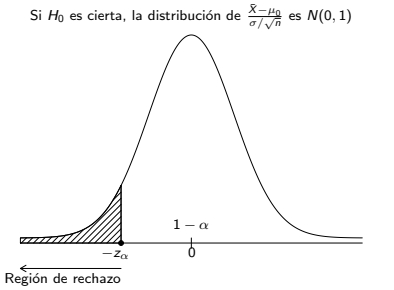
- Supongamos que la varianza σ^2 es conocida
- Se desea contrastar una hipótesis relativa a la media, μ.

Contraste unilateral (hipótesis nula compuesta)

$$H_0 : \mu \geq \mu_0$$
$$H_1 : \mu < \mu_0$$

- El sentido común nos aconseja rechazar la hipótesis nula de que la media poblacional es μ_0 cuando la media muestral $\bar{X}$ sea "considerablemente menor" que μ_0.

Contraste sobre la media de una población normal con varianza conocida



Rechazamos la hipótesis nula $H_0 : \mu \geq \mu_0$ frente a $H_1 : \mu < \mu_0$ si

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

El p-valor

- A medida que el nivel de significación α disminuye es más difícil rechazar la hipótesis nula (manteniendo los mismos datos).
- 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.
- Es decir, si el nivel de significación es menor que p ya no se rechaza H_0.

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

Contraste sobre la media de una población normal con varianza desconocida

Sea $X_1, X_2, \ldots, X_n$ una muestra formada por n variables independientes y con la misma distribución $N(\mu, \sigma)$.

- Supongamos que σ^2 es desconocida
- Se desea contrastar una hipótesis relativa a la media, μ.
- Si H_0 es cierta,

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

- Recuerda que:

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

Contraste sobre la media de una población normal con varianza desconocida

Rechazamos la hipótesis nula $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$ si

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

Rechazamos la hipótesis nula $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$ si

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

Rechazamos la hipótesis nula $H_0 : \mu \geq \mu_0$ frente a $H_1 : \mu < \mu_0$ si

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

t con $n-1$ g.l.

Contrastes referidos a las medias de dos poblaciones normales

- En algunas ocasiones estamos interesados en contrastes sobre la diferencia de medias $\mu_1 - \mu_2$ de dos poblaciones.
- Tenemos 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** (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 que las **varianzas** σ_1^2 **y** σ_2^2 **son conocidas**.

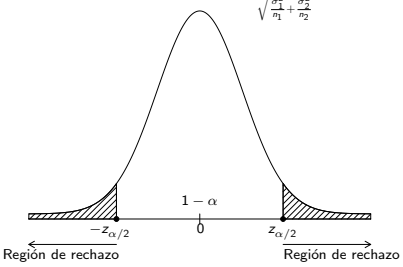
Contraste bilateral (hipótesis nula simple)

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

- El sentido común nos aconseja rechazar la hipótesis nula de que $\mu_1 = \mu_2$ cuando $\bar{X}_1 - \bar{X}_2$ sea muy distinta de 0.

Contrastes referidos a las medias de dos poblaciones normales

Si H_0 es cierta, la distribución de $\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ es $N(0,1)$



Rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

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

Contrastes referidos a las medias de dos poblaciones normales

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$ si

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

Rechazamos la hipótesis nula $H_0 : \mu_1 \geq \mu_2$ frente a $H_1 : \mu_1 < \mu_2$ si

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

Contrastes referidos a las medias de dos poblaciones normales

- En algunas ocasiones estamos interesados en contrastes sobre la diferencia de medias $\mu_1 - \mu_2$ de dos poblaciones.
- Tenemos 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** (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 que las **varianzas** σ_1^2 **y** σ_2^2 **son desconocidas pero iguales**.
- Recuerda que si suponemos que las varianzas de las dos poblaciones son iguales el mejor estimador de la varianza será:

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

Contrastes referidos a las medias de dos poblaciones normales

Rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

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

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$ si

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

Rechazamos la hipótesis nula $H_0 : \mu_1 \geq \mu_2$ frente a $H_1 : \mu_1 < \mu_2$ si

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

t con $n_1 + n_2 - 2$ g.l.

Contrastes referidos a las medias de dos poblaciones normales

- 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.
- Se supone $X_1 \in N(\mu_1, \sigma_1)$ y $X_2 \in N(\mu_2, \sigma_2)$ pero X_1 y X_2 **no son independientes**.
- Consideraremos la variable $D = X_1 - X_2$

Contrastes referidos a las medias de dos poblaciones normales

Rechazamos la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a $H_1 : \mu_1 \neq \mu_2$ si

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

Rechazamos la hipótesis nula $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$ si

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

Rechazamos la hipótesis nula $H_0 : \mu_1 \geq \mu_2$ frente a $H_1 : \mu_1 < \mu_2$ si

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

t con $n-1$ g.l.

Contraste sobre una proporción (muestras grandes)

Rechazamos la hipótesis nula $H_0 : p = p_0$ frente a $H_1 : p \neq p_0$ si

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

Rechazamos la hipótesis nula $H_0 : p \leq p_0$ frente a $H_1 : p > p_0$ si

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

Rechazamos la hipótesis nula $H_0 : p \geq p_0$ frente a $H_1 : p < p_0$ si

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

Bioestadística. Curso 2012-2013
Grado en Medicina
Capítulo 9. Contrastes para datos categóricos

Beatriz Pateiro López

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.
- Ejemplo: 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.

Tablas de contingencia $r \times s$

	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

Tabla de contingencia 2×4 (2 filas, 4 columnas)

Tablas de contingencia 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.

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

Tablas de contingencia 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.

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

Tablas de contingencia 2×2

- Ejemplo de Fundamentals of Biostatistics, Rosner, B. (2000)
- Estudio caso/control:
  - Casos: mujeres con cáncer de mama
  - Controles: mujeres sin cáncer de mama

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

Tablas de contingencia 2×2

- Ejemplo de Fundamentals of Biostatistics, Rosner, B. (2000)
- Estudio caso/control:
  - Casos: mujeres con cáncer de mama
  - Controles: mujeres sin cáncer de mama

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

¿Existe una relación significativa entre el desarrollo de la enfermedad y la edad a la que la mujer tiene el primer hijo?

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:

- Para comprobar afirmaciones acerca de la distribución de una variable aleatoria: Test de bondad de ajuste.
- Para determinar si dos variables son independientes estadísticamente: Test χ^2 de independencia.

Test Chi-cuadrado de independencia

- 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.

Test Chi-cuadrado de independencia

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

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

- Ejemplo de Fundamentals of Biostatistics, Rosner, B. (2000)
- Estudio caso/control:
  - Casos: mujeres con cáncer de mama
  - Controles: mujeres sin cáncer de mama

	Edad al tener el primer hijo		
Tipo	≥ 30	≤ 29	Total
Caso	683 (521.561)	2537 (2698.439)	3220
Control	1498 (1659.439)	8747 (8585.561)	10245
Total	2181	11284	13465

¿Existe una relación significativa entre el desarrollo de la enfermedad y la edad a la que la mujer tiene el primer hijo?

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

- Ejemplo de Fundamentals of Biostatistics, Rosner, B. (2000)
- Estudio caso/control:
  - Casos: mujeres con cáncer de mama
  - Controles: mujeres sin cáncer de mama

	Edad al tener el primer hijo		
Tipo	≥ 30	≤ 29	Total
Caso	683 (521.561)	2537 (2698.439)	3220
Control	1498 (1659.439)	8747 (8585.561)	10245
Total	2181	11284	13465

¿Existe una relación significativa entre el desarrollo de la enfermedad y la edad a la que la mujer tiene el primer hijo?

- Comparamos ahora los datos observados con los datos esperados (entre paréntesis). Si dichos valores son considerablemente distintos, deberíamos rechazar la hipótesis nula de independencia.

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

- El estadístico del contraste es:

$$\chi^2 = \sum_{\text{todas las celdas}} \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}}.$$

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

- El estadístico del contraste es:

$$\chi^2 = \sum_{\text{todas las celdas}} \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}}.$$

- Deberemos rechazar H_0 cuando el valor de χ^2 sea "grande".

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

- El estadístico del contraste es:

$$\chi^2 = \sum_{\text{todas las celdas}} \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}}.$$

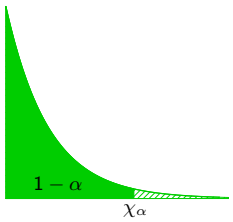
- Deberemos rechazar H_0 cuando el valor de χ^2 sea "grande".
- Bajo H_0, el estadístico se distribuyen aproximadamente según una distribución Chi-cuadrado.
  - Para una tabla $r \times s$: Distribución Chi-cuadrado con $(r-1)(s-1)$ g.l.
  - Para una tabla 2×2: Distribución Chi-cuadrado con 1 g.l.

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

Rechazamos la hipótesis nula H_0 : X e Y son independientes **en tablas** 2×2 **si**

$$\chi^2 = \sum_{\text{todas las celdas}} \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}} \geq \chi_\alpha$$

donde χ_α es el punto que deja a su derecha una probabilidad α en una distribución Chi-cuadrado con 1 grado de libertad



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

Para que la aproximación por la distribución Chi-cuadrado sea buena, es conveniente que las frecuencias esperadas sean grandes.

- En tablas 2×2 se pide que todos los valores esperados 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**.

$$\chi^2_{\text{corregido}} = \sum_{\text{todas las celdas}} \frac{(|\text{observados} - \text{esperados}| - 0.5)^2}{\text{esperados}}.$$

Test Chi-cuadrado de independencia en tablas de contingencia $r \times s$

- Ejemplo estado de salud y capacidad de pago de servicios sanitarios

	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

¿Existe una relación significativa entre el estado de salud y la capacidad que tienen los pacientes de hacer frente al pago de los servicios sanitarios?

Test Chi-cuadrado de independencia en tablas de contingencia $r \times s$

- Ejemplo estado de salud y capacidad de pago de servicios sanitarios

	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

¿Existe una relación significativa entre el estado de salud y la capacidad que tienen los pacientes de hacer frente al pago de los servicios sanitarios?

Test Chi-cuadrado de independencia en tablas de contingencia $r \times s$

- El estadístico del contraste es:

$$\chi^2 = \sum_{\text{todas las celdas}} \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}}.$$

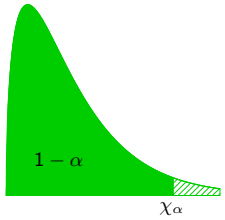
- Deberemos rechazar H_0 cuando el valor de χ^2 sea "grande".
- Bajo H_0, el estadístico se distribuyen aproximadamente según una distribución Chi-cuadrado.
  - Para una tabla de contingencia de r filas y s columnas: Distribución Chi-cuadrado con $(r-1)(s-1)$ g.l.

Test Chi-cuadrado de independencia en tablas de contingencia $r \times s$

Rechazamos la hipótesis nula H_0 : X e Y son independientes **en tablas** $r \times s$ **si**

$$\chi^2 = \sum_{\text{todas las celdas}} \frac{(\text{observados} - \text{esperados})^2}{\text{esperados}} \geq \chi_\alpha$$

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



Bioestadística. Grado en Medicina. Beatriz Pateiro López
Capítulo 9. Contrastes para datos categóricos

Bioestadística. Curso 2012-2013
Grado en Medicina
Capítulo 10. Regresión y correlación

Beatriz Pateiro López

Introducción

- En el primer capítulo nos hemos ocupado de la descripción de variables estadísticas unidimensionales.
- 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.
- Nos centraremos en el estudio de variables estadísticas bidimensionales.
- 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.

Ejemplos

- ¿Existe relación entre la altura en el peso? ¿de qué tipo es esa relación?
- ¿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? ¿cómo?
- ¿Está relacionado el Volumen Expiratorio Forzado con la estatura?

Ejemplo Volumen Expiratorio Forzado y estatura

- 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):
  - X es la estatura de niños de 10 a 15 años de edad.
  - Y es 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

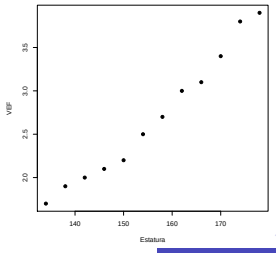
El diagrama de dispersión

- 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 refleja la posible relación entre las variables.
- A mayor relación entre las variables más estrecha y alargada será la nube.

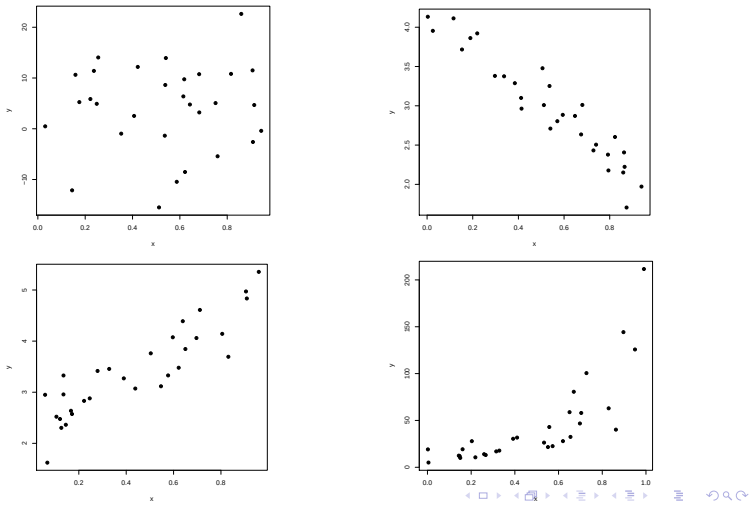
El diagrama de dispersión

- 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 refleja la posible relación entre las variables.
- A mayor relación entre las variables más estrecha y alargada será la nube.

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



Diagramas de dispersión



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.

Covarianza entre X e Y

$$\text{Cov}(X,Y) = s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

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.

Covarianza entre X e Y

$$\text{Cov}(X,Y) = s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

- La covarianza puede interpretarse como una medida de relación lineal entre las variables X e Y.

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.

Covarianza entre X e Y

$$\text{Cov}(X,Y) = s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

- La covarianza puede interpretarse como una medida de relación lineal entre las variables X e Y.
- La covarianza de (X,Y) es igual a la de (Y,X), es decir, $s_{xy} = s_{yx}$.

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.

Covarianza entre X e Y

$$\text{Cov}(X,Y) = s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

- La covarianza puede interpretarse como una medida de relación lineal entre las variables X e Y.
- La covarianza de (X,Y) es igual a la de (Y,X), es decir, $s_{xy} = s_{yx}$.
- La covarianza de (X,X) es igual a la varianza de X, es decir $s_{xx} = s_x^2$

Ejemplo Volumen Expiratorio Forzado y estatura: covarianza

Para los datos del ejemplo sobre el VEF y la estatura se obtiene que:

Ejemplo Volumen Expiratorio Forzado y estatura: covarianza

Para los datos del ejemplo sobre el VEF y la estatura se obtiene que:

- La estatura media es $\bar{x} = 156$ centímetros.
- 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.

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**

Coeficiente de correlación lineal entre X e Y

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

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**

Coeficiente de correlación lineal entre X e Y

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

- La correlación lineal toma valores entre -1 y 1 y sirve para investigar la relación lineal entre las variables.

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**

Coeficiente de correlación lineal entre X e Y

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

- La correlación lineal toma valores entre -1 y 1 y sirve para investigar la relación lineal entre las variables.
- Si toma valores cercanos a -1 diremos que hay una relación inversa entre X e Y.

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**

Coeficiente de correlación lineal entre X e Y

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

- La correlación lineal toma valores entre -1 y 1 y sirve para investigar la relación lineal entre las variables.
- Si toma valores cercanos a -1 diremos que hay una relación inversa entre X e Y.
- Si toma valores cercanos a $+1$ diremos que hay una relación directa entre X e Y.

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**

Coeficiente de correlación lineal entre X e Y

$$r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

- La correlación lineal toma valores entre -1 y 1 y sirve para investigar la relación lineal entre las variables.
- Si toma valores cercanos a -1 diremos que hay una relación inversa entre X e Y.
- Si toma valores cercanos a $+1$ diremos que hay una relación directa entre X e Y.
- Si toma valores cercanos a cero diremos que no existe relación lineal entre X e Y.

Ejemplo Volumen Expiratorio Forzado y estatura: correlación

Para los datos del ejemplo sobre el VEF y la estatura se obtiene que:

Ejemplo Volumen Expiratorio Forzado y estatura: correlación

Para los datos del ejemplo sobre el VEF y la estatura se obtiene que:

- La desviación típica de la estatura es $s_x = 14.422$ centímetros.
- La desviación típica del VEF es $s_y = 0.748$ litros.
- El coeficiente de correlación lineal entre X e Y 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.

Modelo de regresión lineal

- El tipo de relación más sencilla que se establece entre un par de variables es la **relación lineal** $Y = \beta_0 + \beta_1 X$
- 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 regresora** (X) se articula mediante una **recta de regresión**.

Recta de regresión

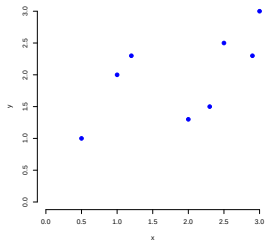
$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Modelo de regresión lineal

Recta de regresión

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

- Dada una muestra $(x_1,y_1),\ldots,(x_n,y_n)$ de la variable bidimensional (X, Y), ¿Cuál es la recta que mejor ajusta los datos?



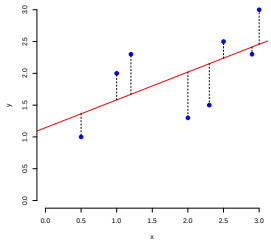
- El objetivo es determinar los valores de los parámetros desconocidos β_0 y β_1 (mediante estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$) de manera que la recta definida ajuste de la mejor forma posible a los datos.

El método de mínimos cuadrados

- El **método de mínimos cuadrados** consiste en encontrar los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ 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 $\hat{y}_i$

El método de mínimos cuadrados

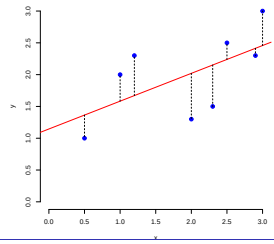
$$M(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$



El método de mínimos cuadrados

Coeficientes estimados por el método de mínimos cuadrados

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Recta de regresión de Y sobre X

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Ejemplo Volumen Expiratorio Forzado y estatura: recta de regresión

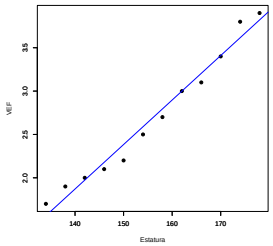
Para los datos del ejemplo sobre el VEF y la estatura se obtiene que:

Ejemplo Volumen Expiratorio Forzado y estatura: recta de regresión

Para los datos del ejemplo sobre el VEF y la estatura se obtiene que:

- $\hat{\beta}_1 = \frac{10.672}{14.422^2} = 0.0513$
- $\hat{\beta}_0 = 2.691 - 156 \cdot 0.0513 = -5.312$
- La recta de regresión será entonces

$$y = \hat{\beta}_0 + \hat{\beta}_1 x = -5.312 + 0.0513x$$



Descomposición de la variabilidad

- La variabilidad de toda la muestra se denomina **variabilidad total (VT)**.

$$\text{VT} = \sum_{i=1}^{n} (y_i - \bar{y})^2 .$$

- La variabilidad total se descompone en dos sumandos:
  - La variabilidad explicada (VE).

$$\text{VE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

  - La variabilidad no explicada (VNE) por la regresión.

$$\text{VNE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Descomposición de la variabilidad

$$\text{VT} = \text{VE} + \text{VNE}.$$

Coeficiente de determinación

- El **coeficiente de determinación** (R^2) se define como la proporción de variabilidad de la variable dependiente que es explicada por la regresión

Coeficiente de determinación

$$R^2 = \frac{\text{VE}}{\text{VT}} = 1 - \frac{\text{VNE}}{\text{VT}}.$$

- En el modelo de regresión lineal simple, el coeficiente de determinación coincide con el cuadrado del coeficiente de correlación.

$$R^2 = r_{xy}^2$$

Ejemplo Volumen Expiratorio Forzado y estatura: coeficiente de determinación

Para los datos del ejemplo sobre el VEF y la estatura se obtiene que:

Ejemplo Volumen Expiratorio Forzado y estatura: coeficiente de determinación

Para los datos del ejemplo sobre el VEF y la estatura se obtiene que:

- $R^2 = 0.9881^2 = 0.976$
- Con el modelo de regresión lineal simple hallado, la variable X es capaz de explicar el 97.6 % de la variación de Y.