ESTADÍSTICA

ingeniería química USC

Curso 2008-2009

Beatriz Pateiro López

ESTADÍSTICA
ingeniería química USC

APUNTES DE TEORÍA
Tema 1: ESTADÍSTICA DESCRIPTIVA
Tema 2: DESCRIPCIÓN ESTADÍSTICA DE DOS VARIABLES
Tema 3: PROBABILIDAD
Tema 4: VARIABLES ALEATORIAS UNIDIMENSIONALES
Tema 5: VECTORES ALEATORIOS: VECTORES BIDIMENSIONALES
Tema 6: MODELOS DE DISTRIBUCIÓN DE PROBABILIDAD
Tema 7: INFERENCIA ESTADÍSTICA: ESTIMACIÓN PUNTUAL E INTERVALOS DE CONFIANZA
Tema 8: CONTRASTE DE HIPÓTESIS

BOLETINES DE EJERCICIOS
Boletín 1: ESTADÍSTICA DESCRIPTIVA
Boletín 2: DESCRIPCIÓN ESTADÍSTICA DE DOS VARIABLES
Boletín 3: PROBABILIDAD
Boletín 4: VARIABLES ALEATORIAS UNIDIMENSIONALES
Boletín 5: INFERENCIA ESTADÍSTICA: ESTIMACIÓN PUNTUAL E INTERVALOS DE CONFIANZA, CONTRASTE DE HIPÓTESIS

PRÁCTICAS CON MATLAB
Práctica 1: INTRODUCCIÓN A MATLAB
Práctica 2: ESTADÍSTICA DESCRIPTIVA CON MATLAB
Práctica 3: DESCRIPCIÓN ESTADÍSTICA DE DOS VARIABLES
Práctica 4: VARIABLES ALEATORIAS UNIDIMENSIONALES I
Práctica 5: VARIABLES ALEATORIAS UNIDIMENSIONALES II
Práctica 6: INFERENCIA ESTADÍSTICA: ESTIMACIÓN PUNTUAL E INTERVALOS DE CONFIANZA

Este material se encuentra disponible en
http://eio.usc.es/pub/pateiro/files/IQ0809Pateiro.pdf

Estadística
Tema 1: ESTADÍSTICA DESCRIPTIVA

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Introducción a la Estadística

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

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

El campo de la estadística tiene que ver con la recopilación, presentación, análisis y uso de datos para tomar decisiones y resolver problemas. Cualquier persona, tanto en su carrera profesional como en la vida cotidiana recibe información en forma de datos a través de periódicos, de la televisión y de otros medios. De manera específica, el conocimiento de la estadística y la probabilidad puede constituirse en una herramienta poderosa para ayudar a los científicos e ingenieros a diseñar nuevos productos y sistemas, a perfeccionar los existentes y a diseñar, desarrollar y mejorar los procesos de producción. Esta sección consiste en una breve introducción a las actividades propias de la Estadística, sus objetivos y las herramientas y argumentos que utiliza. Pretendemos distinguir con claridad las labores de recolección y tratamiento de datos, el cálculo de probabilidades y los razonamientos de inferencia estadística. A continuación exponemos algunos conceptos básicos:

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

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

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

Por tanto, nuestro objetivo es el conocimiento de la población. Podríamos pensar en analizar a todos los individuos de la misma. Sin embargo, esto puede ser inviable por su coste o por el tiempo que requiere. Entonces nos conformamos con extraer una muestra. La muestra proporciona información sobre el objeto de estudio. Lo habitual en nuestro contexto es que en el procedimiento de extracción intervenga el azar.

Ejemplo 1: Se quiere analizar el número de horas de estudio semanal que dedican los estudiantes de la Titulación de Ingeniería Química de esta Universidad. Para ello se pregunta a 50 alumnos de esta titulación.
Población: Todos los estudiantes de Ingeniería Química de esta Universidad.
Variable: Número de horas de estudio semanal.
Muestra: 50 alumnos encuestados.

Ejemplo 2: Se desea estimar el porcentaje de albúmina en el suero proteico de personas sanas. Para ello se analizan muestras de 40 personas, entre 2 y 40 años de edad.
Población: Todas las personas sanas.

Variable: Porcentaje de albúmina en el suero proteico.
Muestra: 40 personas, entre 2 y 40 años de edad.

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

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

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

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

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

2 Tipos de variables

Variables cualitativas: No aparecen en forma numérica, sino como categorías o atributos. *Ejemplos*: sexo, color de los ojos, profesión, potabilidad del agua, tipo de carburante, origen animal de la leche, etc. Se clasifican a su vez en:

Cualitativas nominales: Miden características que no toman valores numéricos. A estas características se les llama modalidades. *Ejemplo*: Si se desea examinar el origen animal de una serie de productos lácteos considerados para un estudio, las modalidades podrían ser: Vaca, Oveja, Cabra,...

Cualitativas ordinales: Miden características que no toman valores numéricos pero sí presentan entre sus posibles valores una relación de orden. *Ejemplos*: nivel de estudios: sin estudios, primaria, secundaria, etc.

Variables cuantitativas: Toman valores numéricos porque son frecuentemente el resultado de una medición. *Ejemplos*: peso (kg.) de una persona, altura (m.) de edificios, temperatura (c^o) corporal, concentración (g 100 ml^{-1}) de inmunoglobina en suero sanguíneo, porcentaje ($0-100\,\%$) de agua recuperada al centrifugar piedra arsénica, nivel (mg Kg^{-1}) de cromo en hierba de centeno, etc. Se clasifican a su vez en:

Cuantitativas discretas: Toman un número discreto de valores (en el conjunto de números naturales). *Ejemplos*: número de hijos de una familia, número de átomos que constituyen una molécula gaseosa, etc.

Cuantitativas continuas: Toman valores numéricos dentro de un intervalo real. *Ejemplos*: altura, peso, concentración de un elemento, tiempo de reaccion de un compuesto quimico, etc.

3 Distribución de frecuencias

La primera forma de recoger y resumir la información contenida en la muestra es efectuar un recuento del número de veces que se ha observado cada uno de los distintos valores que puede tomar la variable. A eso le llamamos frecuencia. Daremos definiciones precisas del concepto de frecuencia en sus distintas formas de presentación. Definimos previamente el **tamaño muestral**, al que denotamos por n, como el número de observaciones en la muestra.

3.1 Descripción de variables cualitativas.

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

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

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

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

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

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

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

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

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

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

Propiedades:

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

Claramente, la suma de las frecuencias absolutas es el número total de datos, n, y la suma de las frecuencias relativas es 1. Observar que el último valor de la distribución de frecuencias absolutas acumuladas coincide con el número de observaciones y que los valores no decrecen. Análogamente, el último valor de la distribución de frecuencias relativas acumuladas es uno.
La información que proporcionan la distribución de frecuencias relativas y la distribución de frecuencias relativas acumuladas es equivalente, pues cada una de ellas puede obtenerse a partir de la otra.
La distribución de frecuencias acumuladas permite conocer la proporción de valores por debajo de cierto valor de la variable, o entre dos valores especificados, o por encima de cierta cantidad.

Ejemplo 3: Dentro de los procesos industriales de gran importancia para el Ingeniero Químico, están los procesos de tratamiento de aguas. Un laboratorio determinó la dureza del agua de 10 muestras obteniendo los siguientes

resultados.

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

Construir la tabla de distribución de frecuencias relativas para la variable X ="Dureza del agua".

Dureza del agua (c_i)	n_i	f_i	N_i	F_i
Agua blanda	4	0,4	4	0,4
Agua dura	2	0,2	6	0,6
Agua muy dura	3	0,3	9	0,9
Agua extremadamente dura	1	0,1	10	1

3.2 Descripción de variables cuantitativas.

Como la variable toma valores numéricos, los datos no se agrupan ahora en **clases** de modo natural. El concepto de continuidad supone la imposibilidad de repetición en los distintos valores que toma la variable. Y es que en variables cuantitativas continuas y mejorando la precisión del aparato de medida, podríamos encontrar diferencias entre dos observaciones cualesquiera. Así, las frecuencias absolutas adoptarían casi siempre el valor uno en cada observación y las frecuencias relativas el valor $1/n$.
Por esta razón, 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. Algunas consideraciones a tener en cuenta:

- *Número de intervalos a considerar:* Para adoptar esta decisión tendremos en cuenta:

  1. Cuantos menos intevalos tomemos, menos información se recoge.
  2. Cuantos más intervalos tomemos, más difícil es manejar las frecuencias.

  Aunque no hay unanimidad al respecto, un criterio bastante extendido consiste en tomar como número de intervalos el entero más próximo a $\sqrt{n}$.

- *Amplitud de cada intervalo:* Lo más común, salvo justificación en su contra, es tomar todos los intervalos de igual longitud. Una amplitud variable de los intervalos podría justificarse por la búsqueda de una descripción más precisa en ciertas zonas de valores. A dichas zonas dedicaríamos más intervalos, con una consiguiente menor longitud.

- *Posición de los intervalos:* Los intervalos deben situarse allí donde se encuentran las observaciones y de forma contigua. Por lo demás, es aconsejable que los restos de intervalos en los extremos derecho e izquierdo del conjunto de observaciones sean similares.

Debemos añadir que para una variable cuantitativa discreta que pueda tomar demasiados valores distintos puede ser conveniente una agrupación por intervalos como en el caso continuo.
A continuación veremos un ejemplo práctico de cómo se construyen los intervalos y la tabla de frecuencias para variables cuantitativas. En la resolución de los ejemplos será útil ordenar la muestra de observaciones y después calcular el **recorrido o rango**, que definimos como la diferencia entre el dato más grande y el más pequeño de la muestra. El recorrido se usa para obtener la amplitud de los intervalos. La ordenación facilita mucho también el recuento de las frecuencias en cada intervalo.

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

Tamaño muestral: $n = 200$.

Número de hijos	n_i	f_i	N_i	F_i
0	50	0,25	50	0,25
1	80	0,40	130	0,65
2	40	0,20	170	0,85
3	20	0,10	190	0,95
4	10	0,05	200	1

Ejemplo 5: Con la finalidad de conocer el comportamiento de algunas variables químicas del suelo y las correspondientes del sedimento provocado por el proceso de erosión hídrica, se analizaron las pérdidas de suelo por escurrimiento de 10 muestras de suelo. Los valores de pérdida de agua (cm^3) de cada muestra son:

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

Calculemos una tabla de frecuencias con estos datos.

Muestra ordenada: $22, 23, 28, 29, 31, 47, 51, 52, 53, 64$.
Recorrido$= 64 - 22 = 42$.
Número de intervalos $\approx \sqrt{10} \approx 3'1623 \approx 3$.
$42/3 = 14$. Tomamos como amplitud de cada intervalo 15 y así conseguimos contener toda la muestra y los extremos de los intervalos resultan manejables.

Intervalo de clase	Marca de clase					Densidad de frecuencia
$[L_i, L_{i+1})$	c_i	n_i	f_i	N_i	F_i	$n_i/(L_{i+1} - L_i)$
$[20, 35)$	27,5	5	0,5	5	0,5	5/15
$[35, 50)$	42,5	1	0,1	6	0,6	1/15
$[50, 65)$	57,5	4	0,4	10	1	4/15

4 Representaciones gráficas

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

4.1 Representaciones gráficas de variables cualitativas

Representaremos las frecuencias absolutas o relativas mediante el **diagrama de barras**. Para ello, situamos los valores de la variable en el eje de abscisas, respetando su orden si lo hubiera, y dibujamos barras verticales sobre ellos con altura proporcional a la frecuencia. Para el aspecto del diagrama de barras es irrelevante si representamos las frecuencias absolutas o relativas. Esto sólo afectaría a la escala del eje de ordenadas.

Ejemplo 6: Utilizaremos el siguiente ejemplo. Realizamos un experimento en el laboratorio cuyo resultado es el color que adopta finalmente un líquido. Los colores posibles son: azul, verde y rojo. Se realiza el experimento 50 veces y en 15 ocasiones resulta el color azul, 25 veces se obtiene el color verde y en 10 ocasiones resulta el color rojo. Entonces tenemos:

Tamaño muestral: $n = 50$.

Color (c_i)	n_i	f_i
Azul	15	0,3
Verde	25	0,5
Rojo	10	0,2

Nótese que no calculamos las frecuencias acumuladas pues el color es una variable nominal. A continuación representamos el diagrama de barras.

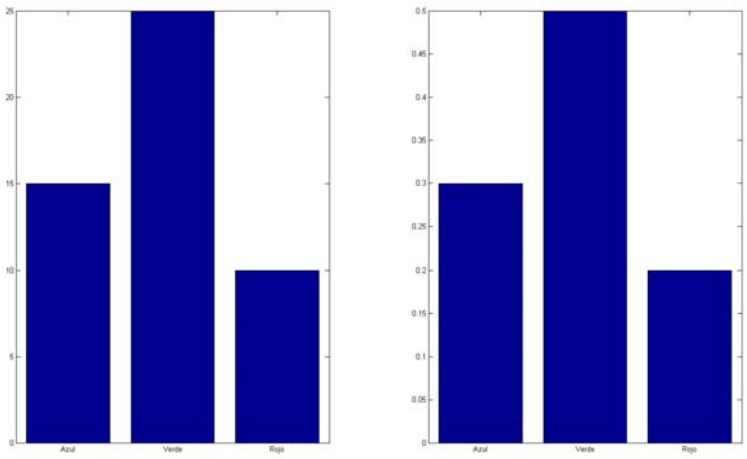

Figura 1: Diagrama de barras. A la izquierda frecuencias absolutas y a la derecha frecuencias relativas.

```
% Guardamos en el vector ni las frecuencias absolutas
ni=[15 25 10];
% Guardamos en el vector fi las frecuencias relativas
fi=[0.3 0.5 0.2];
% Diagrama de barras con las frecuencias absolutas
subplot(1,2,1), bar(ni)
set(gca,'XTickLabel',{'Azul','Verde','Rojo'})
% Diagrama de barras con las frecuencias relativas
subplot(1,2,2), bar(fi)
set(gca,'XTickLabel',{'Azul','Verde','Rojo'})
```

4.2 Representaciones gráficas de variables cuantitativas

Variables cuantitativas discretas

Igual que para las variables cualitativas, las frecuencias de las variables cuantitativas discretas se representan mediante el diagrama de barras.

Asimismo, representaremos las frecuencias acumuladas mediante el **diagrama de frecuencias acumuladas** o **diagrama escalonado**, según se muestra a continuación. Los datos son los correspondientes al Ejemplo 4.

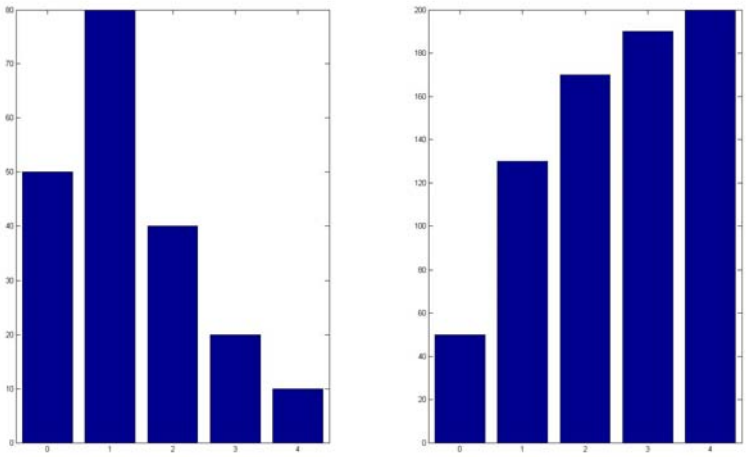

Figura 2: A la izquierda diagrama de barras y a la derecha diagrama de frecuencias acumuladas.

```
% Datos
ci=[0 1 2 3 4];
% Guardamos en el vector ni las frecuencias absolutas
ni=[50 80 40 20 10];
% Guardamos en el vector Ni las frecuencias absolutas acumuladas
Ni=cumsum(ni);
% Diagrama de barras con las frecuencias absolutas
subplot(1,2,1), bar(ci,ni)
% Diagrama de frecuencias acumuladas
subplot(1,2,2), bar(ci,Ni)
```

Variables cuantitativas continuas

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

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

Destacamos que es el área y no la altura de los rectángulos lo que debe ser proporcional a la frecuencia. Así, el eje de ordenadas no refleja la frecuencia, sino que la altura de cada rectángulo representa la **densidad**

de frecuencia sobre ese intervalo, definida como:

$$\text{Densidad de frecuencia} = \frac{\text{frecuencia}}{\text{Amplitud}}$$

Sólo si se toman clases de la misma longitud, las frecuencias son proporcionales a las alturas del histograma de modo que, donde hay más altura hay más datos y donde hay menos altura menos datos. Vamos a dibujar el histograma correspondiente a la distribución de frecuencias obtenida en el Ejemplo 5.

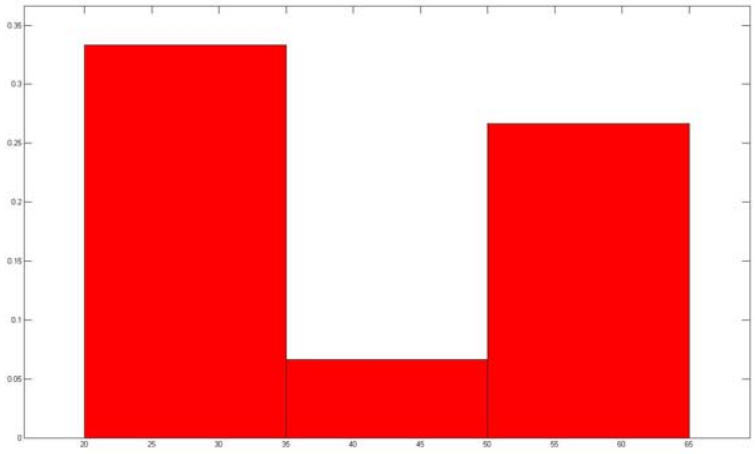

Figura 3: Histograma.

```
% La funcion mihist3 que veremos en clase de prácticas construye
% el histograma y el polígono de frecuencias acumuladas
% x: Contiene los datos
% extremos: Contiene los extremos de los intervalos de clase

x=[22 23 28 29 31 47 51 52 53 64];
extremos=[20; 35; 50; 65];
mihist3([22 23 28 29 31 47 51 52 53 64],[20; 35; 50; 65]);
```

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

El polígono de frecuencias: Se obtiene uniendo mediante segmentos los centros de las bases superiores de los rectángulos del histograma. Proporcionan una representación más suavizada que el histograma.

El polígono de frecuencias acumuladas: Las frecuencias acumuladas se representan mediante el polígono de frecuencias acumuladas. Resulta interesante observar que el polígono de frecuencias acumuladas se obtiene integrando el histograma de izquierda a derecha.

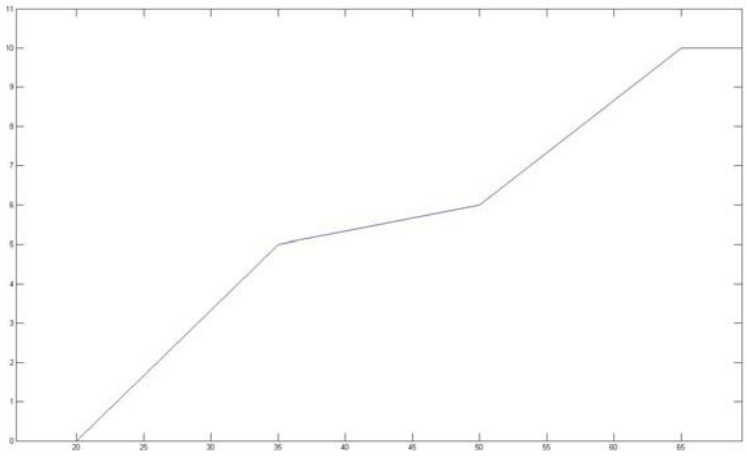

Figura 4: Polígono de frecuencias acumuladas.

Diagrama de tallo y hojas: Los datos se redondean a dos o tres cifras significativas, tomándose como tallo la primera o dos primeras cifras y como hojas las ultimas cifras. El tallo se separa de las hojas por una línea vertical. Así, cada tallo se representa una sola vez y el número de hojas representa la frecuencia. La impresión resultante es la de *acostar* un histograma.

Ejemplo 7: El DDT es un potente insecticida que fue muy empleado a comienzo de los 80. La mezcla técnica de DDT está compuesta básicamente por tres compuestos, entre ellos el pp-DDT. Se tienen los siguientes niveles de pp-DDT en 30 muestras de judías blancas (mg Kg^{-1}).

0.03	0.05	0.08	0.08	0.10	0.11	0.18	0.19	0.20	0.20
0.22	0.22	0.23	0.29	0.30	0.32	0.34	0.40	0.47	0.48
0.55	0.56	0.58	0.64	0.66	0.78	0.78	0.86	0.89	0.96

A continuación se muestra el diagrama de tallo y hojas correspondiente. El punto decimal se situa un dígito a la izquierda de |.

```
0 | 3588
1 | 0189
2 | 002239
3 | 024
4 | 078
5 | 568
6 | 46
7 | 88
8 | 69
9 | 6
```

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

En estas dos últimas secciones del tema estudiamos las medidas que sirven para obtener una descripción muy resumida sobre alguna propiedad concreta del conjunto de datos. Por **medida** entendemos, pues, un número

que se calcula sobre la muestra y que refleja cierta cualidad de la misma. 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.

5.1 Medidas de posición

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

Media aritmética: Se define la media aritmética (o simplemente media) como:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}; \quad \bar{x} = \sum_{i=1}^{m} c_i f_i$$

donde la primera expresión corresponde a tener todos los datos cuantitativos y la segunda corresponde a datos agrupados. Así, en el caso de una variable continua, tenemos dos opciones: o calculamos la media con todos los datos (los sumamos y dividimos por el tamaño muestral), o usamos la tabla de frecuencias considerando las marcas de clase y las frecuencias en cada clase. Los resultados serán diferentes, siendo la segunda opción una aproximación de la primera, con la ventaja de una mayor sencillez de cálculo.

La media aritmética tiene interesantes propiedades:

Propiedades:

1. $\text{mín}(x_i) \leq \bar{x} \leq \text{máx}(x_i)$ y tiene las mismas unidades que los datos originales.
2. Es el centro de gravedad de los datos:
   $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$,
   $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \min_{a \in \mathbb{R}} \sum_{i=1}^{n} (x_i - a)^2$.
3. Si $y_i = a + bx_i \Rightarrow \bar{y} = a + b\bar{x}$. (las transformaciones lineales se comportan bien con la media).

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. Si la variable está agrupada en intervalos de clase, se calcula la clase mediana (aquel intervalo donde la frecuencia relativa acumulada es menor o igual que 0,5 en su extremo inferior y mayor que 0,5 en su extremo superior) para a continuación elegir un representante de este intervalo como mediana (la marca de clase, el valor obtenido por interpolación lineal, *etc.*).

Propiedades:

1. La mediana es la medida de posición central más robusta (*i.e.* más insensible a datos anómalos).
2. La mediana verifica:
   $\sum_{i=1}^{n} |x_i - M_e| = \min_{a \in \mathbb{R}} \sum_{i=1}^{n} |x_i - a|$.

Observa que la media y la mediana tendrán valores similares, salvo cuando haya valores atípicos o cuando la distribución sea muy asimétrica.

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.**

Cuantiles: Sea $p \in (0, 1)$. Se define el cuantil p como el número que deja a su izquierda una frecuencia relativa p. Lo que es lo mismo, la frecuencia relativa acumulada hasta el cuantil p es p. Claro está que los cuantiles sólo se podrán calcular con variables ordinales. Nótese que la mediana es el cuantil 0'5. Para calcular los cuantiles seguiremos las siguientes indicaciones.

- Si la variable es discreta, o si es continua y disponemos de todos los datos:
  Ordenamos la muestra. Tomamos el menor dato de la muestra (primero de la muestra ordenada) cuya frecuencia relativa acumulada es mayor o igual que p. Si se supera p estrictamente, este dato ya es el cuantil p; mientras que si se alcanza con igualdad, el cuantil p es la media de este dato con el siguiente.

  Ejemplo 8: Muestra ordenada: 1, 2, 3'5, 6, 7, 9, 12, 13, 14'5, 15'2.
  Cuantil 0'10=1'5; Cuantil 0'43=7.
- Si la variable es continua y se encuentra agrupada en intervalos de clase:
  Buscamos sobre la tabla de frecuencias el primer intervalo cuya frecuencia relativa acumulada es mayor o igual que p (pensemos que es el intervalo $[L_i, L_{i+1})$) y, dentro de ese intervalo, calculamos el cuantil p por interpolación lineal, esto es:

$$\text{Cuantil } p = L_i + \frac{p \cdot n - N_{i-1}}{n_i} (L_{i+1} - L_i)$$

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 **deciles** son los cuantiles de orden (0.1, 0.2,..., 0.9). Los **percentiles** son los cuantiles de orden $j/100$ donde j=1,2,...,99.

5.2 Medidas de dispersión

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

Recorrido o rango: $R = \text{máx}\, x_i - \text{mín}\, x_i$.

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

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

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2; \quad s^2 = \sum_{i}^{m} (c_i - \bar{x})^2 f_i.$$

Propiedades:

1. $s^2_{a+X} = s^2_X$. La varianza no se ve afectada por cambios de localización.

2. $s^2_{b\cdot X} = b^2 \cdot s^2_X$. La varianza se mide en el cuadrado de la escala de la variable

---

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

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

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

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

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

5.3 Medidas de forma

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

Coeficiente de asimetría de Fisher: Se define como

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

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

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

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

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

Varias de las medidas vistas anteriormente utilizan desviaciones de los datos respecto a la media elevadas a distintos órdenes. Este tipo de coeficientes se denominan **momentos**.

- Se define el **momento respecto al origen de orden r** ($r \geq 0$) como:

$$a_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r.$$

- Se define el **momento central de orden r** ($r \geq 0$) como:

$$m_r = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^r.$$

Casos particulares de los momentos son: $a_1 = \bar{x}$, $m_2 = s^2$, $m_3 = s^3 As_F$ y $m_4 = s^4 K_F$.

Ejemplo 9: Un estudio tiene como objetivo determinar la concentración de pH en muestras de saliva humana. Para ello se recogieron datos de 10 personas obteniéndose los siguientes resultados.

6,59 7,37 7,15 7,08 5,75 5,83 7,12 7,23 7,13 5,60

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

La media se calcula como:

$$\bar{x} = \sum\nolimits_{i=1}^{n} \frac{x_i}{n} = \frac{6{,}59 + 7{,}37 + \ldots + 5{,}60}{10} = 6{,}685.$$

Para calcular la mediana debemos en primer lugar ordenar los datos:

5,60 5,75 5,83 6,59 7,08 7,12 7,13 7,15 7,23 7,37

La mediana se calcula entonces como:

$$Me = \frac{7{,}08 + 7{,}12}{2} = 7{,}1.$$

Calculamos ahora la desviación típica. En primer lugar calculamos la varianza

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = 0{,}43172.$$

Entonces, la desviacíon típica es $s = \sqrt{0{,}43172} = 0{,}657$.
Calculemos ahora los cuartiles. El primer cuartil Q_1 es el cuantil de orden 0,25. Para calcularlo, volvemos a tener en cuenta la muestra ordenada. Tomamos el menor dato de la muestra (primero de la muestra ordenada) cuya frecuencia relativa acumulada es mayor o igual que $p = 0{,}25$, es decir, el valor 5,83. Como se supera $p = 0{,}25$ estrictamente tenemos que $Q_1 = 5{,}83$. Recuerda que $Q_2 = Me = 7{,}1$. Por último, $Q_3 = 7{,}15$. En consecuencia, el recorrido intercuántilico es $RI = Q_3 - Q_1 = 1{,}32$.

Estadística
Tema 2: DESCRIPCIÓN ESTADÍSTICA DE DOS VARIABLES

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Variable estadística bidimensional

En el Tema 1 nos hemos ocupado de la descripción de variables estadísticas unidimensionales, es decir, cada individuo de la muestra era descrito de acuerdo a una única característica. Sin embargo, lo habitual es que tendamos a considerar un conjunto amplio de características para describir a cada uno de los individuos de la población, y que estas características puedan presentar relación entre ellas. Así, si para un mismo individuo observamos simultáneamente k características obtenemos como resultado una variable estadística k-dimensional. Nos centraremos en el estudio de variables estadísticas bidimensionales, es decir, tendremos dos características por cada individuo.

Ejemplo 1: Con frecuencia se obtienen datos bivariados cuando se usan dos técnicas distintas para medir la misma cantidad. Por ejemplo, la concentración de hidrógeno determinada con un método de cromatografía de gases (X), y la concentración determinada con un nuevo método de sensor (Y):

X	47	62	65	70	70	78	95	100	114	118	124	127	140	140	140	150	152	164	198	221
Y	38	62	53	67	84	79	93	106	117	116	127	114	134	139	142	170	149	154	200	215

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. Es claro que el estudio de cada variable bidimensional particular (X, Y) variará según las variables unidimensionales X e Y sean cuantitativas o cualitativas y, de ser cuantitativas, según sean continuas o discretas.

2 Distribuciones de frecuencias

Como en el caso unidimensional, con la obtención de las distribuciones de frecuencias para variables bidimensionales se pretende organizar la información contenida en las observaciones muestrales de la variable (X, Y) de manera que sea más sencilla de interpretar en la práctica.

2.1 Distribución de frecuencia conjunta

Estudiaremos las características (X,Y) de una población de la cual obtenemos una muestra $(x_1,y_1),\ldots,(x_n,y_n)$. Igual que hemos hecho con una sola variable, cada una de estas variables se puede agrupar en modalidades. Supongamos que las modalidades (o datos agrupados) de X son $c_1,\ldots,c_m$ y las de Y son $d_1,\ldots,d_k$.

Frecuencia absoluta: Sea n_{ij} el número de individuos de la muestra que presentan la modalidad c_i de X y la d_j de Y. Este número se conoce como la frecuencia absoluta del par (c_i,d_j).

Frecuencia relativa: Al igual que para variables unidimensionales, las frecuencias relativas se calculan como

$$f_{ij} = \frac{n_{ij}}{n}.$$

Las propiedades de estos números son idénticas al caso unidimensional. La **distribución de frecuencias conjunta** de la variable bidimensional (X, Y) es el resultado de organizar en una tabla de doble entrada las modalidades de las variables unidimensionales X e Y junto con las correspondientes frecuencias absolutas (relativas).

Propiedades:

Frecuencias absolutas	$0 \leq n_{ij} \leq n, \quad (i = 1, \ldots, m, \quad j = 1, \ldots, k)$	$\sum_{i=1}^{m} \sum_{j=1}^{k} n_{ij} = n$
Frecuencias relativas	$0 \leq f_{ij} \leq 1, \quad (i = 1, \ldots, m, \quad j = 1, \ldots, k)$	$\sum_{i=1}^{m} \sum_{j=1}^{k} f_{ij} = 1$

$\mathbf{X \backslash Y}$	$\mathbf{d}_1$	$\ldots$	$\mathbf{d}_j$	$\ldots$	$\mathbf{d}_k$
$\mathbf{c}_1$	$n_{11}(f_{11})$	$\ldots$	$n_{1j}(f_{1j})$	$\ldots$	$n_{1k}(f_{1k})$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$\mathbf{c}_i$	$n_{i1}(f_{i1})$	$\ldots$	$n_{ij}(f_{ij})$	$\ldots$	$n_{ik}(f_{ik})$
$\vdots$	$\vdots$		$\vdots$		$\vdots$
$\mathbf{c}_m$	$n_{m1}(f_{m1})$	$\ldots$	$n_{mj}(f_{mj})$	$\ldots$	$n_{mk}(f_{mk})$

Ejemplo 2: Distribución de frecuencias absolutas del color de ojos (X) de 100 personas y de sus madres (Y)

$\mathbf{X \backslash Y}$	Claros	Oscuros
Claros	28	15
Oscuros	20	37

Ejemplo 3: Distribución de frecuencias relativas de asistencia mensual al cine (X) y al teatro (Y) de una muestra de 200 estudiantes universitarios.

$\mathbf{X \backslash Y}$	0	1	2
1	0.41	0.05	0
2	0.19	0.06	0.02
3	0.10	0.05	0.02
4	0.02	0.07	0.01

Ejemplo 4: Distribución de frecuencias relativas del volumen de ventas (X) y el número de trabajadores (Y) para un grupo de 100 empresas pequeñas y medianas.

$\mathbf{X \backslash Y}$	1-24	25-59	50-74	75-99
1-100	28/100	7/100	1/100	0
101-200	10/100	15/100	6/100	2/100
201-300	4/100	10/100	8/100	9/100

2.2 Distribuciones marginales

Llamaremos **distribuciones marginales** a las distribuciones de frecuencias unidimensionales que resultan de agregar todas las frecuencias que incluyen una determinada modalidad de la variable unidimensional. El nombre de marginal proviene de que esta distribución se obtiene a partir de la distribución conjunta acumulando en los

márgenes de la tabla la suma de las frecuencias de las filas o columnas. Normalmente se denotaran por

$$n_{i\cdot} = \sum_{j=1}^{k} n_{ij} \quad \text{y} \quad f_{i\cdot} = \sum_{j=1}^{k} f_{ij}$$

cuando correspondan a frecuencias marginales de la primera variable y por

$$n_{\cdot j} = \sum_{i=1}^{m} n_{ij} \quad \text{y} \quad f_{\cdot j} = \sum_{i=1}^{m} f_{ij}$$

cuando corresponda a la segunda.

$\mathbf{X}\backslash\mathbf{Y}$	$\mathbf{d}_1$	$\ldots$	$\mathbf{d}_j$	$\ldots$	$\mathbf{d}_k$	$\mathbf{X}$
$\mathbf{c}_1$	$n_{11}(f_{11})$	$\ldots$	$n_{1j}(f_{1j})$	$\ldots$	$n_{1k}(f_{1k})$	$n_{1\cdot}(f_{1\cdot})$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
$\mathbf{c}_i$	$n_{i1}(f_{i1})$	$\ldots$	$n_{ij}(f_{ij})$	$\ldots$	$n_{ik}(f_{ik})$	$n_{i\cdot}(f_{i\cdot})$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
$\mathbf{c}_m$	$n_{m1}(f_{m1})$	$\ldots$	$n_{mj}(f_{mj})$	$\ldots$	$n_{mk}(f_{mk})$	$n_{m\cdot}(f_{m\cdot})$
$\mathbf{Y}$	$n_{\cdot 1}(f_{\cdot 1})$	$\ldots$	$n_{\cdot j}(f_{\cdot j})$	$\ldots$	$n_{\cdot k}(f_{\cdot k})$	n

Ejemplo 5: Color de ojos (X) de 100 personas y de sus madres (Y)

$\mathbf{X}\backslash\mathbf{Y}$	Claros	Oscuros	X
Claros	28	15	43
Oscuros	20	37	57
Y	48	52	

Ejemplo 6: Asistencia mensual al cine (X) y al teatro (Y) de una muestra de 200 estudiantes universitarios.

$\mathbf{X}\backslash\mathbf{Y}$	0	1	2	X
1	0.41	0.05	0	0.46
2	0.19	0.06	0.02	0.27
3	0.10	0.05	0.02	0.17
4	0.02	0.07	0.01	0.1
Y	0.72	0.23	0.05	

Ejemplo 7: Volumen de ventas (X) y el número de trabajadores (Y) para un grupo de 100 empresas pequeñas y medianas.

$\mathbf{X}\backslash\mathbf{Y}$	1-24	25-59	50-74	75-99	X
1-100	28/100	7/100	1/100	0	36/100
101-200	10/100	15/100	6/100	2/100	33/100
201-300	4/100	10/100	8/100	9/100	31/100
Y	42/100	32/100	15/100	11/100	

2.3 Distribuciones condicionadas

La **distribución de X condicionada a Y=$\mathbf{d_j}$** es la distribución unidimensional de X sabiendo que Y ha tomado la modalidad d_j. Para cada modalidad d_j de Y, la frecuencia absoluta de X condicionada a $Y = d_j$ es

$$n_{i/j} = n_{ij},\ (i = 1, \ldots, m).$$

La frecuencia relativa de X condicionada a $Y = d_j$ es

$$f_{i/j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{f_{ij}}{f_{\cdot j}}\ \ (i = 1, \ldots, m).$$

Esto corresponde a dividir la columna de frecuencias absolutas (relativas) de la modalidad d_j por la suma de todos los valores de la columna.
Análogamente se define la distribución de Y condicionada a $X = c_i$.

Ejemplo 8: Distribución de frecuencias (absolutas) condicionadas del color de ojos (X) de 100 personas con madres de ojos claros (Y)

X\Y = Claros	Claros	Oscuros
	28	20

Ejemplo 9: Distribución de frecuencias (relativas) condicionadas del número de asistencias al cine para los estudiantes que no han ido al teatro.

X\Y = 0	1	2	3	4
	0,41/0,72	0,19/0,72	0,10/0,72	0,02/0,72

3 Representaciones gráficas

La representación gráfica más útil de dos variables continuas sin agrupar es el **diagrama de dispersión**. Consiste en representar en un eje de coordenadas los pares de observaciones (x_i,y_i). La nube así dibujada (a este gráfico también se le llama nube de puntos) refleja la posible relación entre las variables. A mayor relación entre las variables más estrecha y alargada será la nube.
Para los datos del Ejemplo 1, se obtiene el diagrama de dispersión de la Figura 3.

```
x=[47 62 65 70 70 78 95 100 114 118 124 127 140 140 140 150 152 164 198 221];
y=[38 62 53 67 84 79 93 106 117 116 127 114 134 139 142 170 149 154 200 215];
plot(x,y,'o');
xlabel('X=Concentración de hidrógeno con un método de cromatografiía de gases');
ylabel('Y=Concentración de hidrógeno con un nuevo método de sensor');
title('Diagrama de dispersión')
```

4 Medidas características

La mayoría de las medidas características estudiadas en el caso unidimensional pueden extenderse al caso bidimensional (multidimensional). Consideremos (x_1, y_1),..., (x_n, y_n) una muestra de n observaciones de una variable bidimensional cuantitativa (X, Y).

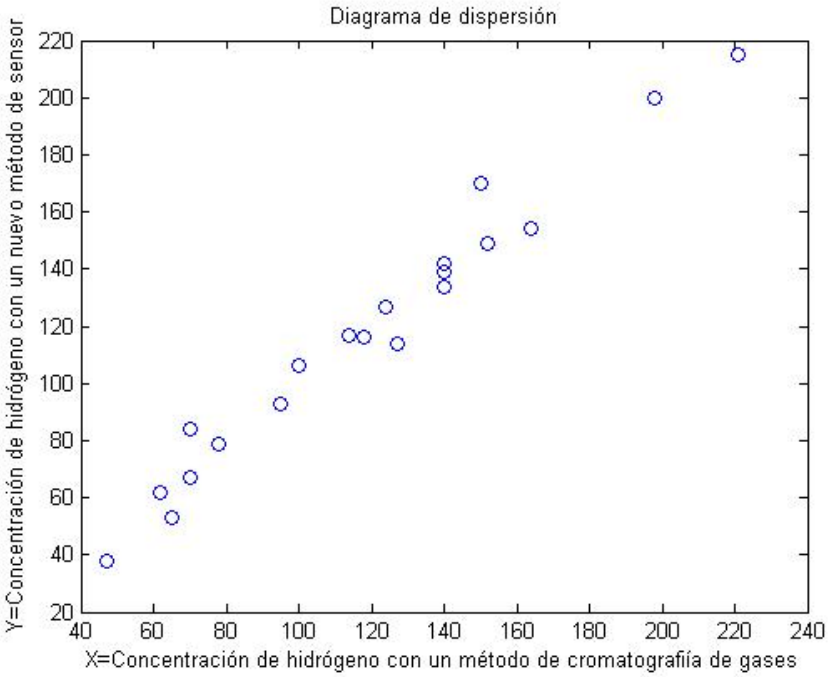

Figura 1: Diagrama de dispersión para los datos del Ejemplo 1.

4.1 Vector de medias

Llamaremos vector de medias de una variable bidimensional (X, Y) al vector $(\bar{x}, \bar{y})$. Siendo $\bar{x}$ la media marginal de la variable X e $\bar{y}$ la media marginal de la variable Y.

4.2 Matriz de varianzas-covarianzas

Llamaremos matriz de varianzas-covarianzas de la variable bidimensional (X, Y) a la matriz

$$S = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$$

donde s_x^2, s_y^2 son las varianzas de las variables X e Y, respectivamente. El término s_{xy} es la **covarianza**, que se define a continuación.

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

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

Propiedades:

1. La covarianza de (X, Y) es igual a la de (Y, X), es decir, $s_{xy} = s_{yx}$
2. La covarianza de (X, X) es igual a la varianza marginal de X, es decir $s_{xx} = s_x^2$

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 muestral, que consiste en tipificar la covarianza dividiéndola por las desviaciones típicas de ambas variables, y se calcula mediante,

$$r(X,Y) = r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

La correlación lineal toma valores entre -1 y 1 y sirve para investigar la relación lineal entre las variables. Así, si toma valores cercanos a -1 diremos que tenemos una relación inversa entre X e Y (esto es, cuando una variable toma valores altos la otra toma valores bajos). Si toma valores cercanos a +1 diremos que tenemos una relación directa (valores altos de una variable en un individuo, asegura valores altos de la otra variable). Si toma valores cercanos a cero diremos que no existe relación lineal entre las variables. Cuando el valor de la correlación lineal sea exactamente 1 o -1 diremos que existe una dependencia exacta entre las variables mientras que si toma el valor cero diremos que son incorreladas.

Propiedades:

1. El coeficiente de correlación toma valores entre -1 y 1.
2. Si existe relación lineal exacta ($Y = a + bX$), el coeficiente de correlación es igual a 1 si la relación es positiva ($b > 0$), e igual a -1 si la relación es negativa ($b < 0$).
3. No depende del orden en que se consideren las variables, es decir, $r_{xy} = r_{yx}$.
4. Si a, b, c, d son constantes, dadas nuevas variables $U = a + bX$ y $V = c + dY$ se verifica que $r_{uv} = r_{xy}$, si $bd > 0$, y $r_{uv} = -r_{xy}$ si $bd < 0$.

---

5 Dependencia lineal. Recta de regresión

En el estudio de variables bidimensionales tiene mucho interés buscar posibles relaciones entre las variables. La más sencilla de estas relaciones es la dependencia lineal donde se supone que la relación entre dos variables X e Y viene dada por la ecuación $Y = a + bX$. Sin embargo, este modelo supone que una vez determinados los valores de los parámetros a y b 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**:

$$Y = a + bX + \varepsilon.$$

Dada una muestra $(x_1,y_1),\ldots,(x_n,y_n)$, el objetivo es determinar los valores de los parámetros desconocidos a y b de manera que la recta definida ajuste de la mejor forma posible a los datos. Aunque existen muchos métodos, el más clásico es el conocido como método de mínimos cuadrados que consiste en encontrar los valores de los parámetros que, dada la muestra de partida, minimizan la suma de los errores al cuadrado. Los coeficientes a y b se determinan minimizando las distancias verticales entre los puntos observados, y_i, y las ordenadas previstas por la recta para dichos puntos ($\hat{y}_i = a + bx_i$). Es decir, el criterio será minimizar

$$M(a,b) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - (a + bx_i))^2.$$

Los valores de los parámetros se obtienen, por tanto, derivando e igualando a cero obteniéndose la solución

$$b = \frac{s_{xy}}{s_x^2}$$

y

$$a = \bar{y} - b\bar{x}$$

que serán llamados **coeficientes de la regresión**. De esta manera obtendremos la ecuación de una recta:

$$y = a + bx = \bar{y} - b\bar{x} + bx = \bar{y} + b(x - \bar{x}) = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

que llamaremos **recta de regresión de Y sobre X** para resaltar que se ha obtenido suponiendo que Y es la variable respuesta y que X es la variable explicativa.
Intercambiando los papeles de X e Y obtendremos una recta de regresión llamada **recta de regresión de X sobre Y** que representada en el mismo eje de coordenadas será en general distinta de la anterior. Solamente coincidirán en el caso de que la relación entre X e Y sea exacta.

Ejemplo 10: Volvamos al Ejemplo 1, donce se recogían datos de la concentración de hidrógeno determinada con un método de cromatografía de gases (X), y la concentración determinada con un nuevo método de sensor (Y). El diagrama de dispersión muestra la recta de regresión de ecuación $y = a + bx = -0{,}9625 + 1{,}0014x$ Haciendo uso de la recta de regresión anterior, si la concentración de hidrógeno determinado con un método de cromatografía de gases es 112 unidades, entonces por el nuevo método será $y = 111{,}15$ unidades.

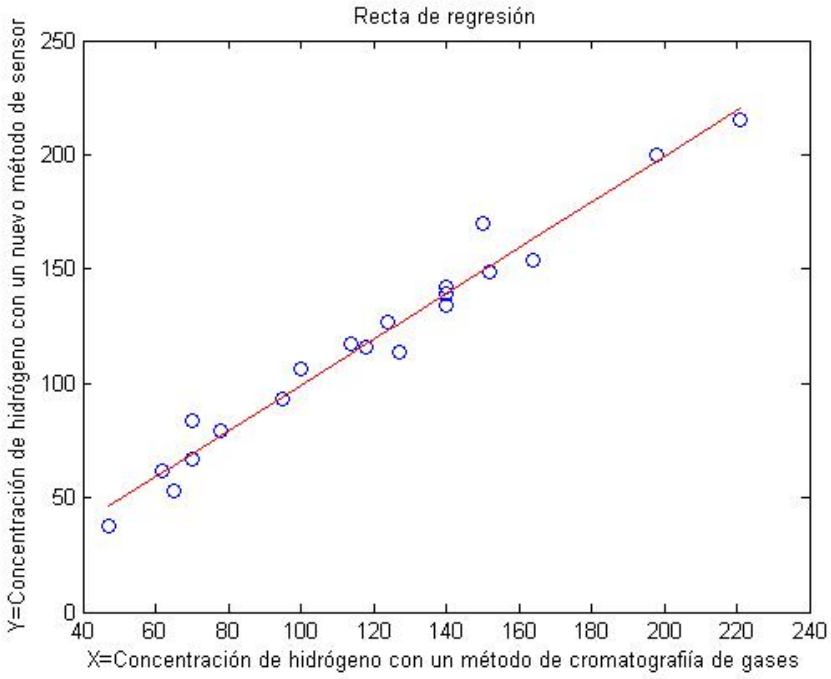

Figura 2: Recta de regresión para los datos del Ejemplo 1.

```
m=polyfit(x,y,1);
yest=polyval(m,x);
plot(x,y,'o',x,yest,'r')
xlabel('X=Concentración de hidrógeno con un método de cromatografiía de gases');
```

```
ylabel('Y=Concentración de hidrógeno con un nuevo método de sensor');
title('Recta de regresión')
```

Una vez resuelto el problema de estimar los parámetros surge la pregunta de si la recta estimada es o no representativa para los datos. Esto se resuelve mediante el **coeficiente de determinación** (R^2) que se define como el cuadrado del coeficiente de correlación lineal. El coeficiente de determinación toma valores entre 0 y 1 y representa el porcentaje de variabilidad de la variable dependiente que es explicada por la regresión.

Ejemplo 11: Para los datos del Ejemplo 1 se puede observar que la recta de regresión no pasa por todos los puntos observados (ver Figura 10). Sin embargo, están muy próximos a ella, el grado de ajuste viene determinado por el coeficiente de determinación $R^2 = 0{,}98522^2 = 0{,}9707$ (el cuadrado del coeficiente de correlación), es decir, con el modelo de regresión lineal simple hallado, la variable X es capaz de explicar el 97,07 % de la variación de Y.

Estadística
Tema 3: PROBABILIDAD

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Introducción histórica

El objetivo de la Estadística es utilizar los datos para inferir sobre las características de una población a la que no podemos acceder de manera completa. Es decir, a partir de la muestra inferir sobre la población. En los temas anteriores, hemos visto como realizar un análisis descriptivo de una muestra de datos y hemos comenzado a intuir que en las distribuciones de frecuencias se repiten ciertos patrones o formas. Esto indica que las observaciones corresponden a un modelo. La **Probabilidad** es la disciplina científica que proporciona y estudia modelos para fenómenos aleatorios en los que interviene el azar y sirve de soporte teórico para la Estadística.
La Teoría de la Probabilidad surgió de los estudios realizados sobre los juegos de azar, y estos se remontan miles de años atrás. Como primeros trabajos con cierto formalismo cabe destacar los realizados por Cardano y Galilei (siglo XVI), aunque las bases de esta teoría fueron desarrolladas por Pascal y Fermat en el siglo XVII. De ahí en adelante grandes científicos han contribuido al desarrollo de la Probabilidad, como Bernouilli, Bayes, Euler, Gauss,... en los siglos XVIII y XIX. Será a finales del siglo XIX y principios del XX cuando la Probabilidad adquiera una mayor formalización matemática, debida en gran medida a la llamada Escuela de San Petesburgo en la que cabe destacar los estudios de Tchebychev, Markov y Liapunov.

2 Conceptos básicos

2.1 Experimento aleatorio

Cuando de un experimento podemos averiguar de alguna forma cuál va a ser su resultado *antes* de que se realice, decimos que el experimento es determinístico. Así, podemos considerar que las horas de salida del Sol, o la pleamar o bajamar son determinísticas, pues podemos leerlas en el periódico antes de que se produzcan. Por el contrario, no podemos encontrar en ningún medio el número premiado en la Lotería de Navidad *antes del sorteo*.
Nosotros queremos estudiar experimentos que no son determinísticos, pero no estamos interesados en todos ellos. Por ejemplo, no podremos estudiar un experimento del que, por no saber, ni siquiera sabemos por anticipado los resultados que puede dar. No realizaremos tareas de adivinación. Por ello definiremos experimento aleatorio como aquel que verifique ciertas condiciones que nos permitan un estudio riguroso del mismo.
Llamamos **experimento aleatorio** al que satisface los siguientes requisitos:

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

Ejemplo 1: Ejemplos de experimentos aleatorios son: $\mathcal{E}_1$ =*Lanzar una moneda al aire*, $\mathcal{E}_2$ =*Lanzar dos veces una moneda*, $\mathcal{E}_3$ =*Lanzar dos monedas a la vez*, $\mathcal{E}_4$ = *Medir en* $mg \cdot kg^{-1}$*la concentración de halofuginona en hígado de pollo*, $\mathcal{E}_5$ =*Determinar la solubilidad del sulfato de bario en gramos por 100 ml de agua*.

2.2 Espacio muestral. Sucesos.

Espacio muestral: Es el conjunto formado por todos los resultados posibles del experimento aleatorio. Lo denotamos por Ω. *Ejemplo:* Si lanzamos una moneda, $\Omega = \{c, +\}$.

Suceso elemental: Es un suceso unitario. Está constituido por un solo resultado del experimento aleatorio. *Ejemplo:* Si lanzamos un dado, $\Omega = \{1, 2, 3, 4, 5, 6\}$, los sucesos elementales son A ="que salga un 1"$= \{1\}$, B ="que salga un 2"$= \{2\}$,...,F ="que salga un 6"$= \{6\}$.

Suceso: Cualquier subconjunto del espacio muestral. *Ejemplo:* Si lanzamos un dado, $\Omega = \{1, 2, 3, 4, 5, 6\}$, podemos considerar muchos sucesos, entre ellos: $A =$"que salga par"$= \{2, 4, 6\}$.

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, que como vemos se pueden enunciar desde el lenguaje común, asignarles probabilidades y efectuar cálculos sobre dichas probabilidades. Observamos que los sucesos no son otra cosa que conjuntos y por tanto, serán tratados desde la Teoría de Conjuntos. Recordamos las operaciones básicas y las dotamos de interpretación para el caso de sucesos.

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

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

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

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

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

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

Sucesos incompatibles: Dos sucesos A y B se dicen incompatibles si no pueden ocurrir a la vez. Dicho de otro modo, que ocurra A y B es imposible. 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 2: Estudiamos el experimento aleatorio consistente en el lanzamiento de un dado, y consideramos los sucesos $A =$"que salga par"$= \{2, 4, 6\}$, $B =$"que sea múltiplo de tres"$= \{3, 6\}$.
El suceso "que salga par y múltiplo de tres" se puede expresar como $A \cap B = \{2, 4, 6\} \cap \{3, 6\} = \{6\}$. De la misma manera, el suceso "que salga par o múltiplo de tres" se puede expresar como $A \cup B = \{2, 4, 6\} \cup \{3, 6\} = \{2, 3, 4, 6\}$.

Ejemplo 3: En los experimentos $\mathcal{E}_1$, $\mathcal{E}_2$ y $\mathcal{E}_3$ del Ejemplo 1 indica cuáles son los sucesos $A =$*sale al menos una cara* y $B =$*no salen cruces*. En el experimento $\mathcal{E}_4$, indica cuáles son los sucesos $C =$*la concentración de halofuginona es menor de* $0.25 mg \cdot kg^{-1}$ y $D =$*la concentración de halofuginona es mayor de* $0.23 mg \cdot kg^{-1}$, y en el experimento $\mathcal{E}_5$, cuál es el suceso $E =$*la solubilidad está entre 0.00021 y 0.00023 gramos por 100 ml de agua*.

Propiedades

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

Ejemplo 4: Indicar en los experimentos $\mathcal{E}_1$, $\mathcal{E}_2$ y $\mathcal{E}_3$ del Ejemplo 3 cuáles son los sucesos $A \cup B$, $A \cap B$. ¿son los sucesos A y B incompatibles?, ¿son los sucesos A y A^c incompatibles?

3 Definiciones de probabilidad

El principal objetivo de un experimento aleatorio suele ser determinar con qué probabilidad ocurre cada uno de los sucesos elementales. A continuación citamos las tres definiciones más manejadas para asignar probabilidades a los sucesos.

3.1 Definición clásica o de Laplace

Nos encontramos ante un experimento, con su colección de sucesos, y nos preguntamos cómo tenemos que actuar para asignarle a cada suceso un número entre 0 y 1 de forma que se respeten los dos axiomas de la definición de probabilidad.
Por desgracia, no existe una solución universal a este problema. En los casos más sencillos podemos hacer deducciones de la propia estructura del experimento, generalmente utilizando su simetría. En otros casos tendremos que combinar la experimentación con la naturaleza teórica del experimento para poder obtener conclusiones.
Cuando el espacio muestral es finito, el problema se reduce a asignar probabilidades a los sucesos elementales. Las probabilidades de los demás sucesos se obtendrán sumando las de los sucesos elementales que lo componen (suma finita).
Sin duda el caso más fácil es aquél en el que no tenemos razones para suponer que unos sucesos sean más probables que otros. Cuando, siendo el espacio muestral Ω finito, todos los sucesos elementales tienen la misma probabilidad, diremos que son **equiprobables** y podremos utilizar la conocida **Regla de Laplace**

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

Ejemplo 5: Lanzamos dos dados y sumamos sus puntuaciones. ¿Cuál es la probabilidad de obtener un 2?, ¿y de obtener un 7?

3.2 Definición frecuentista

Cuando se emplea *coloquialmente* el término probabilidad quiere expresarse el grado de certidumbre o incertidumbre en el resultado de un experimento *antes* de su realización. Para cuantificar la probabilidad nos apoyamos en la experiencia empírica que ya podamos tener de ese experimento. Así, si ya fue observado con anterioridad pudimos haber calculado las frecuencias relativas de los distintos resultados.

$$f_n(A) = \frac{\text{Nº de veces que ha ocurrido } A \text{ en } n \text{ repeticiones}}{n}.$$

En base a esas frecuencias elaboramos nuestra idea de certidumbre sobre el resultado de una futura realización del experimento. Destacamos que las frecuencias se calculan *después* de la realización del experimento y las probabilidades *antes* de la realización del experimento.
Como respaldo a este planteamiento disponemos de la **Ley de Estabilidad de las Frecuencias**, que nos dice que *cuando se repite muchas veces un mismo experimento, las frecuencias relativas de sus posibles resultados tienden a estabilizarse en torno a unos valores*. Aquí usamos que el experimento se puede repetir, siempre en las mismas condiciones, cuantas veces necesitemos.
Así, en el siglo XIX se definió la probabilidad de un suceso como *el límite de sus frecuencias relativas cuando realizamos el experimento muchas veces*.

Esta definición presenta ciertos problemas. Aparte de serias dificultades formales, en la práctica quizás podamos realizar el experimento muchas veces, pero nos será imposible repetirlo indefinidamente.

3.3 Definición axiomática (Kolmogorov 1933)

Las dificultades que presenta la definición frecuentista de probabilidad se han resuelto a principios del siglo XX mediante la utilización de una definición axiomática de la probabilidad, que se basa en que le exigimos unas condiciones de coherencia. La definición, debida al ruso Kolmogorov, es muy parecida a la que damos a continuación.
Sea Ω el espacio muestral, y sea $\mathcal{P}(\Omega)$ el conjunto formado por todos los sucesos. Se define la **probabilidad** como una aplicación $P : \mathcal{P}(\Omega) \longrightarrow [0, 1]$ que cumple las siguientes condiciones:

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

A partir de la definición anterior se 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
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

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

4 Probabilidad condicionada

Supongamos que en el estudio de un experimento aleatorio nos interesa conocer la probabilidad de que ocurra un cierto suceso A. Pero puede ser que dispongamos de información previa sobre el experimento: supongamos que sabemos que el suceso B ha ocurrido. Está claro que ahora la probabilidad de A ya no es la misma que cuando no sabíamos nada sobre B. Por ejemplo, si lanzamos un dado, la probabilidad de que salga 1 es 1/6, pero si disponemos de la información adicional de que el resultado es impar reducimos los casos posibles de 6 a 3 (sólo puede ser un 1, un 3 o un 5), con lo cual la probabilidad es 1/3.
Estamos ahora en condiciones de entender la siguiente definición:
La probabilidad del suceso A **condicionada** al suceso B se define:

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

Fijando el suceso B, la aplicación

$$\begin{array}{rccc} P(\cdot/B): & \mathcal{P}(\Omega) & \longrightarrow & [0,1] \\ & A & \hookrightarrow & P(A/B) \end{array}$$

verifica los axiomas de probabilidad.
También se deduce de manera inmediata que

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

Ejemplo 6: Se ha realizado una encuesta en Santiago para determinar el número de lectores de La Voz y de El Correo. Los resultados fueron que el 25 % lee solamente La Voz, el 20 % sólo El Correo, y el 5 % lee los dos. Si se selecciona al azar un lector de El Correo, ¿cuál es la probabilidad de que lea también La Voz? Y si se ha elegido un lector de La Voz, ¿cuál es la probabilidad de que no lea El Correo?

5 Independencia de sucesos

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

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

Comentarios:

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

Ejemplo 7: Se estima que entre la población de Estados Unidos, el 55 % padece de obesidad, el 20 % es hipertensa, y el 60 % es obesa o hipertensa. ¿Es, de hecho, independiente el que una persona sea obesa de que padezca hipertensión?

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

En esta sección veremos tres teoremas muy importantes, tanto a nivel teórico como para la resolución de ejercicios.

6.1 Regla del producto

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

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

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

Ejemplo 8: La primera aplicación de un insecticida mata al 80 % de los mosquitos. Los supervivientes desarrollan resistencia y en cada aplicación posterior el porcentaje de muertos se reduce a la mitad del verificado en la aplicación inmediatamente anterior: así en la segunda aplicación muere el 40 % de los supervivientes de la primera aplicación, en la tercera aplicación muere el 20 %, etc.

a ¿Cuál es la probabilidad de que un mosquito sobreviva a cinco aplicaciones?

b Idem, dado que sobrevivió a las dos primeras.

6.2 Ley de las probabilidades totales

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

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

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

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

Ejemplo 9: Se sabe que una enfermedad es padecida por el 7 % de los fumadores y por el 2'5 % de los no fumadores. Si en una población de 5.000 habitantes hay 600 fumadores, ¿cuál es la probabilidad de que una persona elegida al azar sufra la enfermedad referida?

6.3 Teorema de Bayes

Por último, tratamos el teorema de Bayes. Consideremos un experimento que se realiza en dos etapas: en la primera, tenemos un sistema completo de sucesos $A_1, A_2, \ldots, A_n$ con probabilidades $P(A_i)$ que denominamos *probabilidades a priori*. En una segunda etapa, ha ocurrido el suceso B y se conocen las probabilidades condicionadas $P(B/A_i)$ de obtener en la segunda etapa el suceso B cuando en la primera etapa se obtuvo el suceso A_i, $\quad i = 1, \ldots, n$.

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. En las condiciones anteriores,

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

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

Estadística
Tema 4: VARIABLES ALEATORIAS UNIDIMENSIONALES

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Introducción

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

2 Variable aleatoria

De manera informal, una **variable aleatoria** es un valor numérico que corresponde al resultado de un experimento aleatorio. Por ejemplo, una variable X como resultado de lanzar una moneda al aire puede tomar el valor 1 si el resultado es cara y 0 si es cruz. De este modo, escribiremos, por ejemplo, $P(X = 1) = 0.5$. Otro ejemplo de variable aleatoria, Y, puede ser el resultado de medir en mg Kg^{-1} la concentración de halofuginona en hígado de pollo. Cuando se han tomado muchísimas observaciones (infinitas), se puede llegar a la conclusión por ejemplo que la probabilidad de que la concentración sea inferior a 0.25 mg Kg^{-1} es igual a 0.8, lo que escribimos con $P(Y < 0.25) = 0.8$.

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

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

Denotamos la variable aleatoria por una letra mayúscula. El conjunto imagen de esa aplicación es el conjunto de valores que puede tomar la variable aleatoria, que serán denotados por letras minúsculas.
Las variables aleatorias son equivalentes a las variables que analizábamos en el tema de Estadística Descriptiva. La diferencia es que en el tema de Estadística Descriptiva se trabajaba sobre una muestra de datos y ahora vamos a considerar que disponemos de toda la población (lo cual es casi siempre imposible en la práctica). Ahora vamos a suponer que podemos calcular las probabilidades de todos los sucesos resultantes de un experimento aleatorio.
De modo idéntico a lo dicho en el tema de Descriptiva, podemos clasificar las variables aleatorias en **discretas** y **continuas** en función del conjunto de valores que pueden tomar. Así, una variable aleatoria será discreta si dichos valores se encuentran separados entre sí. Por tanto será representable por conjuntos discretos, como $\mathbb{Z}$ o $\mathbb{N}$. Una variable aleatoria será continua cuando el conjunto de valores que puede tomar es un intervalo.

2.1 Variables aleatorias discretas.

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

que ocurran estos sucesos,

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

constituyen la distribución de X.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Variables aleatorias continuas.

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

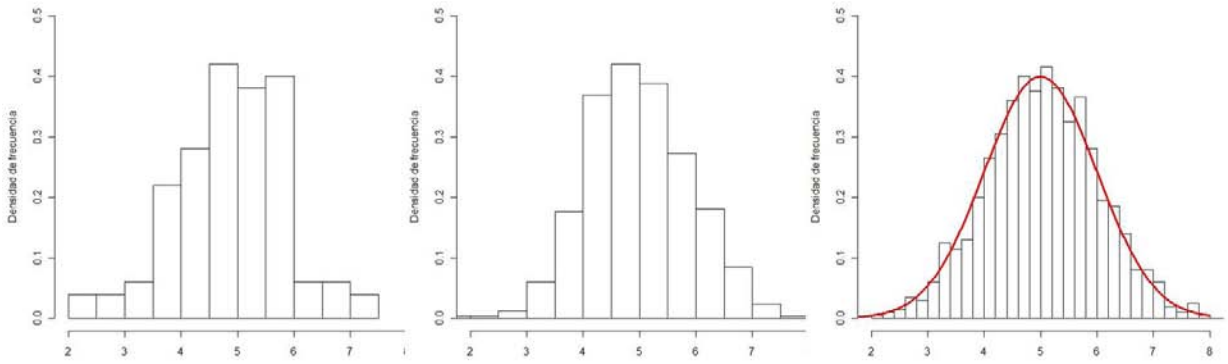

Figura 1: Histograma del diamétro (en mm.) de $n = 100$, $n = 500$ y $n = 1000$ ejes producidos por una empresa. Tomando más observaciones y haciendo más finas las clases, el histograma tiende a estabilizarse en una curva suave (en rojo) que describe la distribución de la variable.

Definición 4. *Llamamos* **función de densidad** *de una variable aleatoria continua X a una aplicación $f : \mathbb{R} \to \mathbb{R}$ no negativa y tal que*

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

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

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

Cualquier función que verifique estas dos propiedades es una función de densidad. La función de densidad se interpreta como el histograma. Sus valores más altos corresponden a las zonas más probables y viceversa. Por

ejemplo, la densidad de la variable $X =$ "Diámetro de un eje" de la Figura 1 indica que lo más probable es que el diámetro tome valores en el intervalo $[4, 6]$. Con menos probabilidad el diámetro estará en los intervalos $[2, 4]$ y $[6, 8]$ y será prácticamente imposible que el diámetro supere los 8 mm. o que sea menor de 2 mm.
Del mismo modo que el histograma representa frecuencias mediante áreas, análogamente, la función de densidad expresa probabilidades por áreas. La probabilidad de que una variable X sea menor que un determinado valor x_0 se obtiene calculando el área de la función de densidad hasta el punto x_0, es decir,

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

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

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

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

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

Ejemplo 7: Se sabe que la proporción de paracetamol en un lote de pastillas es una variable aleatoria continua que tiene como función de densidad,

$$f(x) = \begin{cases} kx, & 0 < x < 100, \\ 0, & \text{en otro caso.} \end{cases}$$

1. Calcular el valor de k para que $f(x)$ sea función de densidad.
2. Calcular la probabilidad de que la proporción de paracetamol sea mayor del 90 %.
3. Calcular la probabilidad de que en un lote determinado la proporción de paracetamol sea inferior al 80 %.
4. ¿Cuál es la probabilidad de que la proporción de paracetamol esté en el intervalo $[80\,\% - 90\,\%]$?

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

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

y por tanto,

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

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

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

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

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

Ejemplo 8: Calcula la función de distribución de la variable $X =$ "Proporción de paracetamol en un lote de pastillas" del Ejemplo 7 y obtén $F(60)$, $F(90)$ y $P(30 \leq X \leq 50)$. Calcular el valor a tal que $P(X \leq a) = 25$.

3 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. Las definiciones son análogas a las introducidas en el tema 1.

3.1 Media o esperanza.

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

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

Ejemplo 9: Calcula la media de pacientes con enfermedades articulares del Ejemplo 2.

Análogamente, la **media poblacional** o **esperanza** de una variable aleatoria continua viene dada por,

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

Ejemplo 10: Calcula la proporción media de paracetamol en un lote de pastillas del Ejemplo 7.

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

3.2 Varianza.

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

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

Ejemplo 11: Calcula la varianza de pacientes con enfermedades articulares del Ejemplo 2.

Análogamente, la **varianza** de una variable aleatoria continua viene dada por,

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

Ejemplo 12: Calcula la varianza de la proporción de paracetamol en un lote de pastillas del Ejemplo 7.

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

3.3 Momentos

Al igual que en el tema 1, podremos definir a nivel poblacional los momentos respecto al origen de orden r y los momentos centrales respecto a la media de orden r. Para ello será de gran utilidad la siguiente propiedad, que se verifica tanto para variables discretas como continuas.

Propiedad.
Sea X una variable aleatoria discreta con valores $x_1, x_2, ..., x_k$. Entonces:

$$\mathbb{E}(g(X)) = \sum_{i=1}^{k} g(x_i)P(X = x_i) = \sum_{i=1}^{k} g(x_i)p_i.$$

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

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

De la anterior propiedad podemos deducir facilmente que, tanto si X es una variable discreta como continua,

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

3.3.1 Momentos respecto al origen de orden r

Se define el **momento respecto al origen de orden r** de una variable aleatoria discreta con valores $x_1, x_2, ..., x_k$como

$$\mathbb{E}(X^r) = \sum_{i=1}^{k} x_i^r p_i.$$

Análogamente, el **momento respecto al origen de orden r** de una variable aleatoria continua viene dada por,

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

3.3.2 Momentos centrales o respecto a la media de orden r

Se define el **momento central o respecto a la media de orden r** de una variable aleatoria discreta con valores $x_1, x_2, ..., x_k$ como

$$\mathbb{E}((X - \mu)^r) = \sum_{i=1}^{k} (x_i - \mu)^r p_i.$$

Análogamente, el **momento central o respecto a la media de orden r** de una variable aleatoria continua viene dada por,

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

4 Desigualdad de Chebychev.

El **teorema de Chebychev** dice que para cualquier variable aleatoria, la probabilidad de que un valor diste de la media menos de k desviaciones típicas es como mínimo $1 - 1/k^2$, es decir,

$$P(-k\sigma \leq X - \mu \leq k\sigma) \geq 1 - \frac{1}{k^2}.$$

Ejemplo 13: Después de medir los diámetros de muchísimos ejes, se llega a la conclusión de que la media poblacional de los diámetros es 5.12 mm. y la desviación típica 0.64 mm. Determinar entre qué valores se encontrará el diámetro de un nuevo eje fabricado con una probabilidad mayor de 0.75

Estadística
Tema 5: VECTORES ALEATORIOS: VECTORES BIDIMENSIONALES

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Concepto de vector aleatorio y su función de distribución.

Pretendemos ahora estudiar varias mediciones simultáneas sobre el resultado de un experimento aleatorio. De esta forma obtenemos vectores cuyas componentes son variables aleatorias como las ya consideradas. En esta ocasión, además vamos a analizar las relaciones entre esas variables.
Por comodidad en la notación sólo trataremos vectores bidimensionales, esto es, la relación entre dos variables aleatorias. Formalmente, un vector aleatorio bidimensional se define simplemente como un par de variables aleatorias definidas sobre el mismo espacio muestral. Lo denotamos

$$(X, Y) : \Omega \longrightarrow \mathbb{R}^2$$

Definición 1. *Se denomina* **función de distribución** *del vector aleatorio* (X, Y) *a*

$$\begin{array}{rccl} F: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (x, y) & \longrightarrow & F(x, y) = P(X \leq x, Y \leq y) \end{array}$$

Propiedades.
La función de distribución de cualquier vector aleatorio verifica las siguientes propiedades, que extienden las ya conocidas para una variable aleatoria.

1. $0 \leq F(x, y) \leq 1 \quad \forall (x, y) \in \mathbb{R}^2$.
2. $\lim\limits_{\substack{x \to +\infty \\ y \to +\infty}} F(x, y) = 1$
   $\lim\limits_{x \to +\infty} F(x, y_0) = 0 \qquad \forall y_0 \in \mathbb{R}$
   $\lim\limits_{y \to +\infty} F(x_0, y) = 0 \qquad \forall x_0 \in \mathbb{R}$
3. F es creciente en cada variable.
   $x_1 \leq x_2 \Rightarrow F(x_1, y_0) \leq F(x_2, y_0) \qquad \forall y_0 \in \mathbb{R}$
   $y_1 \leq y_2 \Rightarrow F(x_0, y_1) \leq F(x_0, y_2) \qquad \forall x_0 \in \mathbb{R}$
4. F es continua por la derecha respecto a cada una de las variables.
   $\lim\limits_{h>0, h \to 0} F(x + h, y) = F(x, y) \qquad \forall (x, y) \in \mathbb{R}^2$
   $\lim\limits_{h>0, h \to 0} F(x, y + h) = F(x, y) \qquad \forall (x, y) \in \mathbb{R}^2$
5. $P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1)$.

Ejemplo 1: Consideremos el vector aleatorio (X, Y) que toma los valores $(2, 0)$, $(0, 2)$ y $(4, 2)$ con igual probabilidad. Calcula $F(1, 1), F(2, 2), F(4, 1), F(2, 4), F(4, 2)$.

Tras definir la función de distribución y estudiar sus propiedades de forma general, trataremos los conceptos fundamentales de vectores aleatorios separando los casos discreto y continuo, que por lo demás sólo se distinguirán en los sumatorios frente a las integrales y las probabilidades frente a las densidades.

2 Vectores aleatorios discretos.

Un vector aleatorio (X, Y) se dice discreto si X e Y son discretas. En ese caso suponemos que X toma los valores $x_1, x_2, \ldots, x_r$ e Y toma los valores $y_1, y_2, \ldots, y_s$. Entonces el vector (X, Y) tomará los valores $(x_i, y_j) \quad 1 \leq i \leq r, \ 1 \leq j \leq s$ con probabilidades $p_{ij} = P(X = x_i, Y = y_j)$. A estas probabilidades las llamamos

probabilidades conjuntas y a la tabla siguiente la llamaremos **distribución de probabilidad conjunta** del vector aleatorio (X, Y).

$X \backslash Y$	y_1	$\dots$	y_j	$\dots$	y_s	
x_1	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1s}	$p_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	p_{i1}	$\cdots$	p_{ij}	$\cdots$	p_{is}	$p_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_r	p_{r1}	$\cdots$	p_{rj}	$\cdots$	p_{rs}	$p_{r\bullet}$
	$p_{\bullet 1}$	$\cdots$	$p_{\bullet j}$	$\cdots$	$p_{\bullet s}$	1

Nótese que a esta tabla le hemos añadido una última columna a la derecha y una última fila en la base. Representan las **distribuciones marginales** de las variables X e Y, respectivamente. La distribución marginal de X es la distribución de probabilidad que tiene la variable X sin tener en cuenta la variable Y. Por eso se obtiene de sumar la fila correspondiente.

$$p_{i\bullet} = P(X = x_i) = \sum_{j=1}^{s} p_{ij}$$

Análogamente definimos la distribución marginal de la variable aleatoria Y y la obtenemos sumando en la columna correspondiente.

$$p_{\bullet j} = P(Y = y_j) = \sum_{i=1}^{r} p_{ij}$$

2.1 Distribuciones condicionadas

Si ya conocemos que la variable X ha tomado el valor x_i, esta información modificará la distribución de probabilidad de la variable Y. De hecho de toda la tabla de valores que puede tomar el vector aleatorio nos restringiremos a la fila de x_i. Sobre ella calcularemos las probabilidades condicionadas

$$P(Y = y_j / X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)} = \frac{p_{ij}}{p_{i\bullet}}$$

y constituirán la distribución de Y condicionada a que $X = x_i$. Como vemos, se obtiene dividiendo la fila de x_i por el total de la fila $p_{i\bullet}$.
Análogamente podemos definir la distribución de X condicionada a que $Y = y_j$.

2.2 Independencia

Diremos que las variables aleatorias X e Y son independientes si cualesquiera dos sucesos relativos respectivamente a X e Y son independientes. Esta definición es aplicable a cualquier vector aleatorio (X, Y). En el caso discreto es equivalente a que la distribución conjunta resulte del producto de las marginales, esto es:

$$\begin{aligned} P(X = x_i, Y = y_j) &= P(X = x_i) \cdot P(Y = y_j) \qquad \forall i \in \{1, \dots, r\} \ \ \forall j \in \{1, \dots, s\} \\ [p_{ij} &= p_{i\bullet} \cdot p_{\bullet j}] \end{aligned}$$

Ejemplo 2: Calcula la distribución de probabilidad conjunta del ejemplo 1, las distribuciones marginales y la distribución de X condicionada a $Y = 2$. ¿Son X e Y independientes?

Ejemplo 3: Considérense las siguientes distribuciones de probabilidad conjunta:

(a)

$X\backslash Y$	10	20
1	1/4	1/4
3	1/4	1/4

(b)

$X\backslash Y$	10	20
1	1/2	0
3	0	1/2

(c)

$X\backslash Y$	10	20
1	0	1/2
3	1/2	0

En cada uno de estos tres casos, representa los valores que puede tomar el vector aleatorio. ¿En qué casos son X e Y independientes?

3 Medidas características de un vector aleatorio.

En base a las distribuciones marginales podemos calcular las medidas ya conocidas (media, varianza, desviación típica,...) para cada una de las variables que componen el vector aleatorio.
Dado el vector aleatorio k-dimensional $(X_1, X_2, \ldots, X_k)$, llamamos **vector de medias** al que tiene por componentes las medias de los X_i, esto es:

$$(\mathbb{E}(X_1), \mathbb{E}(X_2), \ldots, \mathbb{E}(X_k))$$

También podemos estudiar momentos que involucran a dos o más variables del vector aleatorio. Así, por ejemplo, sea (X, Y) el vector aleatorio discreto considerado hasta ahora. La variable aleatoria unidimensional $X \cdot Y$ toma los valores $x_i \cdot y_j \quad 1 \leq i \leq r \; 1 \leq j \leq s$ con probabilidades p_{ij}, y por tanto su media será

$$\mathbb{E}(X \cdot Y) = \sum_{i=1}^{r} \sum_{j=1}^{s} x_i \cdot y_j \cdot p_{ij}$$

Y, en general, para cualquier función $g : \mathbb{R}^2 \to \mathbb{R}$,

$$\mathbb{E}(g(X, Y)) = \sum_{i=1}^{r} \sum_{j=1}^{s} g(x_i, y_j)\, p_{ij}$$

A continuación enunciamos propiedades relativas a la esperanza de la suma y la esperanza del producto:

1. $\mathbb{E}(X_1 + X_2 + \cdots + X_k) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_k)$. (Para cualesquiera variables)
2. Si X e Y son independientes entonces $\mathbb{E}(X \cdot Y) = \mathbb{E}(X) \cdot \mathbb{E}(Y)$.

Definición 2. *Definimos la* **covarianza** *de las variables X e Y como*

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y))] = \sum_{i=1}^{r} \sum_{j=1}^{s} (x_i - \mathbb{E}(X)) \cdot (y_j - \mathbb{E}(Y)) \cdot p_{ij}$$

Es una medida de la relación lineal entre las dos variables, de tal forma que cuando es positiva interpretamos que existe una relación lineal creciente entre ellas y cuando es negativa que dicha relación es decreciente. Además verifica las siguientes propiedades.

1. $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.
2. Si X e Y son independientes entonces $\mathrm{Cov}(X, Y) = 0$. La implicación en sentido inverso no es cierta. Como contraejemplo sirve el ejemplo 1.
3. $\mathrm{Cov}(a + bX, c + dY) = b \cdot d \cdot \mathrm{Cov}(X, Y)$. Por tanto, la covarianza no se ve afectada por los cambios de localización, pero sí por los de escala.
4. $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathrm{Cov}(X, Y)$. De esto se deduce que la varianza de la suma es la suma de varianzas sólo en el caso de que $\mathrm{Cov}(X, Y) = 0$, lo cual ocurre, por ejemplo, si X e Y son independientes.

Llamamos **matriz de varianzas y covarianzas** de un vector aleatorio k-dimensional $(X_1, X_2, \ldots, X_k)$ a

$$\begin{pmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_k) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_k) \\ \vdots & \vdots & & \vdots \\ \text{Cov}(X_k, X_1) & \text{Cov}(X_k, X_2) & \cdots & \text{Var}(X_k) \end{pmatrix}$$

Nótese que es una matriz simétrica y semidefinida positiva.
Hemos dicho que la covarianza depende de la escala. Daremos a continuación una medida de la relación lineal que no depende de la escala y se obtiene simplemente dividiendo la covarianza por las desviaciones típicas de las dos variables. De esta forma acabamos de definir el **coeficiente de correlación**:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

El coeficiente de correlación hereda muchas propiedades de la covarianza. Sigue teniendo el mismo signo de la covarianza y la misma interpretación como medida de dependencia lineal. Al haber eliminado el efecto de la escala verifica además:

$$-1 \leq \rho(X, Y) \leq 1$$

Así, sobre ese rango de valores entre -1 y 1 podemos evaluar la cuantía de la relación lineal. Si la correlación está próxima a 1 hay mucha relación creciente, y si está próxima a -1 hay mucha relación decreciente. Cuando está próxima a cero hay poca dependencia lineal, y cuando vale cero diremos que las variables X e Y están *incorrelacionadas*.

4 Vectores aleatorios continuos.

Definición 3. *Un vector aleatorio (X, Y) se dice continuo si existe una aplicación $f : \mathbb{R}^2 \to \mathbb{R}$ no negativa y tal que*

$$F(x_0, y_0) = P(X \leq x_0, Y \leq y_0) = \int_{-\infty}^{x_0} \int_{-\infty}^{y_0} f(x, y)\, dy\, dx$$

A la función f la llamamos **función de densidad conjunta** *del vector aleatorio (X, Y).*

Podemos hacer un desarrollo análogo al realizado para el caso discreto. Los conceptos son los mismos y únicamente reemplazamos las probabilidades por las densidades y los sumatorios por las integrales. Así, X e Y serán variables continuas y sus densidades marginales se obtienen así:

$$f_X(x_0) = \int_{\mathbb{R}} f(x_0, y)\, dy \quad \forall x_0 \in \mathbb{R} \qquad f_Y(y_0) = \int_{\mathbb{R}} f(x, y_0)\, dx \quad \forall y_0 \in \mathbb{R}$$

La densidad de Y condicionada a que $X = x_0$ y la de X condicionada a que $Y = y_0$ vienen dadas por:

$$f(y/X = x_0) = \frac{f(x_0, y)}{f_X(x_0)} \qquad f(x/Y = y_0) = \frac{f(x, y_0)}{f_Y(y_0)}$$

Las variables X e Y serán independientes si y sólo si las distribuciones condicionadas coinciden con las marginales o, equivalentemente:

$$f(x, y) = f_X(x) \cdot f_Y(y) \quad \forall (x, y) \in \mathbb{R}^2$$

Por último, las medidas se definen de modo natural también para variables continuas y verifican las mismas

propiedades que para el caso discreto. Por ejemplo:

$$E(XY) = \int_{\mathbb{R}} \int_{\mathbb{R}} x \cdot y \cdot f(x,y)\, dx\, dy$$

Ejemplo 4: Sea (X,Y) un vector aleatorio con distribución uniforme en el triángulo de vértices (0,0), (1,0) y (1,1).

(a) Calcular la función de densidad conjunta de (X,Y).

(b) Calcular las funciones de densidad marginales de X y de Y.

(c) Calcular la media y la varianza de X y de Y.

(d) ¿Son X e Y independientes? Calcular el coeficiente de correlación.

(e) Calcular $P\left(Y \geq \frac{1}{2}/X \geq \frac{1}{2}\right)$.

(f) Calcular $P\left(Y \geq \frac{1}{2}/X = \frac{7}{8}\right)$ y $P\left(Y \geq \frac{1}{2}/X = \frac{1}{4}\right)$.

Estadística
Tema 6: MODELOS DE DISTRIBUCIÓN DE PROBABILIDAD

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Introducción

En este tema estudiaremos distribuciones de variables aleatorias que han adquirido una especial relevancia por ser adecuadas para modelizar una gran cantidad de situaciones. Presentaremos en primer lugar los modelos de variables discretas y después los continuos. Caracterizaremos estas distribuciones mediante la distribución de probabilidad en el caso discreto y mediante su función de densidad en el caso continuo. Calcularemos también los momentos (media y varianza) y destacaremos las propiedades de mayor utilidad.

2 Principales modelos de distribuciones discretas

En esta sección examinaremos algunas de las distribuciones de probabilidad de variables aleatorias discretas que más destacan en la teoría estadística y en la práctica.

2.1 Distribución uniforme discreta

Si una variable puede tomar n valores distintos con iguales probabilidades, decimos que ésta tiene una distribución uniforme discreta. Formalmente:

Definición 1. *Una variable aleatoria X tiene una* **distribución uniforme discreta** *y se conoce como variable aleatoria uniforme discreta si y sólo si X toma los valores $x_1 < x_2 < \ldots < x_n$ y su función de probabilidad está dada por:*

$$p_i = P(X = x_i) = \frac{1}{n}, \quad i = 1, \ldots, n.$$

La media y la varianza de esta variable aleatoria son:

- $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$.
- $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$.

Ejemplo 1: Se lanza un dado ordinario. Para $i = 1, \ldots, 6$ definimos la variable aleatoria $X = x_i$ como la cara del dado que cae hacia arriba. Obtén la distribución de probabilidad de esta variable aleatoria, su media y su desviación típica.

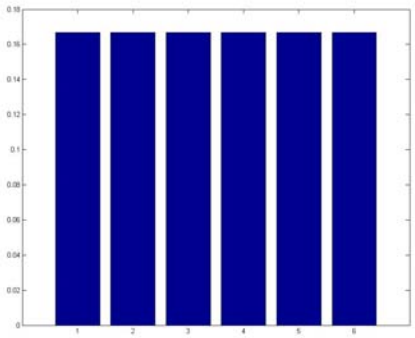

Figura 1: Función de masa de uniforme discreta que toma los valores {1,2,3,4,5,6}.

```
% Función de masa de una Uniforme discreta que toma
% los valores 1,2,3,4,5,6.
masa_unif_discreta=unidpdf(1:6,6);
bar(1:6,masa_unif_discreta)
```

2.2 Distribución de Bernoulli

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

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

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

Definición 2. *Si denotamos por p a la probabilidad de éxito, entonces diremos que la variable X tiene* **distribución de Bernoulli de parámetro p**, *y lo denotamos $X \in Bernoulli(p)$. La distribución de probabilidad de $X \in$ Bernoulli(p) viene dada por*

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

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

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

Como ejemplo, la Figura 2 muestra la función de masa de una variable con distribución de Bernoulli para $p = 0.8$.

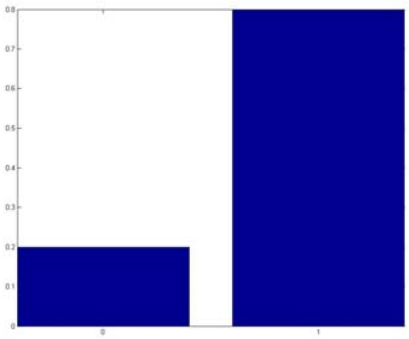

Figura 2: Función de masa de una Bernoulli(0.8).

```
% Función de masa de una Bernoulli(0.8)
masa_bernoulli=binopdf(0:1,1,0.8);
bar(0:1,masa_bernoulli)
```

2.3 Distribución binomial

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

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.

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

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

donde el coeficiente binomial

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

*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).*

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

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

Como ejemplo, la Figura 3 muestra las funciones de masa de una variable con distribución binomial de parámetros $n = 5$ y $p = 1/6$ y una variable con distribución binomial de parámetros $n = 60$ y $p = 1/6$.

```
% Función de masa de una Binomial(5,1/6)
masa_binomial=binopdf(0:5,5,1/6);
subplot(1,2,1), bar(0:5,masa_binomial)
axis([-1 5 0 0.5])
% Función de masa de una Binomial(60,1/6)
masa_binomial=binopdf(0:60,60,1/6);
subplot(1,2,2), bar(0:60,masa_binomial)
axis([-1 60 0 0.15])
```

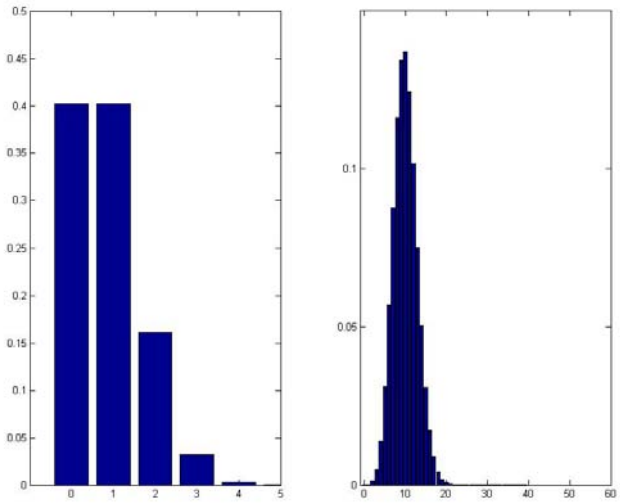

Figura 3: En la izquierda, función de masa de una Binomial(5,1/6). En la derecha, función de masa de una Binomial(60,1/6).

2.4 Distribución de Poisson

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

$$\binom{n}{k} p^k (1-p)^{n-k} \quad \longrightarrow \quad e^{-\lambda} \frac{\lambda^k}{k!}$$

cuando $n \to \infty$ $\left(\lambda \text{ y } k \text{ fijos, } p = \frac{\lambda}{n}\right)$.

Definición 4. *Una variable aleatoria X tiene* **distribución de Poisson de parámetro** λ*, y lo denotamos $X \in Poisson(\lambda)$, si es discreta y*

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

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

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

Como ejemplo, la Figura 4 muestra las funciones de masa de una variable con distribución de Poisson de parámetro $\lambda = 2$ y una variable con distribución de Poisson de parámetro $\lambda = 15$.

```
% Función de masa de una Poisson de parámetro lambda=2
masa_poisson=poisspdf(0:15,2);
subplot(1,2,1), bar(0:15,masa_poisson)
```

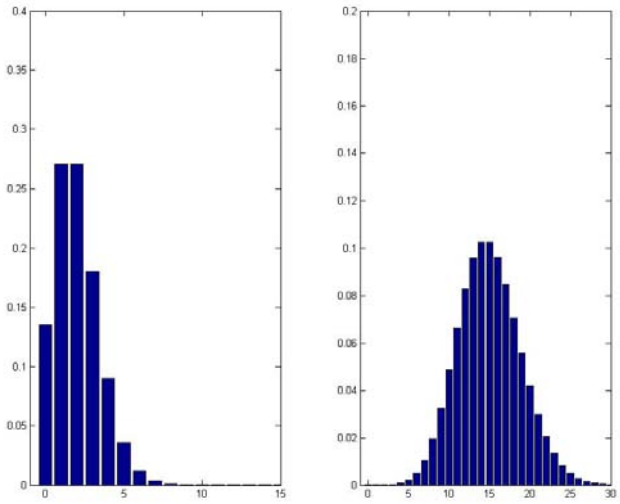

Figura 4: En la izquierda, función de masa de una Poisson(2). En la derecha, función de masa de una Poisson(15).

```
axis([-1 15 0 0.4])

% Función de masa de una Poisson de parámetro lambda=15
masa_poisson=poisspdf(0:30,15);
subplot(1,2,2), bar(0:30,masa_poisson)
axis([-1 30 0 0.2])
```

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

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

Ejemplo 3: La probabilidad de que una persona se desmaye en un concierto es $p = 0.005$. ¿Cuál es la probabilidad de que en un concierto al que asisten 3000 personas se desmayen 18?
La variable $X =$*Número de personas que se desmayan en el concierto* sigue una distribución $Bin(3000, 0.005)$. Queremos calcular

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

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

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

Ejemplo 4: Se sabe que la probabilidad de que un individuo reaccione desfavorablemente tras la inyección de una vacuna es de 0.002. Determina la probabilidad de que en un grupo de 2000 personas vacunadas haya como mucho tres que reaccionen desfavorablemente.

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

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

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

Ejemplo 5: El número de nacimientos en un hospital constituye un proceso de Poisson con intensidad de 21 nacimientos por semana. ¿Cuál es la probabilidad de que se produzcan al menos tres nacimientos la próxima semana?

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

2.5 Distribución hipergeométrica

Consideremos una población formada por N individuos, de los cuales R presentan cierta característica. Extraemos de esa población una muestra **sin reemplazamiento** de n individuos y contamos el número de elementos de la muestra que presentan la característica.
Las similitudes con la distribución binomial son grandes. El experimento consta de n extracciones, que podríamos considerar intentos. La probabilidad de que salga un individuo con la característica en el primer intento es $p = \frac{R}{N}$. Si el muestreo fuera con reemplazamiento los intentos serían independientes y con esa misma probabilidad, resultando así una distribución binomial. Pero al ser **sin reemplazamiento**, el resultado de una extracción afecta a las siguientes.

Definición 5. *Sea $X =$número de individuos con la característica en la muestra obtenida sin reemplazamiento. Diremos que la variable aleatoria X tiene* **distribución Hipergeométrica de parámetros N, n, p***, y lo denotamos $X \in$ Hipergeométrica(N, n, p). La distribución de probabilidad de la Hipergeométrica viene dada por:*

$$P(X = k) = \frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}} \qquad si \;\; k \in \{0, 1, 2, \ldots, n\} \;, \quad k \leq R \;, \quad n - k \leq N - R.$$

La media y la varianza de la Hipergeométrica(N, n, p) son:

- $\mu = np$
- $\sigma^2 = np(1-p)\frac{N-n}{N-1}$

Ejemplo 6: En una urna hay 3 bolas blancas y 5 bolas negras. Extraemos de la urna 6 bolas sin reeemplazamiento. ¿Cuál es la probabilidad de que el número de bolas blancas extraídas sea igual a 2?
La población está formada por $N = 8$ individuos, de los cuales $R = 3$ presentan cierta característica (bola blanca). Extraemos de esa población una muestra **sin reemplazamiento** de $n = 6$ individuos y contamos el

número de elementos de la muestra que presentan la característica.

$$P(X=2)=\frac{\binom{R}{k}\binom{N-R}{n-k}}{\binom{N}{n}}=\frac{\binom{3}{2}\binom{5}{4}}{\binom{8}{6}}=\frac{15}{28}.$$

3 Principales modelos de distribuciones continuas

En esta sección examinaremos algunas de las distribuciones de probabilidad de variables aleatorias continuas que más destacan en la teoría estadística y en la práctica.

3.1 Distribución uniforme

La distribución uniforme es una distribución muy simple cuya función de densidad es simplemente un tramo de línea recta horizontal, denominada densidad uniforme.

Definición 6. *Una variable aleatoria se dice* **uniforme en el intervalo [a,b]**, *y lo denotamos* $X \in Uniforme[a,b]$, *si su función de densidad es*

$$f(x)=\frac{1}{b-a} \qquad si \;\; x\in[a,b]$$

La media y la varianza de una Uniforme[a,b] son:

- La media será el punto medio del intervalo: $\mu=\frac{a+b}{2}$.
- La varianza es: $\sigma^2=\frac{(b-a)^2}{12}$.

Como ejemplo, la Figura 5 muestra la función de densidad de una variable uniforme en el intervalo [5,10]

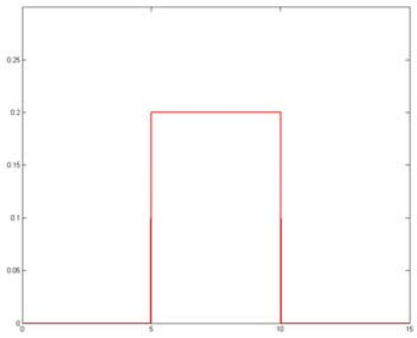

Figura 5: Función de densidad de una Uniforme[5,10].

```
% Densidad de una Uniforme continua en [5,10]
x=linspace(0,15,1000);
unif_continua=unifpdf(x,5,10);
plot(x,unif_continua,'r','LineWidth',2)
axis([0 15 0 0.3])
```

3.2 Distribución exponencial

La distribución exponencial tiene especial utilidad para representar tiempos de vida: duración de una pieza hasta que se avería, longevidad de una persona, etc. Por ello, es una variable continua que toma valores en el intervalo $[0, +\infty)$. La definimos a través de su función de densidad.

Definición 7. *Una variable aleatoria X tiene* **distribución exponencial de parámetro** λ, $\lambda \in (0, +\infty)$, *y lo denotamos $X \in$ Exponencial(λ), si su función de densidad viene dada por:*

$$f(x) = \lambda e^{-\lambda x} \quad si \;\; x \in [0, +\infty)$$

La media y la varianza de una Exponencial(λ) son:

- $\mu = \frac{1}{\lambda}$.
- $\sigma^2 = \frac{1}{\lambda^2}$.

Como ejemplo, la Figura 6 muestra la función de densidad de una variable exponencial de parámetro $\lambda = 1$ y la función de densidad de una variable exponencial de parámetro $\lambda = 1/3$.

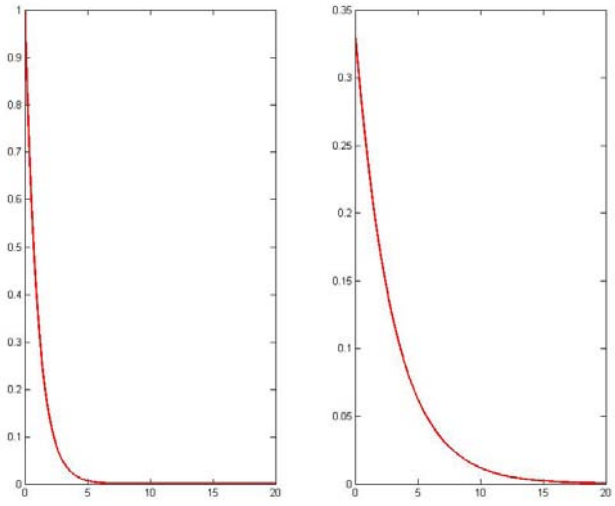

Figura 6: En la izquierda, función de densidad de una Exponencial(1). En la derecha, función de densidad de una Exponencial(1/3)

```
x=linspace(0,20,1000);
% Densidad de una exponencial de parámetro lambda=1
exponencial=exppdf(x,1);
subplot(1,2,1), plot(x,exponencial,'r','LineWidth',2)

% Densidad de una exponencial de parámetro lambda=1/3
exponencial=exppdf(x,3);
subplot(1,2,2), plot(x,exponencial,'r','LineWidth',2)
```

Además, la distribución exponencial rige los tiempos de espera entre acontecimientos consecutivos de Poisson. El parámetro λ se toma como el número (o fracción) de acontecimientos de Poisson que ocurren por unidad de tiempo.

Ejemplo 7: Una secretaria recibe un promedio de 6 llamadas telefónicas por hora durante una jornada de trabajo ordinaria. Expresa el número de llamadas por hora como sucessos de Poisson y expresa el tiempo transcurrido entre 2 llamadas consecutivas que recibe (en horas) como una distribución exponencial.

$$X\text{=Número de llamadas por hora} \in Poisson(6)$$
$$Y\text{=Tiempo transcurrido entre 2 llamadas} \in Exponencial(6)$$

Una de las propiedades más importantes que caracteriza a la distribución exponencial es la llamada **falta de memoria**, que se expresa así:

$$P(X \geq x + \delta / X \geq x) = P(X \geq \delta).$$

Significa que la probabilidad de "duración" un tiempo adicional δ es independiente del tiempo transcurrido x y sólo depende de la "cuantía" de ese tiempo adicional.

3.3 Distribución normal

La distribución normal es la más importante y de mayor uso de todas las distribuciones continuas de probabilidad. Por múltiples razones se viene considerando la más idónea para modelizar una gran diversidad de mediciones de la Física, Química o Biología. Entre estas razones estudiaremos el teorema central del límite, que justifica la utilización de la normal como aproximación para las distribuciones de variables aleatorias bajo ciertas condiciones. En particular, lo usaremos para aproximar la Binomial y la Poisson.
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.

Definición 8. *Una variable aleatoria continua Z se dice que tiene* **distribución normal estándar**, *y lo denotamos $Z \in N(0,1)$, si su función de densidad viene dada por:*

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

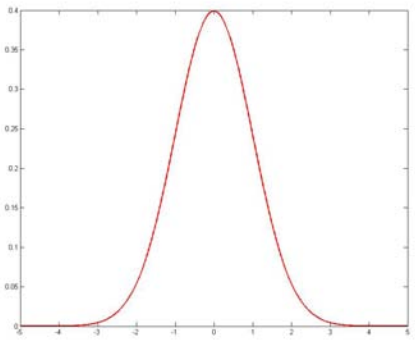

Figura 7: Función de densidad $f(z)$ para $Z \in N(0,1)$.

```
% Densidad de una N(0,1)
x=linspace(-5,5,1000);
densnormal=normpdf(x,0,1);
plot(x,densnormal,'r','LineWidth',2)
```

Propiedades: (Ver Figura 7)

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

Ejemplo 8: Supongamos entonces que $Z \in N(0,1)$. ¿Cómo calcularías $P(Z \leq 1.03)$?

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

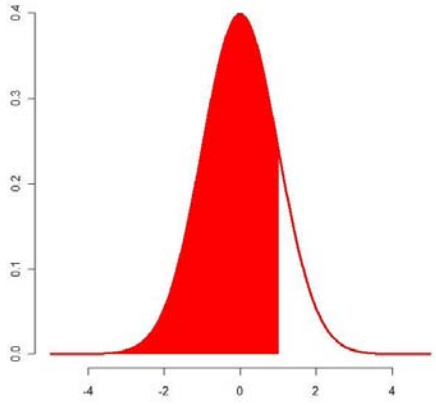

Figura 8: En rojo $P(Z \leq 1.03)$, para $Z \in N(0,1)$.

- La probabilidad inducida vendrá dada por el área bajo la densidad, ver Figura 8.
- Como no existe una expresión explícita para el área existen tablas con algunas probabilidades ya calculadas.
- Las tablas que nosotros utilizaremos proporcionan el valor de la función de distribución, $\Phi(z) = P(Z \leq z)$, de la normal estándar para valores positivos de z, donde z está aproximado hasta el segundo decimal.

Por lo tanto, para calcular $P(Z \leq 1.03)$, en el eje de las x marcamos el valor de Z (en este caso $z = 1.03$) e indicamos la probabilidad como el área que queda debajo de la campana de Gauss. (ver Figura 8). Buscaremos $P(Z \leq 1.03)$ en la tabla en el cruce entre la fila correspondiente a 1.0 y la columna correspondiente a 0.03. Así obtenemos $P(Z \leq 1.03) = 0.8465$.

Ejemplo 9: Supongamos que $Z \in N(0,1)$. Calcula usando las tablas de la normal estándar:

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

Ejemplo 10: Sea Z una variable aleatoria con distribución normal estándar. Halla los valores z_0 tales que

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

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

Definición 9. *Si $Z \in N(0,1)$ entonces*

$$X = \mu + \sigma Z \in N(\mu, \sigma^2)$$

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

Así, la función de densidad de X tendrá la misma forma de campana, será simétrica en torno a la media μ y sus puntos de inflexión serán $\mu - \sigma$ y $\mu + \sigma$. La forma más sencilla de calcular la función de densidad de una $N(\mu, \sigma^2)$ es calculando su función de distribución y después derivandor. La función de distribución de X viene dada por

$$F(x) = P(X \leq x) = P(\mu + \sigma Z \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

donde $\Phi(\cdot)$ es la función de distribución de la Normal estándar.
La función de densidad de una $N(\mu, \sigma^2)$ (ver Figura 9) es entonces

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

```
% Densidad de
% a)N(0,1)
% b)N(2,1)
% c)N(-1,0.5)
% d)N(0,2)
x=linspace(-5,5,1000);
normal_a=normpdf(x,0,1);
normal_b=normpdf(x,2,1);
normal_c=normpdf(x,-1,0.5);
normal_d=normpdf(x,0,2);
```

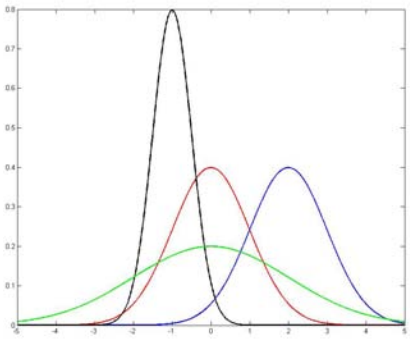

Figura 9: Funciones de densidad de variables normales con distintas medias y varianzas. En rojo densidad de una $N(0,1)$.

```
plot(x,normal_a,'r','LineWidth',2)
hold on;
plot(x,normal_b,'b','LineWidth',2)
plot(x,normal_c,'k','LineWidth',2)
plot(x,normal_d,'g','LineWidth',2)
```

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.

$$\text{Si } X \in N(\mu, \sigma^2) \text{ entonces } Z = \frac{X-\mu}{\sigma} \in N(0,1).$$

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

Ejemplo 11: Supongamos que $X \in N(5,4)$. ¿Cómo calcularías $P(X \leq 1)$?

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

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

4 Aproximación de otras distribuciones por la distribución normal

4.1 Aproximación de la distribución binomial por la distribución normal

Empezaremos aproximando la distribución binomial por la normal. Si mantenemos fija la probabilidad de éxito p e incrementamos el número de intentos n hasta valores muy grandes ($n \to +\infty$), al representar la distribución de probabilidad Binomial(n,p) observamos que adopta una forma de campana muy parecida a la función de densidad normal. Al mismo tiempo, para n grande no disponemos de tablas para las probabilidades binomiales y la fórmula se hace impracticable. Por todo esto, resulta tentador calcular esa probabilidad "como si la variable fuera normal". Sin embargo, necesitamos una justificación de que esto es válido, de que el error cometido es pequeño o se hace pequeño en ciertas condiciones. Esto lo aporta el Teorema de De Moivre-Laplace, precursor del teorema central del límite.

Teorema 1 (Teorema de De Moivre-Laplace)**.** *Tomemos una probabilidad de éxito fija p. Consideremos una sucesión de variables aleatorias $X_n \in Binomial(n,p)$, $n \in \{1,2,3,\ldots\}$ y $Z \in N(0,1)$. Entonces*

$$\lim_{n\to\infty} P\left(\frac{X_n - np}{\sqrt{np(1-p)}} \leq z\right) = P(Z \leq z) \qquad \forall\, z \in \mathbb{R}.$$

Denotaremos esto más brevemente así:

$$\frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{d} Z.$$

Nótese que hemos estandarizado la variable binomial. Utilizaremos el siguiente criterio:

- Si $n \geq 30$, $np \geq 5$ y $nq \geq 5$ entonces la Binomial de parámetros n y p puede ser aproximada por una normal de media $\mu = np$ y varianza $\sigma^2 = np(1-p)$.

Elemento de corrección por continuidad.
Al pasar de la distribución binomial a la normal estamos aproximando una distribución discreta por otra continua. Esto produce trastornos al calcular la probabilidad de un punto (no nula para la discreta y nula para la continua) o a la hora de discernir entre $<$ y $\leq$ en las expresiones. Para solventar esto, aproximamos la probabilidad de un valor de la distribución binomial $k \in \{0,1,2,\ldots,n\}$ por la del intervalo centrado en k de longitud unidad $(k-0.5, k+0.5]$, excepto para el 0 y el n, valores menor y mayor de la distribución discreta, respectivamente, cuyas probabilidades se aproximan por las de los intervalos $(-\infty, 0.5]$ y $(n-0.5, +\infty)$.

Ejemplo 12: Una empresa se dedica a la fabricación de bombas de turbina. La probabilidad de que una bomba pase todos los controles de calidad establecidos es 0.7. Si se fabrican 50 bombas, calcula la probabilidad de que superen los controles de calidad exactamente 38 bombas.

Sea X=Número de bombas que pasan todos los controles de calidad.
Tenemos que $X \in Binomial(50, 0.7)$. Por lo tanto, tendríamos que calcular

$$P(X = 38) = \begin{pmatrix} 50 \\ 38 \end{pmatrix} \cdot 0.7^{38} \cdot 0.3^{12}$$

Como $n \geq 30$, $np \geq 5$ y $nq \geq 5$, podemos aproximar la Binomial por una normal de media $\mu = np = 35$ y varianza $\sigma^2 = np(1-p) = 10.5$, es decir:

$$\begin{aligned} P(X = 38) \approx P(37.5 \leq N(35, 10.5) \leq 38.5) &= P\left(\frac{37.5-35}{\sqrt{10.5}} \leq Z \leq \frac{38.5-35}{\sqrt{10.5}}\right) \\ &= P(0.77 \leq Z \leq 1.08) \\ &= 0.8599 - 0.7794 \\ &= 0.0805 \end{aligned}$$

donde $Z \in N(0,1)$.

4.2 Teorema Central del Límite

En realidad el resultado anterior es un caso particular del denominado Teorema Central del Límite que afirma que, si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes y con la misma distribución X, entonces para n grande, la variable

$$S_n = X_1 + X_2 + \ldots + X_n$$

es aproximadamente normal con media $n\mu$ y varianza $n\sigma^2$ donde μ y σ^2 son la media y varianza de la variable X. Formalmente:

Teorema 2 (Teorema central del límite)**.** *Sea $X_1, X_2, \ldots, X_n, \ldots$ una sucesión de variables aleatorias independientes y con la misma distribución. Denotamos $S_n = X_1 + X_2 + \ldots + X_n$ a la suma, y sea $Z \in N(0,1)$. Entonces*

$$\frac{S_n - \mathbb{E}(S_n)}{\sqrt{\text{Var}(S_n)}} \xrightarrow{d} Z$$

4.3 Aproximación de la distribución de Poisson por la distribución normal

En particular el Teorema Central del Límite implica que la Poisson de parámetro λ puede ser aproximada, bajo ciertas condiciones, por la Normal de media $\mu = \lambda$ y varianza $\sigma^2 = \lambda$. Nosotros utilizaremos el siguiente criterio:

- Si $\lambda \geq 10$ entonces la Poisson de parámetro λ puede ser aproximada por una normal de media $\mu = \lambda$ y varianza $\sigma^2 = \lambda$.

Como la Poisson es discreta y la normal continua, recurriremos al elemento de corrección por continuidad.

5 Propiedades de aditividad para la Binomial, la Poisson y la Normal

Por último, presentamos una serie de propiedades muy útiles sobre la suma de variables independientes.

- Si $X_1 \in \text{Binomial}(n_1, p)$, $X_2 \in \text{Binomial}(n_2, p)$ y son independientes entonces $X_1 + X_2 \in \text{Binomial}(n_1 + n_2, p)$.
- Si $X_1 \in \text{Poisson}(\lambda_1)$, $X_2 \in \text{Poisson}(\lambda_2)$ y son independientes entonces $X_1 + X_2 \in \text{Poisson}(\lambda_1 + \lambda_2)$.
- Si $X_1 \in \text{N}(\mu_1, \sigma_1^2)$, $X_2 \in \text{N}(\mu_2, \sigma_2^2)$ y son independientes, entonces $X_1 + X_2 \in \text{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Estadística
Tema 7: INFERENCIA ESTADÍSTICA: ESTIMACIÓN PUNTUAL E INTERVALOS DE CONFIANZA

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Introducción

En el tema 1 hemos estudiado la Estadística Descriptiva, que se dedica al análisis y tratamiento de datos. A partir de ellos, resume, ordena y extrae los aspectos más relevantes de la información que contienen. Sin embargo, los objetivos de la Estadística son más ambiciosos. No nos conformamos con describir unos datos contenidos en una muestra sino que pretendemos extraer conclusiones para la población de la que fueron extraídos. A esta última tarea la llamamos **Inferencia Estadística**. Obtendremos las muestras de forma aleatoria y por tanto necesitaremos la Teoría de la Probabilidad vista en el tema 3 para elaborar nuestros argumentos. En los temas 4-6, vimos algunos modelos de variables discretas y continuas para una población y sus características más importantes, como la media y varianza poblacionales y otros parámetros. En este tema vamos a construir estimadores de los parámetros de interés a partir de una muestra y además, vamos a estudiar qué propiedades tienen que tener los estimadores para obtener buenas estimaciones.

2 Conceptos básicos.

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

Población. Es el conjunto homogéneo de individuos sobre los que se estudian una o varias características observables. Por ejemplo, la población de un país cuya intención de voto nos interesa. En otros casos (como por ejemplo, al estudiar la probabilidad de explosión en una reacción química), no está tan clara la existencia de una población, entendida como conjunto de individuos. En cualquier caso, el objetivo de la Inferencia Estadística es obtener información sobre una población.
Muestra. Es un subconjunto extraído de la población, al que podemos observar. Múltiples razones nos imposibilitan observar toda la población. Por ese motivo, extraemos una muestra y con ella obtenemos información sobre toda la población.
Tamaño de la población o de la muestra. Es el número de individuos que los forman, en cada caso.

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

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

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

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

   c) **Contrastes de Hipótesis**. Se trata de responder a preguntas muy concretas sobre la población, y se reducen a un problema de decisión sobre la veracidad de ciertas hipótesis. Por ejemplo, nos podemos preguntar si nuestra proporción de votantes superará el 40 %, umbral que nos otorga la mayoría absoluta en el parlamento.

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

3 Planteamiento general del problema de inferencia paramétrica.

Consideramos un experimento aleatorio sobre el cual medimos una cierta variable aleatoria, que denotaremos por X. El objetivo es estudiar la variable aleatoria X, cuya función de distribución F es en mayor o menor grado desconocida.

Ejemplo 1: Provocamos una reacción química y medimos el calor que se desprende X. Nos interesa saber qué valores puede tomar y con qué probabilidades, esto es, su distribución.

Ejemplo 2: Queremos conocer la proporción de individuos con cierta característica en una población. El experimento consiste en extraer uno al azar, y así la distribución de Bernoulli que indica la presencia de la característica tiene como parámetro la proporción desconocida.

Suponemos que la distribución de X, aún siendo desconocida, sigue un modelo como los del tema anterior. En el caso del calor desprendido en la reacción del Ejemplo 1, podría ser normal, y en el caso de la proporción del Ejemplo 2, es claramente de Bernoulli. Así, el problema se reduce a averiguar los parámetros.
Para hacer inferencia, repetimos el experimento n veces en idénticas condiciones y de forma independiente. Una **muestra aleatoria simple** de tamaño n está formada por n variables

$$X_1, X_2, \cdots, X_n$$

independientes y con la misma distribución que X.
Llamamos **realización muestral** a los valores concretos que tomaron las n variables aleatorias *después* de la obtención de la muestra.
Un **estadístico** es una función de la muestra aleatoria, y por tanto nace como resultado de cualquier operación efectuada sobre la muestra. Es también una variable aleatoria y por ello tendrá una cierta distribución, que se denomina **distribución del estadístico en el muestreo**.
Para resolver el problema de estimación puntual, esto es, para aventurar un valor del parámetro poblacional desconocido, escogemos el valor que ha tomado un estadístico calculado sobre nuestra realización muestral. Al estadístico escogido para tal fin le llamamos **estimador** del parámetro. Al valor obtenido con una realización muestral concreta se le llama **estimación**.
El problema radica, por lo tanto, en elegir un "buen estimador", es decir, una función de la muestra con buenas propiedades.
En general, un buen estimador de un parámetro poblacional (media, proporción de individuos que presentan cierta característica,...) va a ser el correspondiente parámetro muestral (media de la muestra, proporción de individuos que presentan la característica en la muestra,...).

4 Estimación puntual de una proporción. Propiedades de un estimador.

Resolvemos ahora dos problemas prácticos de inferencia. El primero consiste en obtener información sobre la proporción de individuos con cierta característica en una población (entendida como conjunto de individuos), mediante la extracción de una muestra **con reemplazamiento**. El segundo consiste en obtener información sobre la probabilidad de ocurrencia de un suceso (éxito), mediante la realización de n intentos independientes y en idénticas condiciones del mismo.
En ambos casos nos ajustamos al planteamiento general de inferencia paramétrica presentado en la sección anterior. La muestra está formada por n variables $X_1, \ldots, X_n$ independientes y con distribución Bernoulli(p). El parámetro p es la proporción poblacional desconocida. Como ya se ha sugerido, el estimador razonable es la

proporción muestral

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

Observa que el numerador tiene distribución Binomial(n, p). Por lo tanto $\hat{p}$ puede tomar los valores $0, 1/n, 2/n, \ldots, 1$ y las probabilidades son las mismas de la Binomial(n, p).
Observamos en primer lugar que $\mathbb{E}(\hat{p}) = p$. Esta nos parece una buena propiedad del estimador.

Definición 1. *Llamamos* **sesgo** *de un estimador $\hat{\theta}$ para un parámetro poblacional θ a*

$$Sesgo\left(\hat{\theta}\right) = \mathbb{E}\left(\hat{\theta}\right) - \theta,$$

y diremos que el estimador es **insesgado** *si su sesgo vale cero.*

Por tanto, $\hat{p}$ es un estimador insesgado de p. Si un estimador presenta sesgo, nos sentimos tentados a efectuar un cambio de localización sobre dicho estimador.
Ahora que sabemos que $\hat{p}$ está centrado en torno a p, nos interesa que su dispersión sea pequeña. Lo ideal sería que $\mathbb{E}(\hat{p}) = p$ y $Var(\hat{p}) = 0$. En ese caso $\hat{p}$ sólo tomaría un valor, que sería p y nunca habría error. En nuestro caso

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

que es distinta de cero, pero $\lim_{n\to\infty} Var(\hat{p}) = 0$. Esto significa que al aumentar el tamaño muestral el estimador se aproxima al parámetro poblacional, lo cual constituye una justificación fundamental del método estadístico.

Definición 2. *Definimos el* **error cuadrático medio** *de un estimador $\hat{\theta}$ para un parámetro poblacional θ como*

$$\mathbb{E}\left((\hat{\theta} - \theta)^2\right) = \left[Sesgo(\hat{\theta})\right]^2 + Var(\hat{\theta})$$

y diremos que dicho estimador es **consistente** *si $\lim_{n\to\infty} \mathbb{E}\left((\hat{\theta} - \theta)^2\right) = 0$.*

Vemos claramente que la proporción muestral $\hat{p}$ es un estimador consistente de p.

5 Concepto de intervalo de confianza. Intervalo de confianza para una proporción.

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

Definición 3. *Un* **intervalo de confianza** *es un intervalo construido en base a la muestra y, por tanto, aleatorio, que contiene al parámetro con una cierta probabilidad, conocida como* **nivel de confianza**.
Sea θ el parámetro desconocido y L_1 y L_2 los extremos del intervalo (que son estadísticos por estar el intervalo de confianza construido en base a la muestra). Se dice que $[L_1, L_2]$ tiene un nivel de confianza $1 - \alpha$, siendo $\alpha \in [0, 1]$, si $P(L_1 \leq \theta \leq L_2) \geq 1 - \alpha$.

El 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.
Construimos ahora un intervalo de confianza para p. Nos basamos en la proporción muestral, $\hat{p}$. Recordamos que la distribución binomial se puede aproximar por la normal cuando n es suficientemente grande, manteniendo p fija. Pero en nuestro caso p está fija y, como en cualquier problema de inferencia, el tamaño muestral n debe

ser moderado o grande. Dado que $\hat{p}$ sólo consiste en dividir a la binomial por un número real, n, su distribución también se puede aproximar por la normal, con su misma media y desviación típica. Por tanto,

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

A ese estadístico le llamamos **estadístico pivote** y al método que estamos usando para construir el intervalo de confianza **método pivotal**.
Denotemos $z_{\alpha/2}$ al número real tal que $P(Z > z_{\alpha/2}) = \alpha/2$, siendo $Z \in N(0,1)$, ver Figura 1. El valor de $z_{\alpha/2}$ se obtiene de las tablas de la normal. Entonces

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

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

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

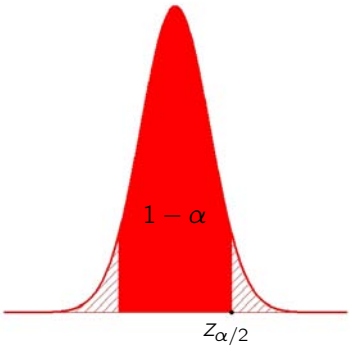

Figura 1: Denotamos $z_{\alpha/2}$ el número real tal que $P(Z > z_{\alpha/2}) = \alpha/2$, siendo $Z \in N(0,1)$.

Ejemplo 3: El fabricante de un determinado tipo de bombillas desea averiguar la proporción de bombillas defectuosas que produce. Para ello selecciona y prueba 200 unidades y descubre un total de 80 unidades defectuosas. ¿Cómo podría estimar la proporción de bombillas defectuosas? Calcula un intervalo de confianza para la proporción al 95 %.

6 Estimación puntual de la media y la varianza de una población normal. Distribuciones en el muestreo de una población normal.

Consideramos ahora el problema de inferencia paramétrica en una población normal. En esta situación disponemos de una muestra aleatoria simple

$$X_1, \ldots, X_n$$

formada por n variables aleatorias independientes y con la misma distribución $N(\mu, \sigma^2)$. El problema de inferencia consiste en averiguar los parámetros μ, media poblacional, y σ, desviación típica poblacional.

6.1 Estimador de la media

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

$$\bar{X} = \frac{1}{n} \sum X_i.$$

- La media de $\bar{X}$ es $\mathbb{E}(\bar{X}) = \mu$.
- La varianza de $\bar{X}$ es $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$.
- Por la propiedad de aditividad de la distribución normal y dado que $\bar{X}$, la media muestral, es la suma de n variables independientes, entonces la media muestral tiene distribución normal $\bar{X} \in N(\mu, \sigma^2/n)$.

De esto se deduce que la media muestral es un estimador insesgado de la media poblacional y que su varianza es la poblacional dividida por n. Por tanto, la dispersión será tanto mayor cuanto mayor sea la de la población y decrece tendiendo a cero cuando el tamaño muestral aumenta. De este modo vemos también que la media muestral es un estimador consistente de la media.

6.2 Estimación de la varianza cuando la media es conocida

Si la media μ es conocida entonces el estimador natural de la varianza es la varianza muestral, definida como la media de las desviaciones al cuadrado de los datos muestrales *respecto a la media de la población*:

$$S_\mu^2 = \frac{1}{n} \sum (X_i - \mu)^2 .$$

Se puede comprobar que $\mathbb{E}(S_\mu^2) = \sigma^2$ y, por lo tanto, la varianza muestral es un estimador insesgado de la varianza.

6.3 Estimación de la varianza cuando la media es desconocida

Si la media μ es desconocida entonces, para calcular la varianza muestral, debemos reemplazar la media de la población por la media de la muestra:

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

Aunque S^2 sería el estimador natural de la varianza cuando μ es desconocida, se puede comprobar que $\mathbb{E}(S^2) = (n-1)\sigma^2/n$ y, por lo tanto S^2 no es insesgado, lo cual no lo hace un estimador apropiado.
Definimos la **cuasivarianza muestral**:

$$S_c^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2 .$$

Lo único que la diferencia de la varianza muestral es la sustitución del denominador n por el denominador $n-1$. La cuasivarianza muestral es, pues, un estimador alternativo de la varianza. Se puede comprobar que $\mathbb{E}(S_c^2) = \sigma^2$, esto es, que la cuasivarianza muestral es un estimador insesgado de la varianza de la población.

7 Intervalos de confianza para la media y varianza de una población normal.

7.1 Media con varianza conocida

La distribución de la media muestral permite obtener como pivote

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

y extraer de este pivote un intervalo de confianza para la media cuando la varianza es conocida, de la forma:

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

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

7.2 Media con varianza desconocida

Podemos construir un pivote para la media cuando la varianza es desconocida de la siguiente manera:

$$\frac{\bar{X}-\mu}{S_c/\sqrt{n}} \in T_{n-1}.$$

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

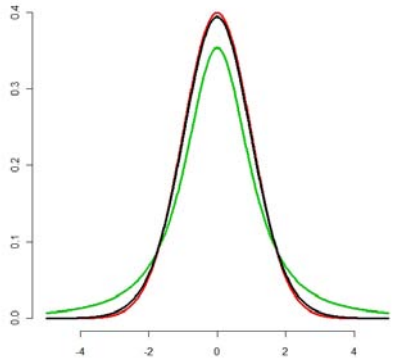

Figura 2: En verde densidad de una T de Student con 2 grados de libertad, en rojo densidad de una N(0,1) y en negro densidad de una T de Student con 20 grados de libertad

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 $k \to \infty$.

Del pivote anterior se deduce un intervalo de confianza para la media cuando la varianza es desconocida, de la forma:

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

siendo $t_{\alpha/2}$ el valor que deja una probabilidad $\alpha/2$ a su derecha en la distribución T_{n-1}, ver Figura 3.

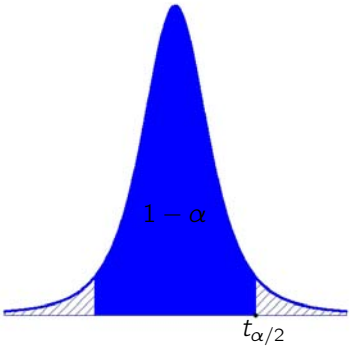

Figura 3: Denotamos $t_{\alpha/2}$ el número real tal que $P(T_k > t_{\alpha/2}) = \alpha/2$, siendo T_k una variable T de Student con k grados de libertad.

El precio que tenemos que pagar por no conocer la varianza es que, como $t_{\alpha/2} > z_{\alpha/2}$, el intervalo de confianza para la media con varianza desconocida suele resultar más amplio que el construido con varianza conocida.

7.3 Varianza con media conocida

Podemos construir un pivote para la varianza cuando la media es conocida así:

$$\frac{nS_\mu^2}{\sigma^2} \in \chi_n^2$$

La distribución χ_n^2: La distribución Chi-cuadrado (o ji-cuadrado) con n grados de libertad es un modelo de variable aleatoria continua. En la Figura 4 se representa la función de densidad de variables Chi-cuadrado para diferentes grados de libertad.

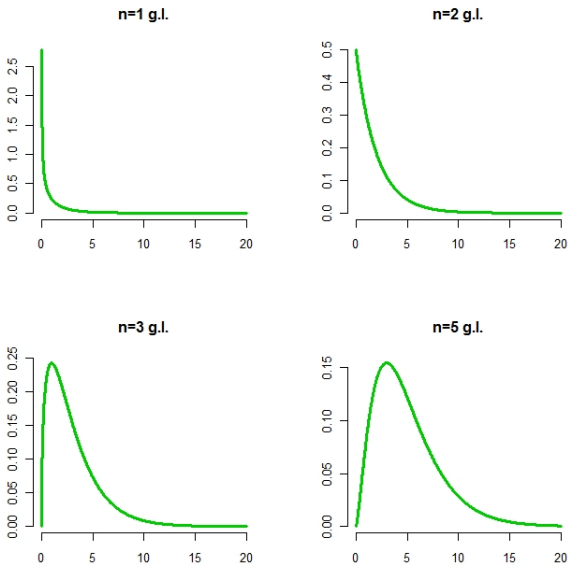

Figura 4: En verde densidades de variables χ_n^2 para distintos valores de n.

Propiedades:

1. Si $Z_1, \ldots, Z_n \in N(0,1)$ son variables aleatorias independientes, entonces $X = Z_1^2 + \cdots + Z_n^2$ tiene distribución Chi-cuadrado con n grados de libertad.
2. La variable Chi-cuadrado toma valores $[0, +\infty)$.
3. La distribución T Chi-cuadrado es asimétrica.

Del pivote anterior se deduce un intervalo de confianza para la varianza cuando la media es conocida, de la forma:

$$\left(\frac{nS_\mu^2}{\chi^2_{\alpha/2}} , \frac{nS_\mu^2}{\chi^2_{1-\alpha/2}} \right)$$

siendo $\chi^2_{\alpha/2}$ $\chi^2_{1-\alpha/2}$ los valores que dejan probabilidades respectivas $\alpha/2$ y $1-\alpha/2$ a la derecha en la distribución χ^2_n, ver Figura 5.

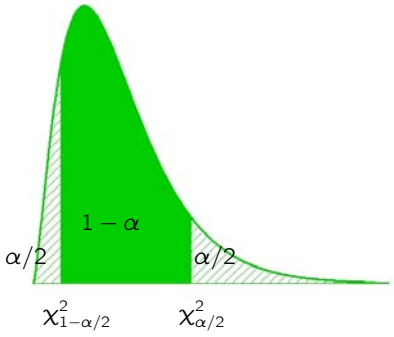

Figura 5: Denotamos $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ los valores que dejan probabilidades respectivas $\alpha/2$ y $1-\alpha/2$ a su derecha en la distribución χ^2_n.

7.4 Varianza con media desconocida

En base al pivote

$$\frac{(n-1)S_c^2}{\sigma^2} \in \chi^2_{n-1}$$

obtenemos el intervalo de confianza para la varianza con media desconocida así:

$$\left(\frac{(n-1)S_c^2}{\chi^2_{\alpha/2}} , \frac{(n-1)S_c^2}{\chi^2_{1-\alpha/2}} \right)$$

siendo $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ los valores que dejan probabilidades respectivas $\alpha/2$ y $1-\alpha/2$ a su derecha en la distribución χ^2_{n-1}.

Estadística
Tema 8: CONTRASTE DE HIPÓTESIS

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Introducción

Los procedimientos de inferencia que hemos realizado hasta ahora se resumen en dos: la **estimación puntual** y los i**ntervalos de confianza**. Con la estimación puntual se obtienen valores concretos que sirven de estimaciones de los parámetros poblacionales de interés, por ejemplo, estimamos la media poblacional, μ, con la media muestral, $\bar{x}$. Con los intervalos de confianza se obtienen regiones aleatorias que contienen a los parámetros de interés con cierta probabilidad, por ejemplo, el intervalo de confianza con nivel de confianza $1-\alpha$ para la media μ de una población normal es $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, cuando la desviación σ es conocida. La otra gran tarea de la Inferencia Estadística consiste en responder a preguntas muy concretas sobre la población. Por ejemplo, ¿hay la misma proporción de vehículos diésel que de gasolina? Como veremos se plantean en términos de unas hipótesis que debemos aceptar o rechazar. Y esta decisión la tomaremos en base a una realización muestral. Cuando los datos muestrales discrepen mucho de la hipótesis (en nuestro ejemplo, cuando la proporción muestral de vehículos diésel sea muy distinta de la proporción de vehículos de gasolina) rechazaremos la hipótesis.

2 Planteamiento y resolución de un contraste de hipótesis.

Se tiene una hipótesis de trabajo y una muestra de observaciones, y se trata de decidir si la hipótesis planteada es compatible con lo que se puede aprender del estudio de los valores muestrales, es decir, decidir si la muestra que se obtuvo está de acuerdo con la hipótesis de trabajo.
En nuestro ejemplo la hipótesis que queremos contrastar sería $H_0: \ p = 0.5$, siendo p la proporción de vehículos diésel. Observamos que esta hipótesis y su alternativa, que sería $H_a: \ p \neq 0.5$, no son tratadas de igual modo. Damos por cierta la hipótesis H_0 y vemos si la muestra aporta pruebas en su contra.
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_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_a es porque hemos encontrado pruebas significativas a partir de la muestra. Más adelante definimos con precisión qué entendemos por pruebas significativas.
Una **Hipótesis simple** es la que está constituida por un único punto. En el ejemplo, $H_0: \ p = 0.5$ es una hipótesis simple. Si una hipótesis consta de más de un punto la llamaremos **hipótesis compuesta**. La hipótesis "la proporción de vehículos diésel es menor o igual que la de vehículos de gasolina" sería compuesta, pues viene expresada así: $H_0: \ p \leq 0.5$.
Volviendo al problema de decisión que supone el contraste de hipótesis, prestemos atención la siguiente tabla.

Decisión \ Realidad	H_0 cierta	H_0 falsa
Aceptar	Correcto	Error tipo II
Rechazar	Error tipo I	Correcto

Observamos que se puede tomar una decisión correcta o errónea. Llamamos **error de tipo I** al que cometemos cuando rechazamos la hipótesis nula, siendo cierta. **Error de tipo II** es el que cometemos cuando aceptamos la hipótesis nula, siendo falsa.
Nivel de significación: Es la probabilidad del error de tipo I. Lo denotamos por α:

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

Potencia: Es la probabilidad de detectar que una hipótesis es falsa. La denotamos por β:

$$\beta = P(\text{Rechazar } H_0/H_0 \text{ es falsa}) = 1 - P(\text{Error de tipo II})$$

Debemos adoptar un criterio que, en base a la muestra, nos permita decidir si aceptamos o rechazamos la hipótesis nula. Obviamente, queremos minimizar las probabilidades de los errores de tipo I y II. Pues bien, la forma de minimizar la probabilidad del error de tipo I (el nivel de significación) es mediante un criterio que acepte H_0 la mayor parte de las veces. Sin embargo, así se incrementa la probabilidad del error de tipo II, esto es, disminuye la potencia del test. Una forma de proceder ante un problema con dos objetivos como es éste, consiste en fijar el nivel de significación y escoger el criterio que nos proporcione la mayor potencia posible.
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) y la **región de aceptación** que tiene probabilidad $1-\alpha$ (bajo H_0).

- Si el valor del estadístico cae en la región de aceptación, no existen razones suficientes para rechazar la hipótesis nula con un nivel de significación α, y el contraste se dice **estadísticamente no significativo**, es decir no existe evidencia a favor de H_a.
- 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_a.

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

1. Especificar las hipótesis nula H_0 y alternativa H_a.
2. Elegir un estadístico de contraste apropiado, $T(X_1, ..., X_n)$, que sea una medida de la discrepancia entre la hipótesis y la muestra.
3. Fijar el nivel de significación α en base a cómo de importante se considere rechazar H_0 cuando realmente es cierta.
4. Prefijado y elegido $T(X_1, ..., X_n)$, construir las regiones de aceptación y rechazo, según se trate de un contraste uni o bilateral.
5. Tomar la muestra $\{x_1, ..., x_n\}$ y evaluar el estadístico de contraste $T(x_1, ..., x_n)$
6. Concluir si el test es estadísticamente significativo (se rechaza H_0) o no al nivel de significación α según el valor del estadístico $T(x_1, ..., x_n)$ se ubique en la región de rechazo o no, respectivamente.

3 Contraste de hipótesis sobre una proporción.

Queremos contrastar hipótesis como las propuestas como ejemplos en la sección anterior. La muestra en la que basaremos nuestra decisión está constituida por n variables independientes con distribución Bernoulli(p), como en el planteamiento general de Inferencia Paramétrica sobre una proporción. Contrastaremos dos tipos de hipótesis sobre p:

- Hipótesis simple. $H_0: \ p = p_0$, siendo p_0 una proporción conocida. En el ejemplo, hemos considerado $H_0: \ p = 0.5$, esto es p_0 era 0.5.
- Hipótesis compuesta. $H_0: \ p \leq p_0$ ó $H_0: \ p \geq p_0$, siendo p_0 una proporción conocida.

Rechazaremos la hipótesis simple $H_0: \ p = p_0$ si la proporción muestral discrepa mucho de p_0, tanto por ser mucho mayor como por ser mucho menor. Estandarizando $\hat{p}$ obtenemos un estadístico con distribución conocida y tabulada. Sobre este estadístico construimos la región de aceptación y la región de rechazo (o región crítica).

Pues bien, hemos dicho que fijamos el nivel de significación y escogemos el criterio que maximiza la potencia. Supongamos entonces que hemos determinado ya α. Si la hipótesis fuera cierta, esto es, $p = p_0$ entonces el estadístico tiene distribución dependiente del parámetro p_0 (conocido). Podemos buscar entonces $z_{\alpha/2}$ en las tablas de la normal de forma que la región crítica tenga probabilidad α, pues ésta sería la probabilidad de que, siendo $p = p_0$, el estadístico cayera en esa región y en consecuencia se rechazara la hipótesis. El criterio final sería (ver Figura 1):

$$\text{Rechazamos} \quad H_0 : \ p = p_0 \qquad \text{si} \quad \frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0\,(1-p_0)}{n}}} > z_{\alpha/2}$$

En este caso de hipótesis nula simple, la región crítica se descompone en dos trozos y, por ello, hablamos de contraste **bilateral**. Si la hipótesis nula fuera compuesta, por ejemplo $H_0 : \ p \leq p_0$, sólo rechazaríamos cuando $\hat{p}$ fuera mucho mayor que p_0, y la región crítica tendría un único trozo. En esta ocasión diremos que se trata de un contraste de hipótesis **unilateral**. El criterio sería (ver Figura 1):

$$\text{Rechazamos} \quad H_0 : \ p \leq p_0 \qquad \text{si} \quad \frac{\hat{p} - p_0}{\sqrt{\frac{p_0\,(1-p_0)}{n}}} > z_{\alpha}$$

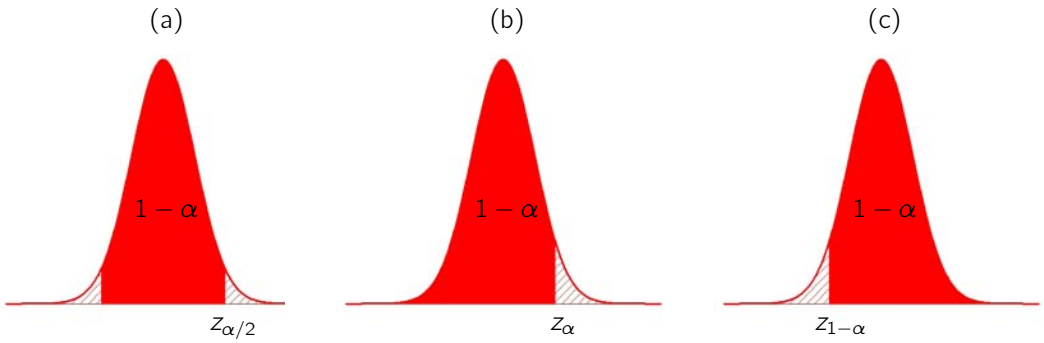

Figura 1: Función de densidad de una N(0,1). (a) Región de aceptación y rechazo para el estadístico del contraste $H_0 : p = p_0$. (b) Región de aceptación y rechazo para el estadístico del contraste $H_0 : p \leq p_0$. (c) Región de aceptación y rechazo para el estadístico del contraste $H_0 : p \geq p_0$.

Por último, en muchas ocasiones, en lugar de fijar el nivel de significación, se proporciona la probabilidad que contendría una región crítica limitada por el valor observado del estadístico. A esta probabilidad le llamamos **nivel crítico** y viene a representar el mayor nivel de significación que permite aceptar la hipótesis nula.

Ejemplo 1: Una empresa farmacéutica quiere comercializar un medicamento que cura cierta dolencia. Se sabe que el 40 % de los pacientes se curan sin tomar este medicamento. La empresa debe probar que su medicamento es eficaz y para ello administra el medicamento a 100 pacientes, de los cuales se curan 50.

Ejemplo 2: En un ecosistema dos especies de aves A y B se encuentran en equilibrio, con igual proporción de ambas. Se teme que los últimos acontecimientos hayan alterado el equilibrio, y para comprobarlo, se toma una muestra de 1600 aves, de las cuales 720 son de la especie A. ¿Podemos concluir que se ha alterado el equilibrio?

Ejemplo 3: Las normas de calidad no permiten que la proporción de unidades defectuosas supere el 5 %. Una inspección toma una muestra de 400 unidades y encuentra 16 defectuosas. ¿Constituye este resultado una prueba significativa de que no se respeta la norma de calidad?

Como comentario general, debemos ser conscientes de que cuando rechazamos la hipótesis nula en base a una muestra es porque nos ha aportado pruebas significativas a un nivel α de que esa hipótesis no es cierta. Por pruebas significativas a un nivel α entendemos que si la hipótesis fuera cierta, la probabilidad de que el resultado muestral discrepara tanto de ella sería tan pequeña como α. Sin embargo, cuando aceptamos una hipótesis nula no es porque haya pruebas a su favor, sino porque no las hubo en su contra.

4 Contrastes de hipótesis para dos proporciones

Supongamos que una población contiene una proporción p_1 de individuos que presentan cierta característica, mientras que otra población contiene una proporción p_2 de dicha característica.
Extraemos una muestra aleatoria simple en cada población

$$\begin{aligned} X_{11}, \ldots, X_{1n_1} &\in \text{Bernoulli}\,(p_1) \qquad \text{independientes} \\ X_{21}, \ldots, X_{2n_2} &\in \text{Bernoulli}\,(p_2) \qquad \text{independientes} \end{aligned}$$

y además ambas muestras son extraídas de manera independiente entre sí.
Estimaremos cada proporción poblacional mediante la correspondiente proporción muestral:

$$\begin{aligned} \hat{p}_1 &= \frac{X_{11} + \ldots + X_{1n_1}}{n_1} \\ \hat{p}_2 &= \frac{X_{21} + \ldots + X_{2n_2}}{n_2} \end{aligned}$$

Pensemos en el contraste de la hipótesis nula de que las dos proporciones poblacionales son iguales. Si dicha hipótesis nula $H_0 : p_1 = p_2$ es cierta, entonces

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \sim N(0,1)$$

Rechazaremos la hipótesis nula de igualdad de las proporciones cuando las proporciones muestrales sean muy distintas, y si a eso añadimos un nivel de significación α prefijado, debemos actuar así:

$$\text{Rechazar } H_0 : p_1 = p_2 \qquad \text{si} \quad \frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} > z_{\alpha/2}$$

El contraste unilateral consistiría en:

$$\text{Rechazar } H_0 : p_1 \leq p_2 \qquad \text{si} \quad \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} > z_{\alpha}$$

5 Contrastes en una población normal

Queremos contrastar hipótesis relativas a la media y la varianza de una población $N(\mu, \sigma^2)$. Para ello, tomamos una muestra aleatoria simple:

$$X_1, \ldots, X_n \in N(\mu, \sigma^2) \qquad \text{independientes}$$

5.1 Contraste sobre la media

5.1.1 Media con varianza conocida

Supongamos que la varianza σ^2 es conocida, y se desea contrastar una hipótesis relativa a la media, μ, por ejemplo, que la media toma cierto valor conocido $H_0 : \mu = \mu_0$. Si dicha hipótesis nula $H_0 : \mu = \mu_0$ es cierta, entonces

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

El sentido común nos aconseja rechazar la hipótesis nula de que la media poblacional es μ_0 cuando la media muestral sea muy distinta de μ_0. Si además debemos respetar un nivel de significación α prefijado, debemos actuar así:

$$\text{Rechazar } H_0 : \mu = \mu_0 \qquad \text{si} \quad \frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2}$$

El contraste unilateral consistiría en:

$$\text{Rechazar } H_0 : \mu \geq \mu_0 \qquad \text{si} \quad \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < -z_{\alpha}$$

5.1.2 Media con varianza desconocida

Podemos repetir toda la argumentación anterior, con la salvedad de que, cuando la varianza es desconocida, no podemos usar σ y en su lugar debemos emplear un estimador adecuado, por ejemplo, S_c. Sabemos que este cambio afecta a la distribución, que pasa a ser T de Student. Así, si $H_0 : \mu = \mu_0$ es cierta, entonces

$$\frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} \in T_{n-1}$$

y la regla de decisión será:

$$\text{Rechazar } H_0 : \mu = \mu_0 \qquad \text{si} \quad \frac{|\bar{X} - \mu_0|}{S_c/\sqrt{n}} > t_{\alpha/2}$$

De igual modo, el contraste unilateral consistiría en:

$$\text{Rechazar } H_0 : \mu \geq \mu_0 \qquad \text{si} \quad \frac{\bar{X} - \mu_0}{S_c/\sqrt{n}} < -t_{\alpha}$$

En la Figura 2 se muestran las regiones de aceptación y rechazo de los contrastes sobre la media de una población con varianza desconocida.

5.2 Contraste sobre la varianza

5.2.1 Varianza con media conocida

Si la hipótesis nula $H_0 : \sigma^2 = \sigma_0^2$ es cierta, entonces

$$\frac{nS_\mu^2}{\sigma_0^2} \in \chi_n^2$$

y, por tanto, la regla de decisión será:

$$\text{Rechazar } H_0 : \sigma^2 = \sigma_0^2 \qquad \text{si} \quad \frac{nS_\mu^2}{\sigma_0^2} > \chi^2_{\alpha/2} \;\; \text{o} \;\; \frac{nS_\mu^2}{\sigma_0^2} < \chi^2_{1-\alpha/2}$$

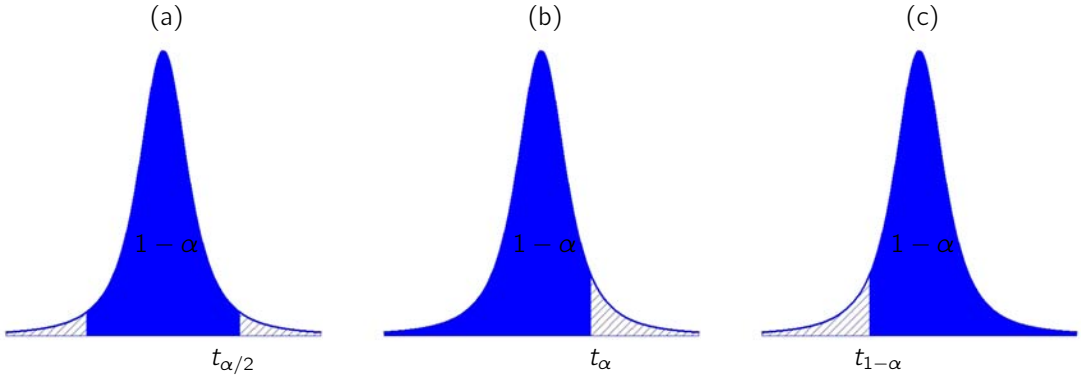

Figura 2: Función de densidad de una T_{n-1}. (a) Región de aceptación y rechazo para el estadístico del contraste $H_0 : \mu = \mu_0$. (b) Región de aceptación y rechazo para el estadístico del contraste $H_0 : \mu \leq \mu_0$. (c) Región de aceptación y rechazo para el estadístico del contraste $H_0 : \mu \geq \mu_0$.

mientras que para el contraste unilateral será:

$$\text{Rechazar } H_0 : \sigma^2 \leq \sigma_0^2 \qquad \text{si } \frac{nS_\mu^2}{\sigma_0^2} > \chi_\alpha^2$$

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

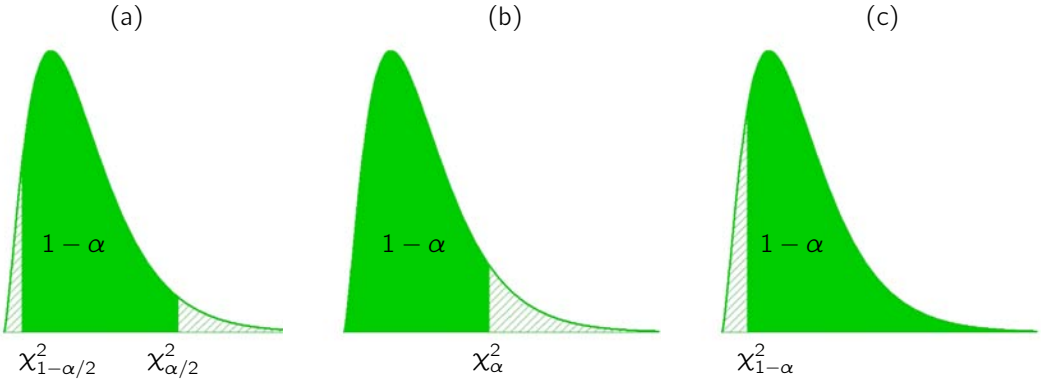

Figura 3: Función de densidad de una χ_n^2. (a) Región de aceptación y rechazo para el estadístico del contraste $H_0 : \sigma^2 = \sigma_0^2$. (b) Región de aceptación y rechazo para el estadístico del contraste $H_0 : \sigma^2 \leq \sigma_0^2$. (c) Región de aceptación y rechazo para el estadístico del contraste $H_0 : \sigma^2 \geq \sigma_0^2$.

5.2.2 Varianza con media desconocida

Si la hipótesis nula $H_0 : \sigma^2 = \sigma_0^2$ es cierta, entonces

$$\frac{(n-1)S_c^2}{\sigma_0^2} \in \chi_{n-1}^2$$

y, por tanto, la regla de decisión será:

$$\text{Rechazar } H_0 : \sigma^2 = \sigma_0^2 \qquad \text{si } \frac{(n-1)S_c^2}{\sigma_0^2} > \chi^2_{\alpha/2} \text{ o } \frac{(n-1)S_c^2}{\sigma_0^2} < \chi^2_{1-\alpha/2}$$

mientras que para el contraste unilateral será:

$$\text{Rechazar } H_0 : \sigma^2 \leq \sigma_0^2 \qquad \text{si } \frac{(n-1)S_c^2}{\sigma_0^2} > \chi^2_{\alpha}$$

6 Contrastes referidos a dos poblaciones normales

6.1 Contrastes referidos a dos poblaciones normales basados en muestras independientes

Consideremos el siguiente ejemplo:

Ejemplo 4: El Verapamil y el Nitroprusside son dos productos utilizados para reducir la hipertensión. Para compararlos, unos pacientes son tratados con Verapamil y otros con Nitroprusside. Los resultados obtenidos se muestran en la siguiente tabla, donde:

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

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

Admitiendo normalidad y sabiendo que ambas variables tienen la misma desviación típica, ¿se puede aceptar la igualdad de medias?.

La situación descrita en el Ejemplo 4 responde al siguiente modelo general. Pensemos en dos poblaciones normales, con sus respectivas medias y varianzas: $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$. Queremos contrastar hipótesis que comparen sus medias, μ_1 y μ_2.
Extraemos una muestra aleatoria simple en cada población

$$\begin{aligned} &X_{11}, \ldots, X_{1n_1} \in N\left(\mu_1, \sigma_1^2\right) && \text{independientes} \\ &X_{21}, \ldots, X_{2n_2} \in N\left(\mu_2, \sigma_2^2\right) && \text{independientes} \end{aligned}$$

y además **ambas muestras son extraídas de manera independiente entre sí**.

6.1.1 Contraste sobre la igualdad de medias con varianzas conocidas

Supongamos que conocemos las varianzas de ambas poblaciones σ_1^2 y σ_2^2 y que queremos contrastar la hipótesis $H_0 : \mu_1 = \mu_2$. Fijamos el nivel de significación y escogemos el criterio que maximiza la potencia. Supongamos entonces que hemos determinado ya α. Si dicha hipótesis nula $H_0 : \mu_1 = \mu_2$ es cierta, entonces

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

El sentido común nos aconseja rechazar la hipótesis nula de que las medias son iguales cuando $\bar{X}_1 - \bar{X}_2$ sea muy distinta de cero. Por tanto, rechazaremos $H_0 : \mu_1 = \mu_2$ (ver Figura 4) si

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

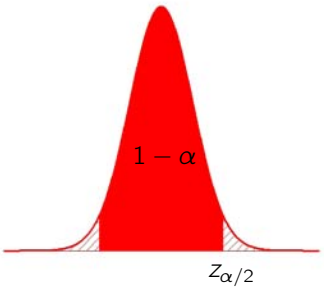

Figura 4: Función de densidad de una N(0,1). Región de aceptación y rechazo para el estadístico del contraste $H_0 : \mu_1 = \mu_2$ en el caso de dos poblaciones normales basados en muestras independientes con varianzas conocidas.

6.1.2 Contraste sobre la igualdad de medias con varianzas desconocidas pero iguales

Supongamos ahora que desconocemos las varianzas de ambas poblaciones pero que podemos asumir que dichas varianzas son iguales. Queremos contrastar la hipótesis $H_0 : \mu_1 = \mu_2$. Si suponemos que las varianzas de las dos poblaciones son iguales el mejor estimador de la varianza será:

$$S_T^2 = \frac{(n_1 - 1)S_{c1}^2 + (n_2 - 1)S_{c2}^2}{n_1 + n_2 - 2},$$

que no es más que una adecuada ponderación de los mejores estimadores de cada población (en la ecuación anterior, S_{c1}^2 y S_{c2}^2 denotan la cuasivarianza muestral de primera y segunda población, respectivamente). Se puede demostrar que

$$\frac{\bar{X}_1 - \bar{X}_2}{S_T\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \in T_{n_1+n_2-2}.$$

Por los mismos argumentos que en casos anteriores, rechazaremos $H_0 : \mu_1 = \mu_2$ (ver Figura 5) si

$$\frac{|\bar{X}_1 - \bar{X}_2|}{S_T\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha/2}$$

donde $t_{\alpha/2}$ es el valor que deja una probabilidad $\alpha/2$ a su derecha en la distribución $T_{n_1+n_2-2}$.

6.1.3 Contraste sobre la igualdad de medias con varianzas desconocidas y desiguales

Supongamos ahora que desconocemos las varianzas de ambas poblaciones y que no podemos asumir que dichas varianzas son iguales. Queremos contrastar la hipótesis $H_0 : \mu_1 = \mu_2$. Si las varianzas de ambas poblaciones no

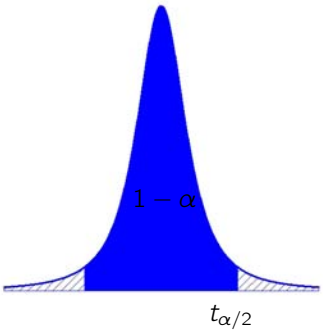

Figura 5: Función de densidad de una $T_{n_1+n_2-2}$. Región de aceptación y rechazo para el estadístico del contraste $H_0 : \mu_1 = \mu_2$ en el caso de dos poblaciones normales basados en muestras independientes con varianzas desconocidas pero iguales.

pueden suponerse iguales, entonces el estadístico de contraste que debemos utilizar es de la forma:

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_{c1}^2}{n_1} + \frac{S_{c2}^2}{n_2}}} \sim N(0,1)$$

siendo válida esta aproximación cuando las dos muestras son grandes (Criterio: $n_1 > 30$ y $n_2 > 30$). Por lo tanto, rechazaremos $H_0 : \mu_1 = \mu_2$ si

$$\frac{|\bar{X}_1 - \bar{X}_2|}{\sqrt{\frac{S_{c1}^2}{n_1} + \frac{S_{c2}^2}{n_2}}} > z_{\alpha/2}$$

donde de nuevo $z_{\alpha/2}$ es el valor que deja una probabilidad $\alpha/2$ a su derecha en la distribución N(0,1).

Observación: En la práctica, cuando las varianzas teóricas no se conocen, antes de contrastar la igualdad de medias, se deberá realizar un contraste de igualdad de varianzas. Si éste resulta significativo, entonces se resolverá el contraste de igualdad de medias considerando varianzas iguales y viceversa.

6.2 Contrastes referidos a dos poblaciones normales basados en muestras apareadas

Consideremos ahora el siguiente ejemplo:

Ejemplo 5: La siguiente tabla proporciona los resultados de la determinación de la concentración de paracetamol (%) en pastillas por dos métodos diferentes. Para ver si existen diferencias entre los resultados obtenidos por

los dos métodos se analizan diez pastillas de diez lotes diferentes,

Lote	Ensayo espectométrico UV	Espectroscopía de reflectancia en el infrarojo cercano
1	84.63	83.15
2	84.38	83.72
3	84.08	83.84
4	84.41	84.20
5	83.82	83.92
6	83.55	84.16
7	83.92	84.02
8	83.69	83.60
9	84.06	84.13
10	84.03	84.24

Nos gustaría contrastar si existe diferencia significativa entre los resultados obtenidos con los dos métodos.

La situación descrita en el Ejemplo 5 es ligeramente diferente a lo visto hasta este momento. En este caso se tienen dos muestras:

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

observadas en los mismos individuos $1, ..., n$, es decir, cada par de observaciones (X_{1i}, X_{2i}), se toma del mismo i-ésimo individuo, para $i = 1, ..., n$.
Se supone que las muestras se han obtenido de poblaciones normales, $X_1 \in N\left(\mu_1, \sigma_1^2\right)$ e $X_2 \in N\left(\mu_2, \sigma_2^2\right)$, pero teniendo en cuenta que ahora X_1 y X_2 no son independientes. A las muestras obtenidas de esta manera se les denomina **muestras apareadas**.

6.2.1 Contraste para comparar dos medias con muestras apareadas

Para abordar contrastes para comparar dos medias en esta situación se considera la variable, $D = X_1 - X_2$, que sigue una distribución normal (puesto que es una combinación lineal de variables normales).

- $\mu_D = \mathbb{E}(X_1 - X_2) = \mu_1 - \mu_2$
- $Var(D) = \sigma_D^2 = Var(X_1 - X_2)$

Tendremos entonces una muestra $D_1 = X_{11} - X_{21}, \ldots, D_n = X_{1n} - X_{2n}$. Estimaremos

- μ_D mediante $\bar{D}$
- σ_D^2 mediante S_{cD}^2

Queremos contrastar la hipótesis $H_0 : \mu_1 = \mu_2$. Fijamos el nivel de significación y escogemos el criterio que maximiza la potencia. Supongamos entonces que hemos determinado ya α. Si dicha hipótesis nula $H_0 : \mu_1 = \mu_2$ es cierta, entonces

$$\frac{\bar{D}}{S_{cD}/\sqrt{n}} \in T_{n-1}$$

El sentido común nos aconseja rechazar la hipótesis nula de que las medias son iguales cuando $\bar{D}$ sea muy distinta de cero. Como siempre, rechazaremos $H_0 : \mu_1 = \mu_2$ si

$$\frac{|\bar{D}|}{S_{cD}/\sqrt{n}} > t_{\alpha/2}$$

donde $t_{\alpha/2}$ es el valor que deja una probabilidad $\alpha/2$ a su derecha en la distribución T_{n-1}.

Estadística
Boletín 1: ESTADÍSTICA DESCRIPTIVA

Curso 2008/2009

Ej. 1 — Se midió el tiempo en segundos que tarda en consumirse cierto combustible, con los siguientes resultados:

$$11, 14, 7, 18, 12, 8, 10, 16, 12, 6$$

Obtén la tabla de frecuencias resultante de agrupar esos 10 datos en los intervalos $[5, 10)$, $[10, 15)$ y $[15, 20)$. Calcula la mediana con los 10 datos y con la tabla de frecuencias obtenida anteriormente.

Ej. 2 — ¿Bajo qué condiciones la varianza muestral vale cero?

Ej. 3 — Se han medido los pesos y las alturas de diez personas, obteniéndose los resultados siguientes (los pesos vienen expresados en kilogramos y las alturas en centímetros):

Peso:	68	65	75	84	81	62	88	70	72	78
Altura:	161	165	168	178	185	158	182	163	172	176

1. Calcular la media del peso y la altura.
2. ¿Cuál de las dos variables está más dispersa, el peso o la altura?
3. ¿Cómo se verían afectadas las contestaciones a los apartados anteriores si las alturas se midiesen en metros?

Ej. 4 — De un total de n números la fracción p son unos y la fracción $1 - p$ son ceros. Hallar los tres primeros momentos respecto a la media de estos números.

Ej. 5 — Las longitudes, medidas en metros, en una muestra de varillas de acero, han sido:

$$16, 5, 8, 19, 9, 6, 13, 9, 11, 14$$

Obtén la tabla de frecuencias resultante de agrupar estos datos en los intervalos $[4, 8)$, $[8, 12)$ y $[12, 20)$. Representa el histograma de frecuencias. Sobre la tabla de frecuencias, calcula la media, la desviación típica y el coeficiente de variación. ¿Qué valores habrían tomado estas tres medidas si la longitud hubiera sido medida en centímetros?

Ej. 6 — Se han medido mediante pruebas adecuadas los coeficientes intelectuales de un grupo de 20 alumnos, viniendo los resultados agrupados en seis intervalos de amplitud variable. Estas amplitudes son $C_1 = 12$, $C_2 = 12$, $C_3 = 4$, $C_4 = 4$, $C_5 = 12$, $C_6 = 20$. Si las frecuencias relativas acumuladas correspondientes a cada uno de los intervalos son: $F_1 = 0{,}15$, $F_2 = 0{,}15$, $F_3 = 0{,}55$, $F_4 = 0{,}8$, $F_5 = 0{,}95$, $F_6 = 1$. Se pide:

1. Formar la tabla de distribución de frecuencias (absolutas, relativas, absolutas acumuladas, relativas acumuladas), sabiendo que el extremo inferior del primer intervalo es 70.
2. Dibujar el histograma y el polígono de frecuencias. Calcular las medidas de posición central.
3. ¿Entre que dos percentiles está comprendido un coeficiente intelectual de 98.4? Encontrar el valor de ambos percentiles. ¿De qué tipo de medida se trata?

Ej. 7 — Se considera el número de materias suspensas que constan en el expediente de un grupo de alumnos. Con estos datos se construyó la siguiente distribución de frecuencias:

Materias suspensas	N° de estudiantes
0	13
1	16
2	25
3	16
4	9
5	5
6	4

1. Representar las frecuencias y sus acumuladas.
2. Calcular la moda, la media, la mediana, la desviación típica y el coeficiente de variación.

Estadística
Boletín 2: DESCRIPCIÓN ESTADÍSTICA DE DOS VARIABLES

Curso 2008/2009

Ej. 1 — Se ha medido el contenido de oxígeno (Y) en mg/litro de un lago a una profundidad de X metros, obteniéndose los siguientes datos:

Profundidad en m.	15	20	30	40	50	60	70
Contenido oxígeno en mg/l	6.5	5.6	5.4	6	4.6	1.4	0.1

Calcular la recta de regresión de Y sobre X y el coeficiente de correlación lineal.

Ej. 2 — La siguiente tabla contiene 6 lecturas del colorímetro (X) efectuadas sobre 6 disoluciones y sus 6 concentraciones de resina correspondientes (Y) determinadas por análisis químico (medidas en mg/100ml):

X	Y
8	0.12
50	0.71
81	1.09
102	1.38
140	1.95
181	2.50

1. Obtener la ecuación de la recta ajustada a la muestra bidimensional por el método de mínimos cuadrados.
2. Estudiar el grado de asociación lineal de la muestra anterior.
3. Supongamos que sobre una séptima disolución sabemos que la lectura del colorímetro fue 95, pero hemos extraviado su correspondiente medida de la concentración de resina. Haz una predicción de dicha concentración.

Ej. 3 — Se han examinado una serie de soluciones estándar de fluoresceína en un fluorímetro lo que condujo a las siguientes intensidades de fluorescencia (en unidades arbitrarias):

Intensidad de fluorescencia	2.1	5	9	12.6	17.3	21	24.7
Concentración en pg/ml	0	2	4	6	8	10	12

1. Queremos predecir la intensidad de fluorescencia a partir de la concentración. Obtener la recta mínimo cuadrática correspondiente.
2. Estudiar el grado de asociación lineal de la muestra anterior.
3. Obtener una predicción de la intensidad de fluorescencia para una solución cuya concentración es de 7 pg/ml. ¿Es fiable el resultado obtenido?.

Ej. 4 — De una variable estadística bidimensional se conoce que:

- La recta de regresión de Y sobre X es $Y = 2 + 0{,}5X$
- La recta de regresión de X sobre Y es $X = -4 + 2Y$
- $s_x = 3$

1. Halla la covarianza y la varianza de Y.
2. Si $\bar{x} = 2$, determina $\bar{y}$ y el momento respecto al origen de orden 2 de Y.

Ej. 5 — La siguiente tabla muestra la distribución conjunta de frecuencias relativas de la variable X, que representa el número de tarjetas de crédito que posee una persona, y la variable Y, que refleja el número de compras semanales pagadas con tarjeta de crédito.

X\Y	0	1	2	3	4
1	0.08	0.13	0.09	0.06	0.03
2	0.03	0.08	0.08	0.09	0.07
3	0.01	0.03	0.06	0.08	0.08

1. Si se sabe que en el estudio han participado 300 personas, hallar la distribución conjunta de frecuencias absolutas.
2. Hallar la distibución marginal de Y. ¿Cuál es el número medio y la desviación típica del número de compras semanales pagadas con tarjeta de crédito?
3. Obtener la distibución del número de tarjetas de crédito que poseen las personas de dicho estudio. ¿Cuál es el número más frecuente de tarjetas de crédito que posee una de estas personas?
4. Calcular la distribución del número de compras semanales pagadas con tarjetas de crédito que realizan las personas que poseen tres tarjetas. ¿Cuál es la media de esta distribución?

Ej. 6 — Se han estudiado el cociente intelectual de 100 niños (X) y sus calificaciones en Matemáticas (Y) obteniéndose los siguientes resultados:

$$\bar{x} = 110 \quad \bar{y} = 2{,}5 \quad s_x = 10 \quad s_y = 0{,}5.$$

Además se sabe que el coeficiente de correlación entre ambas variables es de 0.85.

1. ¿Qué nota se puede predecir para un niño con un cociente intelectual de 125?
2. ¿Cuánto vale los momentos de segundo orden respecto al origen de ambas variables?
3. ¿Cuál es la ecuación de la recta de regresión de X sobre Y?

Estadística
Boletín 3: PROBABILIDAD

Curso 2008/2009

Ej. 1 — Una moneda se lanza tres veces. Se pide:

1. Construir el espacio muestral.
2. Asignar probabilidades a los sucesos elementales.
3. Expresar a través de los sucesos elementales y calcular la probabilidad de los siguientes sucesos:
   A=Los tres lanzamientos producen el mismo resultado.
   B=El mismo resultado aparece dos veces exactamente.
   C=Al menos dos veces sale cara.
   D=Exactamente dos veces sale cara.
   E=Aparece cara en los lanzamientos primero y segundo.

Ej. 2 — Sean A, B y C sucesos tales que $P(A) = 0.5$, $P(B) = 0.3$, $P(C) = 0.6$, $P(A \cap B) = 0.2$, $P(A \cap C) = 0.3$, $P(B \cap C) = 0.1$, $P(A \cap B \cap C) = 0.01$.
Calcular: $P(A/B)$, $P(B/A)$, $P(A/C)$, $P(B^c/A)$, $P((A \cap B)/C)$ y $P((A \cup B)/C)$.

Ej. 3 — A un congreso asisten cien personas, de las cuales 60 hablan sólo inglés, 30 sólo francés y los 10 restantes ambos idiomas. Calcular la probabilidad de que se entiendan dos congresistas elegidos al azar.

Ej. 4 — Sean A, B y C sucesos arbitrarios de un experimento aleatorio. Se consideran los siguientes sucesos:
E_1= al menos dos de los sucesos A, B, C ocurren.
E_2= exactamente dos de los sucesos A, B, C ocurren.
E_3= al menos uno de los sucesos A, B, C ocurre.
E_4= exactamente uno de los sucesos A, B, C ocurre.
E_5= no más de dos sucesos A, B, C ocurren.

1. Expresar E_1, E_2, E_3, E_4 y E_5 en función de A, B y C.
2. Suponiendo que los sucesos A, B y C son independientes y sus probabilidades son 0.5, 0.2 y 0.3, respectivamente, calcula la probabilidad de E_5.

Ej. 5 — La probabilidad de que un chico acuda a una fiesta el sábado es 1/4, y la probabilidad de que su novia acuda a la fiesta es 1/3. Hallar la probabilidad de que:

1. Ambos acudan a la fiesta.
2. Al menos uno acuda a la fiesta.
3. Ninguno vaya a la fiesta.
4. Solamente la chica acuda a la fiesta.

Ej. 6 — En una urna hay dos bolas blancas y una negra. Dos individuos han de sacar sucesivamente y sin reemplazamiento una bola de la urna. ¿Cuál de los dos tiene mayor probabilidad de sacar la bola negra?

Ej. 7 — Tenemos dos urnas. La urna U_1 contiene 3 bolas blancas y 2 negras, y la urna U_2 contiene 1 bola blanca y 3 negras. Con probabilidad 1/3, extraeremos una bola al azar de la urna U_1 y con probabilidad 2/3 extraeremos una bola al azar de la urna U_2. Si al final nos comunican que la bola obtenida es blanca, ¿cuál es la probabilidad de que provenga de la urna U_1?, ¿y de que provenga de la urna U_2?

Ej. 8 — Se lanzan dos monedas y, a continuación, se lanza un dado tantas veces como caras se hayan obtenido. Hallar la probabilidad de que la suma de puntuaciones sea 6.

Ej. 9 — Una población está formada por tres grupos étnicos: A (30 %), B (10 %) y C (60 %). Los porcentajes del carácter "ojos claros" son, respectivamente, 20 %, 40 % y 5 %. Calcular:

1. La probabilidad de que un individuo elegido al azar tenga ojos claros.
2. La probabilidad de que un individuo de ojos oscuros sea de A.
3. Si un individuo, elegido al azar, tiene los ojos claros, ¿a qué grupo es más probable que pertenezca?

Ej. 10 — En un laboratorio se toman 200 medidas del contenido de mercurio en muestras de polvo utilizando un polarógrafo a varias distancias. Los resultados se clasifican por nivel de mercurio en cuatro clases (bajo, medio-bajo, medio-alto y alto) y por distancia al polarógrafo en tres clases (cerca, intermedio, lejos). El número de observaciones en cada grupo se clasifican en la tabla siguiente:

Nivel de H_g \ Distancia	Cerca	Intermedio	Lejos
Bajo	8	26	6
Medio-bajo	16	40	14
Medio-alto	6	62	12
Alto	0	2	8

Si se escoge una medida al azar, calcular:

1. Probabilidad de que tenga un nivel alto de mercurio.
2. Probabilidad de que haya sido medida lejos del polarógrafo.
3. Probabilidad de que se clasifique dentro del grupo "nivel bajo de mercurio - Cerca del polarógrafo"
4. Probabilidad de tener un nivel medio-bajo de mercurio condicionado a estar cerca del polarógrafo.
5. ¿Son los sucesos "nivel bajo de mercurio" y "distancia intermedia al polarógrafo" independientes?

Ej. 11 — Un ladrón en la plaza Roja, al huir de un policía, puede hacerlo por la calle Fray Rosendo Salvado, República del Salvador o San Pedro de Mezonzo, con probabilidades 0.25 , 0.6 y 0.15, respectivamente. La probabilidad de ser alcanzado si huye por la calle Fray Rosendo Salvado es 0.4 , si huye por la calle República del Salvador es 0.5 y si huye por la calle San Pedro de Mezonzo es 0.6.

1. Calcula la probabilidad de que la policía alcance al ladrón
2. Si el ladrón ha sido alcanzado, ¿cuál es la probabilidad de que haya sido en la calle Fray Rosendo Salvado?

Ej. 12 — De una urna que contiene 8 bolas blancas y 7 negras, hacemos una extracción de 2 bolas, sin reemplazamiento. En el supuesto de que hayamos visto que una de estas bolas es negra ¿Cuál es la probabilidad de que la otra también lo sea?.

Ej. 13 — Se lanzan dos dados, A y B. A es un dado corriente, mientras que B tiene en sus caras {1, 1, 1, 2, 2, 3}. Calcula la probabilidad de los siguientes sucesos,

1. La suma de los puntos obtenidos es 3.
2. En ambos dados se obtiene el mismo resultado.
3. Obtener un 1 con el dado A, sabiendo que el resultado de B ha sido distinto del obtenido en A.

Ej. 14 — Demostrar que si A y B son sucesos independientes, entonces también lo son A^c y B^c.

Estadística

Boletín 4: VARIABLES ALEATORIAS UNIDIMENSIONALES

Curso 2008/2009

Ej. 1 — Se lanza un dado. A continuación se lanza una moneda y si sale cara se suma uno a la puntuación del dado, y si sale cruz se deja igual la puntuación del dado. Sea X la variable aleatoria resultante de esa operación.

1. Determinar la distribución de probabilidad de X.
2. Calcular la media, la varianza y la desviación típica de X.
3. Determinar la probabilidad de que a lo sumo se obtenga un tres.

Ej. 2 — Sea X el tiempo de supervivencia de cierto tipo de resistencias. La función de densidad de X viene dada por:

$$f(x) = -\frac{x}{2} + 1 \quad \text{si } 0 \leq x \leq 2$$

1. Comprueba que f es una función de densidad y represéntala gráficamente.
2. Halla la probabilidad de supervivencia más allá de 1.
3. Determina la función de distribución correspondiente.

Ej. 3 — Una máquina fabrica discos cuyos radios se distribuyen con densidad

$$f(x) = k(x-1)(3-x) \quad \text{si } 1 \leq x \leq 3.$$

La variable X se mide en metros.

1. Hallar k.
2. Hallar la densidad para los radios de los discos medidos en centímetros.
3. Calcular la densidad para el diámetro de los discos.
4. Calcular la densidad para el área de los discos.

Ej. 4 — Se lanzan tres monedas. Sea X =“número de caras”. Se pide:

1. Distribución de probabilidad de X.
2. Función de distribución de X y su representación gráfica.
3. Media, varianza y desviación típica.
4. Probabilidad de que salgan a lo sumo dos caras.
5. Probabilidad de que salgan al menos dos caras.

Ej. 5 — Calcular k para que

$$f(x) = kx^2 \quad \text{si } 0 < x < 1$$

sea una función de densidad. Hallar:

1. Función de distribución.
2. Media y varianza.
3. El valor de a tal que $P(X \leq a) = 1/4$.

Ej. 6 — Para establecer el precio a pagar por cada litro de leche, una central lechera ha dividido, atendiendo al contenido de materia grasa por litro, la leche recibida en su factoría en tres categorías:

Categoría ligera: contenido de materia grasa inferior al 4 %

Categoría media: contenido de materia grasa entre el 4 % y el 5 %

Categoría extra: contenido de materia grasa superior al 5 %

El porcentaje de materia grasa por litro de leche recibido es una variable aleatoria con función de densidad:

$$f(x) = \begin{cases} \frac{2}{9}(6-x) & \text{si } x \in [3,6] \\ 0 & \text{si } x \notin [3,6] \end{cases}$$

Esta empresa paga el litro de leche a 30 pesetas para la categoría ligera, 35 pesetas para la categoría media y 40 pesetas para la categoría extra. Obténgase el precio medio del litro de leche pagado por esta empresa.

Ej. 7 — Sea X una variable aleatoria continua con función de densidad

$$f(x) = \begin{cases} k(1+x^2) & \text{si } x \in (0,3) \\ 0 & \text{si } x \notin (0,3) \end{cases}$$

1. Hallar la constante k.
2. Hallar la probabilidad de que X esté comprendida entre 1 y 2.
3. Hallar la probabilidad de que X sea menor que 1.

Estadística

Boletín 5: MODELOS DE DISTRIBUCIÓN DE PROBABILIDAD

Curso 2008/2009

Ej. 1 — En un examen entran 10 temas, se preguntan tres y para aprobar hay que contestar correctamente al menos dos. Un estudiante sabe 7 temas. ¿Qué probabilidad tiene de aprobar?

Ej. 2 — El tiempo de espera de un cliente hasta recibir el producto que ha solicitado sigue una distribución exponencial de media 40 días. Se pide:

1. La probabilidad de que tenga que esperar más de 40 días.
2. Lleva 40 días esperando, ¿cuál es la probabilidad de que llegue en los próximos 5 días? Calcula la misma probabilidad si llevase sólo 10 días esperando.

Ej. 3 — Una caja de cincuenta cerillas contiene diez defectuosas. Para inspeccionar la calidad de la caja, se toman siete cerillas de la misma.

1. ¿Cuál es la probabilidad de que no haya ninguna cerilla defectuosa entre las siete inspeccionadas?
2. Calcula la media y la varianza del número de cerillas defectuosas entre las siete inspeccionadas.

Ej. 4 — Una compañía de explotación petrolífera va a perforar 10 pozos, y cada uno de ellos tiene una probabilidad 0.1 de producir petróleo en forma comercial. A la compañía le cuesta 1 millón de euros perforar cada pozo. Un pozo comercial saca petróleo por valor de 50 millones de euros.

1. Calcular la media de la ganancia que obtendrá la compañía por los 10 pozos, así como su desviación típica.
2. Calcular la probabilidad de que la compañía pierda dinero con la operación.

Ej. 5 — Los errores en un aparato que transmite información, constituyen un proceso de Poisson con intensidad de 0.1 errores por minuto. ¿Cuál es la probabilidad de que en una hora haya como mucho un error?

Ej. 6 — El departamento de investigación de un fabricante de acero cree que una de las máquinas de rolado de la compañía está produciendo láminas de metal con espesores variables. El espesor es una variable aleatoria uniforme con valores entre 150 y 200 mm. Cualquier lámina que tenga menos de 160 mm. de espesor deberá desecharse, pues resulta inaceptable para los compradores.

1. Calcula la media y la desviación típica del espesor de las láminas producidas por esta máquina.
2. Calcula la función de densidad y represéntala.
3. Calcula la fracción de las láminas de acero producidas por esta máquina que se desechan.

Ej. 7 — Los empleados de cierto laboratorio tienen un horario oficial establecido de 7 horas y media al día, aunque trabajan entre 7 horas y 7 horas y 45 minutos al día dependiendo de diversos factores.

1. Calcula el tiempo que se puede esperar que trabaje al día un empleado.
2. ¿Cuál es la probabilidad de que un día cualquiera un empleado incumpla su horario?

3. En un departamento de ese laboratorio trabajan 5 empleados de forma independiente, ¿cuál es la probabilidad de que un día cualquiera sólo uno de esos empleados incumpla su horario?

Ej. 8 — Los errores en el peso proporcionado por la báscula de un laboratorio son normales de media 0 y desviación 1 kg. Calcula la probabilidad de que la diferencia entre el peso real de un material y el proporcionado por la báscula no supere los 500 gr. (bien por exceso o bien por defecto).

Ej. 9 — El consumo diario de carburante de cierta maquinaria sigue una distribución normal de media 7.31 y desviación 2.36 litros.

1. Calcula el porcentaje de días que el consumo supera los 9 litros.
2. ¿Cuántos litros consume como mínimo el 5 % de los días de mayor gasto?

Ej. 10 — Según recomendaciones de un estudio de salud laboral, no se deberían pasar más de 2 horas seguidas trabajando con el ordenador. Una gran empresa sabe que el tiempo máximo diario que están sus empleados trabajando con el ordenador sin realizar ninguna pausa es normal con una media de 3 horas y media y una desviación de 48 minutos, ¿qué porcentaje de empleados incumple esa recomendación?

Ej. 11 — En una universidad se ha observado que el 60 % de los estudiantes que se matriculan lo hacen en una carrera de Ciencias, mientras que el otro 40 % lo hacen en carreras de Humanidades. Si un determinado día se realizan 20 matrículas, calcular la probabilidad de que:

1. Haya igual número de matrículas en Ciencias y en Humanidades.
2. El número de matrículas en Ciencias sea menor que en Humanidades.
3. Haya al menos 8 matrículas en Ciencias.
4. No haya más de 12 matrículas en Ciencias.

Ej. 12 — Supongamos que la probabilidad de tener una unidad defectuosa en una línea de ensamblaje es de 0.05. Si el conjunto de unidades terminadas constituye un conjunto de ensayos independientes

1. ¿Cuál es la probabilidad de que entre diez unidades dos se encuentren defectuosas?
2. ¿Y de que a lo sumo dos se encuentren defectuosas?
3. ¿Cuál es la probabilidad de que por lo menos una se encuentre defectuosa?

Ej. 13 — Una empresa electrónica observa que el número de componentes que fallan antes de cumplir 100 horas de funcionamiento es una variable aleatoria de Poisson. Si el número promedio de estos fallos es ocho,

1. ¿Cuál es la probabilidad de que falle un componente en 25 horas?
2. ¿Y de que fallen no más de dos componentes en 50 horas?

Ej. 14 — Supóngase que X se distribuye como $N(\mu, \sigma^2)$, de manera que $P(X \leq 0) = 1/3$ y $P(X \leq 1) = 2/3$.

1. ¿Cuáles son los valores de μ y σ^2?
2. ¿Y si $P(X \leq 1) = 3/4$?

Estadística

Boletín 6: INFERENCIA ESTADÍSTICA: ESTIMACIÓN PUNTUAL, INTERVALOS DE CONFIANZA Y CONTRASTE DE HIPÓTESIS

Curso 2008/2009

Ej. 1 — En 20 días lectivos y a la misma hora se ha observado el número de terminales de una universidad conectados a Internet. Los resultados son:

1027, 1023, 1369, 950, 1436, 957, 634, 821, 882, 942,
904, 984, 1067, 570, 1063, 1307, 1212, 1045, 1047, 1178.

Se pide:

1. Calcular el intervalo de confianza al 95 % para el número medio de terminales conectados a Internet.
2. Calcular el intervalo de confianza al 90 % para la varianza del número de terminales conectados a Internet.

Ej. 2 — Una compañía asegura que sus tornillos miden por término medio tres centímetros. Se sabe que el proceso de producción sigue una distribución normal y padece una desviación típica de 0.1 cm. ¿Se puede dudar de la veracidad de ese tamaño medio si en una muestra de 25 tornillos la media fue de 3.5 cm?

Ej. 3 — Se lanza una moneda cien veces.

1. ¿Cuál es la probabilidad de que la proporción muestral de caras se encuentre entre 0.45 y 0.55?
2. ¿Cuántas veces habría que lanzar la moneda para que la proporción muestral de caras se encuentre entre 0.45 y 0.55 con una probabilidad de al menos el 95 %?

Ej. 4 — Una empresa desea conocer la proporción de clientes dispuestos a demandar el producto que ofrece. Para ello consultó, al azar, a cien de ellos, obteniendo los siguientes resultados: 30 estarían dispuestos a demandar y el resto no.

1. Obtener la estimación puntual de la proporción poblacional de demandantes.
2. Calcular la probabilidad de que la proporción muestral de demandantes difiera de la correspondiente proporción poblacional en menos de 0.15.

Ej. 5 — Se toma una muestra aleatoria de diez alumnos de una población escolar. Se considera, por experiencias anteriores, que la estatura de un alumno tiene distribución normal de media 167 cm y desviación típica 3.2 cm. Se pide:

1. Probabilidad de que la media muestral de las alturas de los diez alumnos sea inferior a 165 cm.
2. Probabilidad de que la cuasivarianza muestral de las alturas de los 10 alumnos sea superior a 16.50 cm^2.

Ej. 6 — Cierta empresa se ha propuesto comercializar un aparato para analizar la concentración en sangre de una sustancia. Los fabricantes son conocedores de que su método presenta un error de medición cuya desviación típica es de 2.4 mg/l. Sin embargo, dado que desconocen la media, se han decidido a tomar una muestra que les permita estimarla. A continuación consta tal muestra de los errores de medición (en mg/l):

0.51, -2.75, 1.83, 2.97, -0.82, 2.32, -0.69, -2.19,
1.47, -1.54, 0.30, -1.25, 0.18, -0.21, -1.95, -3.67.

Elabora una estimación de la media y construye un intervalo de confianza a un nivel del 99 % para dicha estimación, suponiendo que los errores siguen una distribución normal.

Ej. 7 — Los siguientes datos representan los tiempos (en minutos) de montaje para 20 unidades seleccionadas aleatoriamente:

9.8	10.4	10.6	9.6	9.7	9.9	10.9	11.1	9.6	10.2
10.3	9.6	9.9	11.2	10.6	9.8	10.5	10.1	10.5	9.7

Supóngase que el tiempo necesario para montar una unidad es una variable aleatoria normal. A partir de esta muestra, ¿existe alguna razón para creer, a un nivel de 0.05, que la media del tiempo de montaje es mayor de 10 minutos?

Ej. 8 — La cantidad de horas que duermen los escolares cada noche varía mucho. Consideremos la siguiente muestra de las horas que duermen cada noche 16 alumnos de un instituto.

6.9, 7.6, 6.5, 6.2, 7.8, 7.0, 5.5, 7.6,
7.3, 6.6, 7.1, 6.9, 6.8, 6.5, 7.2, 5.8

1. Calcula una estimación puntual para la media de horas que se duerme cada noche y para la desviación típica. ¿Qué estimadores utilizas? ¿Por qué?
2. Suponer que la población sigue una distribución normal.
   a) Determinar un intervalo de confianza del 80 % para la media de horas que se duerme cada noche.
   b) Determinar un intervalo de confianza del 90 % para la varianza.

Ej. 9 — Se pretende conocer la media y la varianza del tiempo de eliminación de un medicamento. Para ello, se han observado los tiempos en una muestra de pacientes, obteniéndose los siguientes datos (en horas):

5.64, 7.83, 6.92, 5.31, 8.85, 7.94, 6.04, 5.19,
7.33, 8.24, 7.68, 6.47, 6.09, 8.75, 5.87, 7.28.

Supón que los datos proceden de una distribución normal y, en base a ello, confecciona estimaciones para la media y la varianza, así como intervalos de confianza a un nivel del 90 % para las mismas.

Ej. 10 — En una región han registrado las profundidades que tuvieron que alcanzar los pozos hasta obtener agua (en metros): 21, 19, 29, 30, 28, 22, 26, 25, 28, 22. Proporciona una estimación de la media. Suponiendo que la profundidad tiene distribución normal, construye un intervalo de confianza a un nivel del 95 % para esa estimación de dos formas: sabiendo que la desviación típica es de 4 metros y sin conocer la desviación típica.

Ej. 11 — El responsable del control de calidad de una factoría está interesado en determinar si la distribución de la tensión de ruptura (en K/cm^2) de cierto metal cumple los requisitos para ser empleado en la construcción de buques. Estos requisitos se traducen en que la media de la tensión debe ser de 454 K/cm^2, con una desviación de 9 K/cm^2. Para ver si se cumplen dichos requisitos, se seleccionan de forma aleatoria 21 muestras del metal sometiéndolas a presión hasta su ruptura. Las tensiones obtenidas dan lugar a una media muestral de 443.81 K/cm^2 y a una cuasidesviación típica muestral igual a 9.4 K/cm^2. Suponiendo normalidad,

1. Realizar los test de hipótesis pertinentes ($\alpha = 0.1$) para verificar si se cumplen los requisitos.
2. Determinar un intervalo de confianza de nivel 0.90 para la tensión media y otro para su desviación típica. Explicar cómo se podrían haber resuelto los contrastes del apartado anterior a partir de estos intervalos.

Ej. 12 — Se cree que los jóvenes adolescentes que fuman comienzan a hacerlo a una edad más temprana que las chicas adolescentes fumadoras. ¿Los siguientes datos apoyan esta suposición?. (Suponer que la distribución de la variable edad a la que empiezan a fumar hombres y mujeres, es normal).

Hombres	Mujeres
$n = 31$	$m = 13$
$\bar{x} = 11.3$ años	$\bar{y} = 12.6$ años
$s_x^2 = 4$ años^2	$s_y^2 = 3.5$ años^2

Ej. 13 — Para estudiar el efecto del ejercicio físico sobre el nivel de triglicérido, se ha realizado el siguiente experimento con 11 individuos: previo al ejercicio, se tomaron muestras de sangre para determinar el nivel de triglicérido por 100 mililitros de sangre, de cada sujeto. Después los individuos fueron sometidos a un programa de ejercicios que se centraba diariamente en carreras y marchas. Al final del periodo de ejercicios, se tomaron nuevamente muestras de sangre y se obtuvo una segunda lectura del nivel de triglicérido. De este modo, se dispone de dos conjuntos de observaciones del nivel de triglicérido por 100 mililitros de sangre de los sujetos: (suponer normalidad),

Sujeto	1	2	3	4	5	6	7	8	9	10	11
Previo	68	77	94	73	37	131	77	24	99	629	116
Posterior	95	90	86	58	47	121	136	65	131	630	104

¿Hay pruebas suficientes para afirmar que el ejercicio físico produce cambios en el nivel de triglicérido?

Ej. 14 — Hallar un intervalo de confianza del 99 % para μ, número medio de microgramos de partículas en suspensión por metro cúbico de aire, en base a los valores de una muestra aleatoria simple de tamaño $n = 5$, dada por $\{58, 70, 57, 61, 59\}$, en los siguientes casos:

1. X, número de microgramos de partículas en suspensión por metro cúbico de aire, está normalmente distribuida con varianza 9.
2. X, número de microgramos de partículas en suspensión por metro cúbico de aire, está normalmente distribuida con varianza desconocida.

Ej. 15 — Una empresa de metalurgia está interesada en la temperatura media que alcanza cierta máquina utilizada en el proceso de fabricación. Para su estimación se obtienen 10 mediciones en grados centígrados: 41.60, 41.84, 42.34, 41.95, 41.86, 42.18, 41.72, 42.26, 41.81, 42.04.

1. Obtener el intervalo de confianza al 95 % para la temperatura media supuesto que $\sigma = 0.30$ grados.
2. Deducir el tamaño muestral necesario para conseguir un intervalo de confianza al 95 % con una longitud menor o igual que 0.1 grados.
3. Determinar el intervalo de confianza al 95 % para la temperatura media supuesto que desconocemos el valor de σ.

Estadística
Práctica 1: INTRODUCCIÓN A MATLAB

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Objetivos de la práctica

El objetivo de esta práctica es familiarizarse con el funcionamiento y terminología básicas en MATLAB. Repasaremos:

- El entorno de desarrollo. descripción y localización de las herramientas más habituales.
- Fundamentos de MATLAB: estructuras de datos y operaciones básicas.
- Representaciones gráficas.

2 Introducción

MATLAB es un lenguaje de computación técnica de alto nivel y un entorno interactivo para desarrollo de algoritmos, visualización de datos, análisis de datos y cálculo numérico. MATLAB cuenta con una amplia gama de aplicaciones que incluyen procesamiento de señales e imágenes, comunicaciones, diseño de sistemas de control, sistemas de prueba y medición, modelado y análisis financiero y biología computacional. Los conjuntos de herramientas complementarios (colecciones de funciones de MATLAB para propósitos especiales, que están disponibles por separado) amplían el entorno de MATLAB permitiendo resolver problemas especiales en estas áreas de aplicación.
Además, MATLAB contiene una serie de funciones para documentar y compartir el trabajo. Se puede integrar código de MATLAB con otros lenguajes y aplicaciones, y distribuir los algoritmos y aplicaciones que desarrollo usando MATLAB.

Características principales:

- Lenguaje de alto nivel para cálculo técnico
- Entorno de desarrollo para la gestión de código, archivos y datos
- Herramientas interactivas para exploración, diseño y resolución de problemas iterativos
- Funciones matemáticas para álgebra lineal, estadística, análisis de Fourier, filtraje, optimización e integración numérica
- Funciones gráficas bidimensionales y tridimensionales para visualización de datos
- Herramientas para crear interfaces gráficas de usuario personalizadas
- Funciones para integrar los algoritmos basados en MATLAB con aplicaciones y lenguajes externos, tales como C/C++, FORTRAN, Java, COM y Microsoft Excel.

3 El escritorio de MATLAB

En general, cuando se inicia MATLAB, aparece el escritorio de trabajo junto con una serie de herramientas que nos permiten manejar los ficheros, variables y aplicaciones asociadas a MATLAB. En la Figura 2 aparece un ejemplo del escritorio de MATLAB. Se puede cambiar la visualización del escritorio según las necesidades abriendo o cerrando nuevas ventanas, redistribuyendo las herramientas, etc. Consulta la ayuda del programa para adaptar el escritorio a tus necesidades.

El paquete de ayuda de MATLAB: Para acceder al paquete de ayuda de MATLAB pulsa en el menú superior el botón **Help**. Una vez desplegada la ayuda selecciona en la pestaña de Contenidos el tema que deseas consultar. Además, si tienes dudas sobre los argumentos de una función o sobre su funcionamiento puedes utilizar el comando `help` en la consola de comandos. Por ejemplo, escribe `help mean`. ¿Qué hace la función `mean`? ¿Qué argumentos toma?

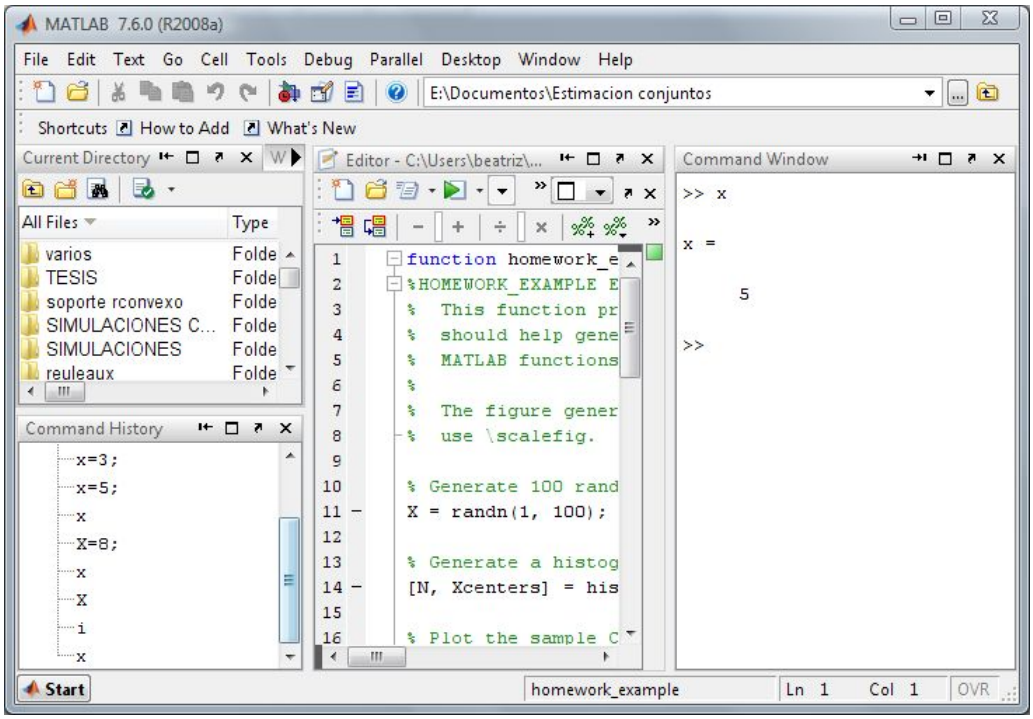

Figura 1: *Vista del escritorio de MATLAB con ventana de comandos, navegador de ficheros, historia de comandos, editor, etc.*

4 Fundamentos de MATLAB

4.1. Operaciones elementales

La notación para las operaciones matemáticas elementales es la siguiente:

$\wedge$	exponenciación
$*$	multiplicación
$/$	división
$+$	suma
$-$	resta

Veamos un ejemplo muy sencillo:

```
>> x=3+5
x =
      8
```

Si no se asigna el resultado a ninguna variable, MATLAB lo asigna por defecto a la variable `ans` (answer):

```
>> 3-8
ans =
    -5
```

En ocasiones, es interesante no presentar el resultado en la pantalla (por ejemplo, cuando se trata de una lista de datos muy larga). Eso se consigue poniendo un punto y coma al final de la instrucción.

```
>> y=3+5;
>>
```

Veamos ahora un listado con las funciones elementales:

`sin`	seno
`cos`	coseno
`tan`	tangente
`exp`	exponencial
`log`	logartimo natural
`sqrt`	raíz cuadrada
`abs`	valor absoluto

Edición de la línea de comandos: Con las flechas del teclado se pueden recuperar las órdenes anteriores, sin tener que volver a teclearlas. Así, en el caso de una equivocación en un comando complicado en vez de volver a teclear todo, puede recuperarse la instrucción pulsando la tecla "flecha hacia arriba", desplazarse hasta el error y arreglarlo.

4.2. Estructuras de datos: vectores y matrices

Un vector se define introduciendo los componentes, separados por espacios o por comas, entre corchetes:

```
» v=[sqrt(3) 0 -2]
v =
      1.7321    0   -2.0000
```

Para definir un vector columna, se separan las filas por puntos y comas:

```
» w=[1;0;1/3]
w =
      1.0000
      0
      0.3333
```

La operación transponer (cambiar filas por columnas) se designa por el apóstrofe:

```
» w'
ans =
      1.0000    0    0.3333
```

Las operaciones matemáticas elementales pueden aplicarse a los vectores:

```
» v*w
ans =
      1.0654
» v+w'
ans =
      2.7321    0   -1.6667
```

Para crear un vector de componentes equiespaciados se emplean los dos puntos:

```
» x=4:2:10
x =
      4  6  8  10
```

Utiliza esta función para construir el vector $v = (8, 6, 4, 2)$.

> Vectores de ceros y unos: Con la función `zeros` se puede crear un vector en el que todas las componentes sean ceros. La función `ones` sirve para crear vectores en los que todas las componentes sean unos. Consulta la ayuda y utiliza dichas funciones para crear vectores de ceros y unos. ¿Cómo aprovecharías dichas funciones para crear el vector $v = (5, 5, 5, 5, 5)$? Consulta la ayuda de la función `linspace` para obtener el mismo resultado.

Para introducir matrices, se separa cada fila con un punto y coma:

```
» M = [1 2 3 ;4 5 6 ;7 8 9]
M =
      1  2  3
      4  5  6
      7  8  9
```

Para referirse a un elemento de la matriz se hace así:

```
» M(3,1)
ans =
      7
```

Para referirse a toda una fila o a toda una columna se emplean los dos puntos. Por ejemplo, la segunda columna de la matriz se obtiene así:

```
» v1=M(:,2)
v1 =
      2
      5
      8
```

Para obtener la primera fila haremos

```
>> M(1,:)
ans =
     1     2     3
```

Con las matrices también funcionan las operaciones matemáticas elementales. Así

```
» M^2
ans =
      30  36  42
      66  81  96
      102 126 150
```

Si se quiere operar en los elementos de la matriz, uno por uno, se pone un punto antes del operador. Si se quiere elevar al cuadrado cada uno de los elementos de M, entonces

```
» M.^2
ans =
      1  4  9
      16 25 36
      49 64 81
```

> Longitud de un vector y dimensión de una matriz: Consulta la ayuda de las funciones `length`, `size` y `numel` y aplícalas a diferentes vectores y matrices. ¿Qué calcula cada una de ellas?

Define las matrices

$$A = \begin{pmatrix} 2 & 6 & 4 \\ 1 & 5 & 8 \end{pmatrix} \quad B = \begin{pmatrix} 4 & 1 & 3 \\ 2 & 2 & 1 \end{pmatrix} \quad C = \begin{pmatrix} 3 & 6 \\ 4 & 7 \end{pmatrix}$$

¿Qué estás calculando si ejecutas en MATLAB `A*B`, `A*C`, `A.*B`, `A.*C`? ¿Por qué algunas de estas operaciones no tienen sentido?

4.3. Operadores relacionales y lógicos

Los principales operadores relacionales aparecen recogidos en la siguiente tabla.

==	igual
~=	distinto
>	mayor
>=	mayor o igual
<	menor
<=	menor o igual

Estos operadores se pueden combinar utilizando los operadores lógicos:

&	y
|	o
~	no

Veamos algunos ejemplos de como usar estos operadores.

```
>> v=[4 5 7 2 1 6]
v =
     4     5     7     2     1     6
>> v==4
ans =
     1     0     0     0     0     0
>> v~=4
```

```
ans =
     0     1     1     1     1     1
>> ~(v==4)
ans =
     0     1     1     1     1     1
>> v>4
ans =
     0     1     1     0     0     1
>> v>=4
ans =
     1     1     1     0     0     1
>> v>2&v<6
ans =
     1     1     0     0     0     0
>> w=[4 6 7 1 5 7]
w =
     4     6     7     1     5     7
>> v==w
ans =
     1     0     1     0     0     0
>> x=[2 4]
x =
     2     4
>> v==x
??? Error using ==> eq
Matrix dimensions must agree.
>> v(v>4)
ans =
     5     7     6
```

5 Representaciones gráficas con MATLAB

Las posibilidades que ofrece MATLAB para hacer representaciones gráficas son muy grandes. Veremos a continuación cómo realizar gráficos sencillos.

> Representaciones gráficas con MATLAB: Para más información sobre todas las posibilidades que ofrece MATLAB a la hora de hacer representaciones gráficas consulta en la ayuda el capítulo dedicado a este tema. También puedes consultar la ayuda de la función `graph2d`.

Veamos cómo se puede representar la función seno entre 0 y 10. Para empezar creamos una variable x que va de cero a 10.

```
» x=0:0.1:10;
```

y a continuación, calculamos $\sin(x)$ almacenando el resultado en la variable y:

```
» y=sin(x);
```

Para trazar el gráfico, se emplea la función `plot`:

```
» plot(x,y)
```

y se obtiene en otra ventana el gráfico.

Otras funciones para cambiar el aspecto del gráfico: Consulta la ayuda de la función `plot` y utiliza las funciones a las que se hace referencia en dicha ayuda para cambiar el aspecto de tu gráfico, añadiendo un título, nombre a los ejes, cambiando el aspecto de los puntos, etc.

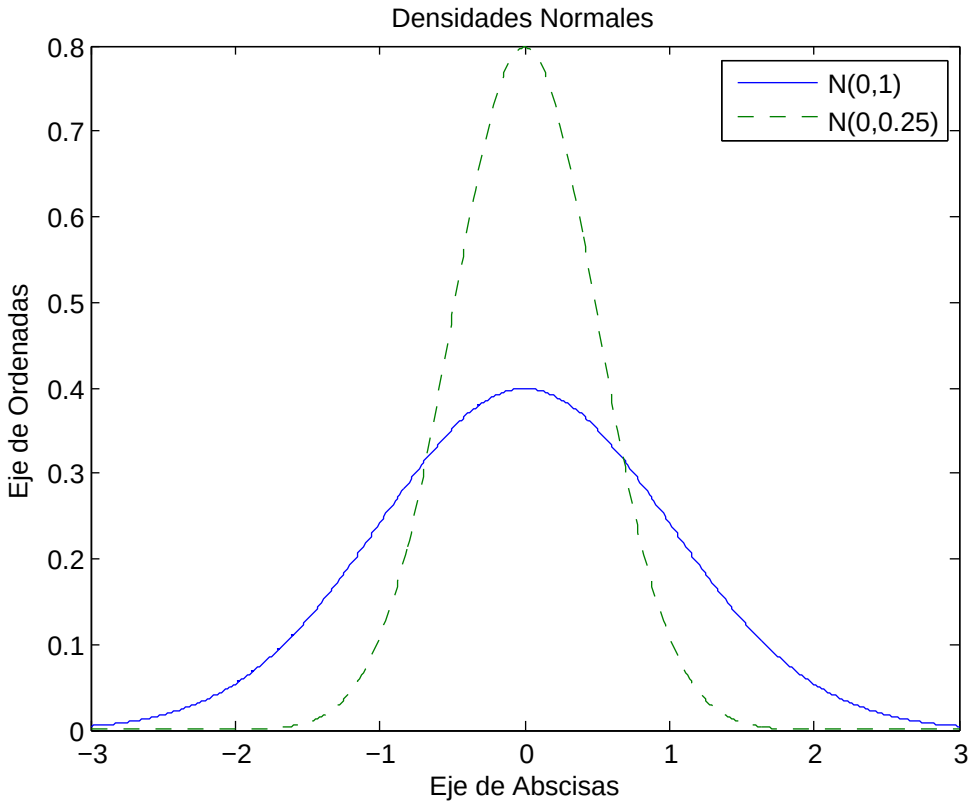

Figura 2: *Intenta reproducir este gráfico. En azul se representa la función* $f(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$. *En verde se representa la función* $f(x) = \frac{1}{\sqrt{0{,}5\pi}} e^{\frac{-x^2}{0{,}5}}$.

Estadística
Práctica 2: ESTADÍSTICA DESCRIPTIVA CON MATLAB

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Objetivos de la práctica

El principal objetivo de esta práctica es conocer los procedimientos de estadística descriptiva que nos ofrece MATLAB y aplicarlos a un conjunto de datos. Repasaremos:

- Principales funciones para importar datos.
- Tablas de frecuencias para variables cualitativas y cuantitativas (discretas o continuas).
- Representaciones gráficas.
- Medidas características

2 Importando datos

La forma más sencilla de importar datos desde MATLAB es mediante el menú de importar datos (File->Import Data). Utilizando dicho menú podemos leer automáticamente datos numéricos almacenados en cualquier fichero de texto. Como ejemplo, el fichero **altura.txt** contiene las alturas de los alumnos de Ingeniería Química del curso 2008/2009. Guarda dichos datos en una variable llamada `altura` .
Los datos recogidos en clase (sexo, altura, peso, número de hermanos y equipo de los alumnos de Ingeniería Química del curso 2008/2009) se encuentran guardados en el archivo **IQ0809.csv**.

> Ficheros CSV: Los ficheros CSV (del inglés comma-separated values) son un tipo de documento sencillo para representar datos en forma de tabla, en los que las columnas se separan por un carácter delimitador (coma, punto y coma,...) y las filas se separan por saltos de línea.

Abre el fichero **IQ0809.csv** con un editor de texto y comprueba su estructura. Si intentas importar los datos de **IQ0809.csv** en MATLAB a través del menú de importar datos, verás que no es posible. Si el fichero de datos que queremos importar contiene una mezcla de datos numéricos y alfanuméricos, la opción más simple para importar dichos datos es la función `textscan`. Previamente tendrás que abrir el fichero **IQ0809.csv** mediante la función `fopen` para acceder a su lectura.

```
» fid=fopen('IQ0809.csv')
```

Ahora, para leer el contenido del fichero **IQ0809.csv**, utiliza la función `textscan`. Debes especificar el identificador del fichero `fid` y los formatos de las columnas de datos (`%s` para variables carácter, `%f` para formato double, `%d` para formato entero, ...). Además si se utiliza un carácter distinto del espacio en blanco como delimitador de columnas, utiliza el parámetro `'delimiter'` para especificar dicho delimitador. Por último, usando el parámetro `'headerlines'` podemos especificar el número de líneas de cabecera que debemos ignorar. Completa el siguiente comando con los argumentos necesarios para importar los datos de **IQ0809.csv**.

```
» IQ=textscan(fid, ...
```

> Cell array: El resultado de importar los datos del fichero **IQ0809.csv** es un objeto de tipo **cell**. Para acceder a los elementos de dicho objeto utiliza { }.

Una vez importado el fichero, guarda en `sexo, altura, peso, nher` y `equipo` los datos de las variables. Guarda en `nind` y `nvar` el número de individuos (tamaño muestral) y el número de variables recogidas.

3 Tablas de frecuencias y gráficas para variables cualitativas

Veamos ahora diversas maneras de hacer estadística descriptiva con este conjunto de datos: (1) mediante tablas de frecuencias, (2) mediante gráficos, (3) mediante el uso de medidas de centralización, dispersión y forma. Empezamos con las variables cualitativas. Vemos en primer lugar como obtener las frecuencias absolutas de la variable `equipo`, que es una variable cualitativa nominal. Para obtener las frecuencias absolutas utilizamos la función `tabulate`, como sigue:

```
>> equipo=nominal(equipo);
>> neq=getlabels(equipo);
>> tabulate(equipo);
      Value    Count   Percent
   Atletico        2      2.82%
  Barcelona       24     33.80%
      Betis        1      1.41%
      Celta        5      7.04%
      Depor       20     28.17%
     Madrid       15     21.13%
   Numancia        1      1.41%
   Sporting        1      1.41%
   Valencia        1      1.41%
         nc        1      1.41%
```

Comenta los resultados. ¿Qué representan las columnas `Count` y `Percent`? ¿Cómo calcularías las frecuencias relativas a partir de dichas columnas? Comprueba que se verifican las propiedades de frecuencias relativas y absolutas.

> Suma de los elementos de un vector: El comando `sum` permite obtener la suma a lo largo de un vector columna o un vector fila.

Una vez calculadas las frecuencias absolutas y relativas de cada uno de los equipos de la muestra, podemos empezar a hacer resúmenes gráficos. Por ejemplo, para hacer un diagrama de barras utilizaremos la función `bar`.

> Representaciones gráficas de variables cualitativas: La función `bar(x,y)` realiza un diagrama de barras. Podemos influir en la anchura de los rectángulos representados. La función `stem(x,y)` representa segmentos verticales rematados con una circunferencia en su extremo superior, en las posiciones del eje de abscisas contenidas en el vector **x** y cuyas alturas vienen dadas por el vector **y**.

Por ejemplo, si guardamos en el vector `frel` las frecuencias relativas de los equipos de fútbol de la muestra,

```
>> bar(frel)
>> set(gca,'XTickLabel',neq)
```

representará el diagrama de barras que aparece en la Figura 1. El último comando nos permite situar en el eje de abcisas los nombres de los equipos.
Ejercicio: Realiza un estudio similar con el resto de variables cualitativas recogidas.

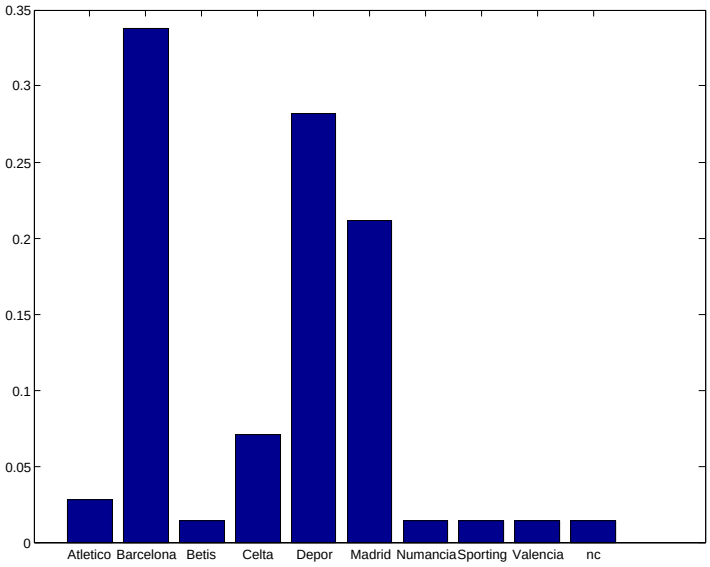

Figura 1: *La función* `bar` *representa rectángulos.*

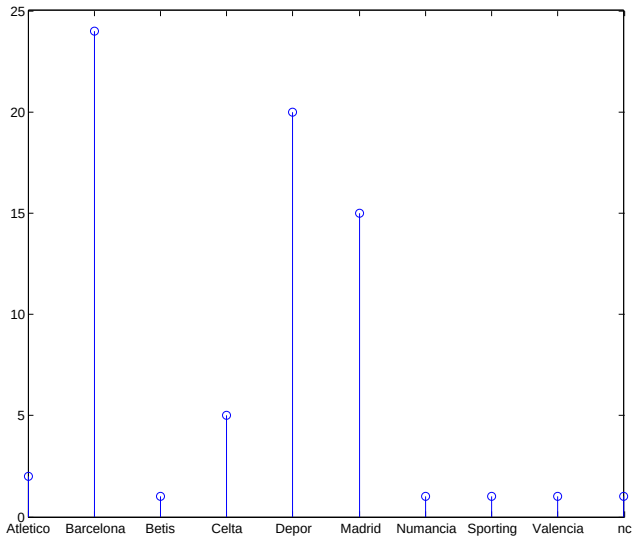

Figura 2: *La función* `stem` *representa segmentos verticales.*

4 Tablas de frecuencias y gráficas para variables cuantitativas discretas

Consideremos ahora la variable "Número de hermanos", que es una variable cuantitativa discreta. Podemos volver a utilizar la función `tabulate` para obtener una tabla de frecuencias absolutas y porcentajes. Fíjate que ahora, el resultado de la función `tabulate` es una matriz.

```
>> tabulate(nher)
```

```
  Value     Count    Percent
      0        13     18.31%
      1        45     63.38%
      2        12     16.90%
      3         1      1.41%
```

Además ahora tiene sentido calcular las frecuencias acumuladas (tanto absolutas como relativas). Utiliza la función `cumsum` para calcularlas. Comprueba que se verifican las propiedades de frecuencias relativas y absolutas acumuladas. Utiliza la función `bar` para realizar las representaciones gráficas correspondientes, entre ellas un diagrama de frecuencias acumuladas como el que se muestra en la Figura 3.

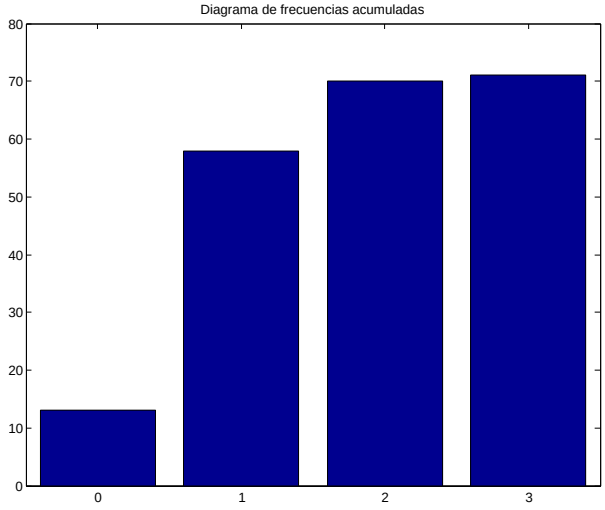

Figura 3: *Diagrama de frecuencias acumuladas para la variable "Número de hermanos".*

5 Tablas de frecuencias y gráficas para variables cuantitativas continuas

Para variables cuantitativas continuas las cosas se complican algo debido a que tenemos que agrupar los valores de las variables. Consideremos como ejemplo la variable "Altura". Existen varias opciones para definir los intervalos en los cuales agruparemos los datos. Fíjate en el siguiente código. ¿Qué es lo que hace?

```
>> ninter=round(sqrt(nind));
>> aumento=range(altura)*0.15;
>> extremos=linspace(min(altura)-aumento,max(altura)+aumento,ninter+1)
```

Una vez determinados los intervalos que tendremos en cuenta, ¿cómo determinarías el intervalo al que pertenece cada observación?

```
>> intervalo=zeros(nind,1);
>> for i=1:ninter
intervalo(extremos(i)<=altura&altura<extremos(i+1))=i;
end
```

```
>> tabulate(intervalo)
  Value    Count   Percent
      1        2      2.82%
      2       12     16.90%
      3       16     22.54%
      4       13     18.31%
      5       17     23.94%
      6        7      9.86%
      7        3      4.23%
      8        1      1.41%
```

La función histc: MATLAB ofrece el comando `histc` para construir la agrupación en intervalos de clase y calcular las frecuencias. Se utiliza así

- **ni = histc(y,ext)**: para el vector **y** cuenta el número de valores de **y** que caen entre los elementos del vector **ext**.

El histograma: MATLAB ofrece el comando `hist` para construir la agrupación en intervalos de clase, calcular las frecuencias y representarlas mediante un histograma. Presenta las siguientes posibilidades:

- **ni = hist(y)**: Reparte la muestra contenida en el vector **y** en diez intervalos de igual longitud y devuelve en el vector **ni** las frecuencias absolutas de cada intervalo.
- **ni = hist(y,m)**: Utiliza **m** intervalos.
- **ni = hist(y,x)**: Utiliza las marcas de clase especificadas por el vector **x**.
- **[ni, x] = hist(y)**: Devuelve en **x** las marcas de clase.
- **hist(y)**: Sin argumentos de salida, produce la figura con el histograma. **Ojo!** las alturas que representa son frecuencias absolutas

Ejercicio: Programa una función `histograma` que represente el histograma de una variable cuantitativa continua de forma que las alturas de los rectángulos representen las densidades de frecuencias.

6 Medidas características

6.1. Medidas de posición

MATLAB ofrece comandos que permiten calcular directamente algunas medidas de posición. Para el cálculo de la media podemos usar la función `mean`, que efectúa la media aritmética simple de los elementos de un vector, esto es, los suma y divide entre el número de ellos. De este modo nos permite calcular la media de una muestra de una variable continua. ¿Cuál es la altura media de los alumnos de Ingeniería Química? ¿Y el peso medio?

Datos agrupados: Si deseamos ponderar por las frecuencias, en lugar de la función `mean`, debemos efectuar el producto escalar del vector de valores distintos por el vector de frecuencias.

Respecto a la mediana, MATLAB ofrece la función `median`, que permite calcular la mediana de un vector de observaciones como el valor central o la media de los dos centrales (según proceda) en las observaciones ordenadas. Al igual que la función `mean`, no tiene en cuenta frecuencias y por tanto es aplicable sólo cuando disponemos de todos los datos de una muestra de una variable continua.

Para calcular la moda, MATLAB no ofrece ninguna función. Sin embargo, podemos obtener la moda solicitando el máximo de las frecuencias, mediante la función `max`, que devuelve el máximo valor a lo largo de un vector. Por ejemplo, calcula la moda para la variable "Número de hermanos".

MATLAB ofrece la posibilidad de calcular los percentiles de un vector de valores mediante la función `prctile`. La sintaxis sería:

```
>> prctile(x,p)
```

y devuelve un valor que sería mayor que el $p\,\%$ de los valores del vector x. Por ejemplo, si $p = 50$, estamos pidiendo la mediana de x.
Utiliza la función `prctile` para calcular los cuartiles de las variable "Altura" y "Peso".

6.2. Medidas de dispersión

Programa una función que calcule la varianza de un vector de datos y úsala para calcular la varianza de las variables "Peso" y "Altura". Calcula también la desviación típica de dichas variables.

La función std: El lenguaje MATLAB ofrece la funcion `std` para el cálculo de la desviación tipica de un vector de datos. Igual que en casos anteriores, no tiene en cuenta frecuencias y maneja como antes las matrices y vectores. Compara el resultado obtenido con la función que has programado y con la función `std`. Ambos resultados no coinciden porque la función `std` devuelve

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

El origen de esta medida (conocida como cuasidesviacion tipica y su cuadrado como cuasivarianza) se verá justificado en el Tema 7 de Inferencia Estadistica.

6.3. Medidas de forma

Programa una función que te permita calcular los momentos centrales de cualquier orden y utiliza dicha función para obtener lo coeficientes de asimetría y kurtosis de las variables "Altura" y "Peso".

6.4. El diagrama de caja o boxplot

La información obtenida a partir de las medidas de centralización, dispersión y forma se puede usar para realizar **diagramas de caja (boxplots)** que visualmente nos proporcionen la información de cómo están distribuidos los datos. El diagrama de caja consta de una caja central que está delimitada por la posición de los cuartiles Q_3 y Q_1. Dentro de esa caja se dibuja la línea que representa la mediana. También ocasionalmente se puede representar la media dentro de la caja. De los extremos de la caja salen unas líneas que se extienden hasta los puntos $LI = \text{máx}\{\text{mín}(x_i), Q_1 - 1{,}5\,(RI)\}$ y $LS = \text{mín}\{\text{máx}(x_i), Q_3 + 1{,}5\,(RI)\}$ que representarían el rango razonable hasta el cual se pueden encontrar datos. Los datos que caen fuera del intervalo (LI, LS) se consideran

datos atípicos y se representan individualmente. La función para obtener esta representación en MATLAB es `boxplot`.

```
>> boxplot(altura)
```

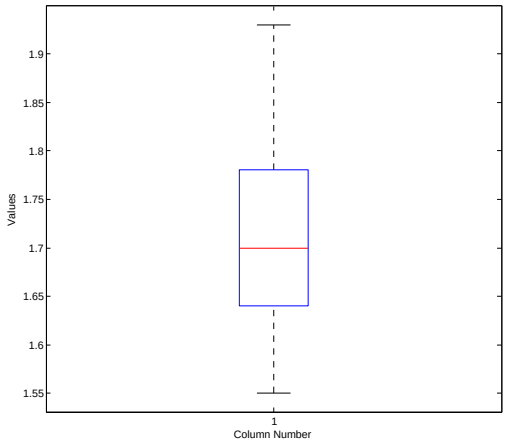

Figura 4: *Diagrama de caja para la variable "Altura".*

7 Ejercicios

- Calcula la altura media y peso medio de los chicos y de las chicas de la clase.
- ¿Cuál es el equipo favorito de los chicos? ¿Y el de las chicas?
- Obtén un diagrama de caja como el que se muestra en la Figura 5 e intrepreta el resultado.

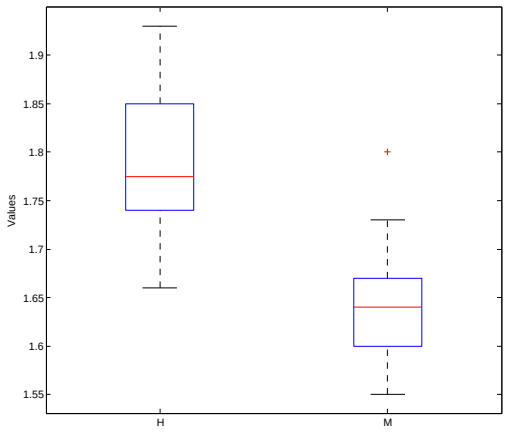

Figura 5: *Diagrama de caja para la variable "Altura" agrupado por Sexo.*

Estadística
Práctica 3: DESCRIPCIÓN ESTADÍSTICA DE DOS VARIABLES

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Objetivos de la práctica

El objetivo de esta práctica es aprender a utilizar MATLAB como herramienta para la estimación y discusión de modelos de regresión. Repasaremos:

- Gráficos de dispersión.
- Funciones para ajustar un modelo de regresión lineal simple.
- Covarianza, coeficiente de correlación lineal, coeficiente de determinación.

2 Gráficos de dispersión

La situación general es la siguiente. Disponemos de una variable aleatoria respuesta Y, que supondremos relacionada con otra variable X, que llamaremos explicativa o independiente. A partir de una muestra de n individuos para los que se dispone de los valores de ambas variables, $\{(X_i, Y_i), i = 1, \ldots, n\}$, podemos visualizar gráficamente la relación existente entre ambas. Así, utilizando la función *plot* de MATLAB podemos realizar un *gráfico de dispersión*, en el que los valores de la variable X se disponen en el eje horizontal y los de Y en el vertical. En la Figura 2 se muestran ejemplos de gráficos de dispersión.

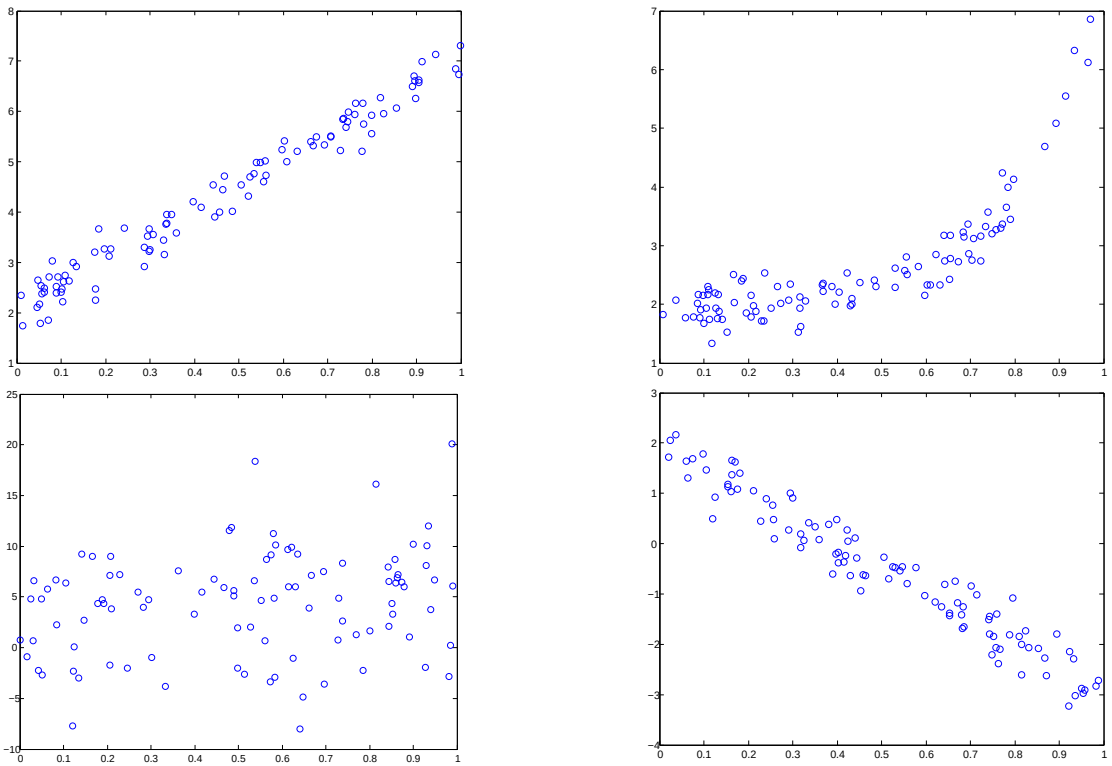

Figura 1: *Ejemplos de gráficos de dispersión.*

- ¿Qué conclusiones podrías sacar a partir de las gráficas sobre la relación entre las variables X e Y en cada ejemplo?
- Los puntos (X_i, Y_i) de la gráfica inferior izquierda han sido generados a partir del modelo lineal $Y_i = a + bX_i + \varepsilon_i$. ¿A qué crees que se debe que casi no se aprecie la relación lineal?
- ¿Existe relación lineal entre las variables X e Y representadas en la gráfica superior derecha? ¿Qué tipo de relación crees que existe?
- En la primera y última gráfica, los puntos (X_i, Y_i) han sido generados a partir del modelo lineal $Y_i = a + bX_i + \varepsilon_i$. ¿En que se diferencian ambos ejemplos?
- ¿Podrías determinar ejemplos reales en los que la relación entre variables se ajuste a alguna de las gráficas mostradas?

Consideremos ahora el siguiente ejemplo real, que ya hemos comentado en clase de teoría.
Se han obtenido veinte mediciones de la concentración de hidrógeno determinada con un método de cromatografía de gases (X), y la concentración determinada con un nuevo método de sensor (Y):

X	47	62	65	70	70	78	95	100	114	118	124	127	140	140	140	150	152	164	198	221
Y	38	62	53	67	84	79	93	106	117	116	127	114	134	139	142	170	149	154	200	215

Realiza el gráfico de dispersión correspondiente a la variable bidimensional (X,Y). ¿Observas algún tipo de relación entre X e Y?

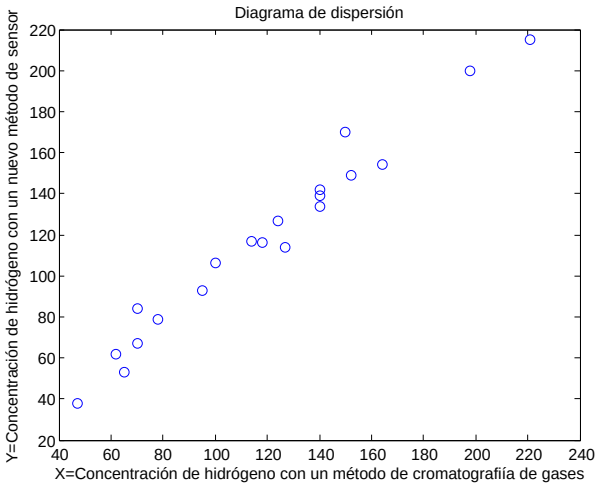

Figura 2: *Diagrama de dispersión*.

3 Medidas características de variables bidimensionales

Como vimos en clase de teoría, dada una variable bidimensional podemos calcular medidas características. Para el ejemplo anterior, calcula el vector de medias. Recuerda que la matriz de varianzas-covarianzas de la variable

bidimensional (X, Y) se definía como la matriz

$$S = \begin{pmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{pmatrix}$$

donde s_x^2, s_y^2 son las varianzas de las variables X e Y, respectivamente. El término s_{xy} es la **covarianza**, que se define como

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

Calcula la covarianza entre las variables X e Y del ejemplo que estamos tratando. La función

```
» cov(X,Y)
```

devuelve la matriz de varianzas-covarianzas, lo que nos permite conocer que la covarianza entre X e Y calculada por la función `cov` para esta muestra es 2.1569. ¿Es el mismo resultado que has obtenido programando directamente la fórmula de la covarianza?

La función cov: El lenguaje MATLAB ofrece la funcion `cov` para el cálculo de la matriz de varianzas-covarianzas de una variable bidimensional. La función `cov` devuelve entre otros la covarianza entre dos variables calculada como:

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

4 El modelo de regresión lineal simple con MATLAB

Para estudiar el grado de relación lineal que existe entre dos variables, calculamos el coeficiente de correlación lineal. Recuerda que

$$r(X, Y) = r_{xy} = \frac{s_{xy}}{s_x s_y}.$$

En MATLAB se calcula mediante

```
» corrcoef(X,Y)
```

que devuelve la matriz de correlación entre ambas variables. En este caso el coeficiente es 0.9852, próximo a 1 lo que indica una fuerte relación lineal creciente entre ambos métodos de medición.
Para obtener la recta de regresión realizaremos un ajuste por el método de mínimos cuadrados.

```
» m=polyfit(X,Y,1)
```

devuelve la recta de regresión

$$y = a + bx = -0{,}9625 + 1{,}0014x.$$

Comprueba que los coeficiente de regresión obtenidos responden a las ecuaciones

$$b = \frac{s_{xy}}{s_x^2}$$

y

$$a = \bar{y} - b\bar{x}.$$

A partir de la recta de regresión se pueden obtener las predicciones para la variable Y a partir de los valores conocidos de la variable X, sustituyendo convenientemente o bien utilizando el comando `polyval` de MATLAB.

> La función polyval: La función `polyval(m,X)` evalúa el polinomio con coeficientes almacenados en el vector m en todos los valores de la variable X

Por ejemplo, si la concentración de hidrógeno determinada con un método de cromatografía de gases es 112 unidades, entonces por el nuevo método será

```
>> polyval(m,112)
ans =

  111.1908
```

Almacena en un vector `yest` las predicciones para todos los valores de la variable X y representa gráficamente los valores reales y los valores pronosticados como se muestra en la Figura 3

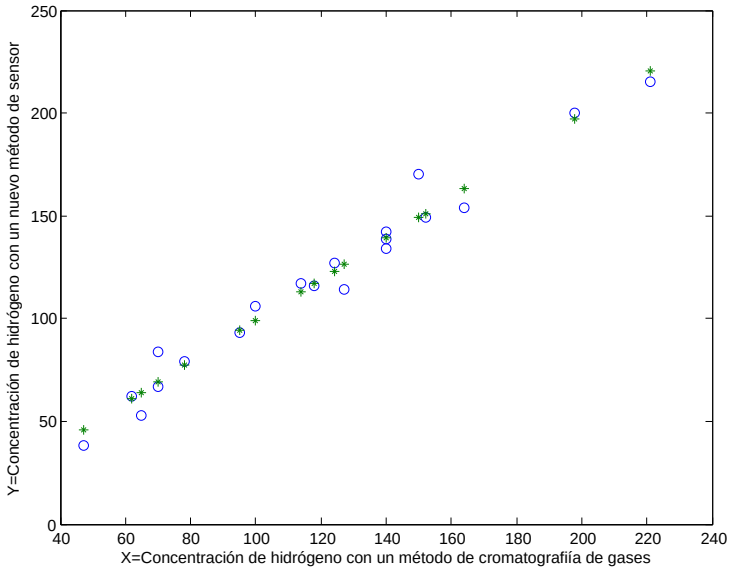

Figura 3: *Valores reales y pronosticados por la recta de regresión.*

Representa la recta de regresión y demuestra que dicha recta pasa por la media muestral.

4.1. El coeficiente de determinación

Una vez resuelto el problema de estimar los parámetros surge la pregunta de si la recta estimada es o no representativa para los datos. Esto se resuelve mediante el **coeficiente de determinación** (R^2) que se define como el cuadrado del coeficiente de correlación lineal.
El coeficiente de determinación toma valores entre 0 y 1 y representa el porcentaje de variabilidad de la variable dependiente que es explicada por la regresión. Calcula el coeficiente de determinación para el ejemplo de la concentración de hidrógeno.

Estadística
Práctica 4: VARIABLES ALEATORIAS UNIDIMENSIONALES

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Objetivos de la práctica

Hasta ahora hemos supuesto que disponíamos de un conjunto de datos que nos venía dado, pero hemos reflexionado muy poco acerca de cómo se obtienen estos datos. Se denomina "experimento" al proceso por el que obtenemos observaciones. Notar que podemos distinguir entre dos tipos diferentes de experimentos: deterministas y aleatorios.

1. Los experimentos deterministas son aquellos tales que siempre que se repitan bajo condiciones análogas, se obtiene el mismo resultado. Es decir, son totalmente predecibles.

2. Los experimentos aleatorios son aquellos tales que siempre que se repitan bajo condiciones análogas, se obtienen resultados diferentes, pero que se conocen previamente. Es decir, dentro de los posibles resultados, el resultado del experimento es impredecible.

Los experimentos que nos interesan son los que producen resultados impredecibles, es decir, los experimentos aleatorios. ¿Cúal es el tiempo de reacción de un determinado proceso químico (por ejemplo, tiempo de reacción del cloro en agua)? Dicho tiempo depende de multitud de factores que ocasionan que tengamos un amplio rango de valores posibles, pero antes de realizar el proceso es imposible de determinar con exactitud. Una variable aleatoria se define entonces como el resultado de realizar un experimento aleatorio. En el ejemplo, podemos definir la variable aleatoria $X =$*"Tiempo de reacción del cloro en agua"*. Este experimento se puede repetir tantas veces como se quiera. Si realizamos este experimento 100 veces y tomamos los tiempos de reacción, obtenemos una muestra de valores de la variable aleatoria de tamaño muestral 100. La población correspondería a todas las posibles veces que podemos intentar medir el tiempo de reacción del cloro en agua que, en principio, son infinitas.
Podemos dividir las variables aleatorias en discretas y continuas:

1. Variables aleatorias discretas son las que toman valores de un conjunto de valores discretos. Por ejemplo, el número de zapato, el número de hermanos, el resultado de lanzar un dado o el número de aciertos en una quiniela son variables aleatorias discretas.

2. Variables aleatorias continuas son las que toman valores de la recta real. Por ejemplo, el tiempo de reacción del cloro en agua o el índice de masa corporal son variables aleatorias continuas.

Pero gracias a nuestra experiencia sabemos que los valores de ciertos experimentos se repiten unos más que otros. Por ejemplo, sabemos que es más frecuente tener 7 aciertos en la quiniela que 14. Esto ya lo sabemos ya que hemos visto como obtener frecuencias absolutas y relativas. El concepto de probabilidad procede de estas frecuencias. Gracias a la probabilidad, podemos relacionar los conceptos de población y muestra e inferir si los resultados sobre una muestra pueden ser extrapolados al conjunto de la población.
El objetivo de esta práctica es simular, con la ayuda de MATLAB, algunos experimentos aleatorios sencillos y relacionarlos con modelos de probabilidad conocidos.

2 Distribución de Bernoulli

Vamos a comprobar los resultados del lanzamiento de una moneda. Sólo tenemos dos posibles resultados para cada lanzamiento: cara o cruz. El ejercicio es el siguiente. Vamos a escribir una función que simule los lanzamientos de una moneda. Para ello, utilizamos la siguiente función, donde los valores "C" corresponden a caras y los valores "X" corresponden a cruces:

```
function [res] = moneda(n)
% moneda(n)
% Esta función simula n lanzamientos de una moneda

% La funcion unidrnd genera datos de la Uniforme discreta
simulo=unidrnd(2,n,1);
res(simulo==1)='C';
res(simulo==2)='X';
return
```

Esta función proporciona n resultados del lanzamiento de una moneda al aire. Probamos los resultados para 1 y 5 lanzamientos.

```
>> moneda(1)
ans =
X

>> moneda(5)
ans =
CCXXX

>> moneda(5)
ans =
XCXXX
```

Aumentando el número de lanzamientos: Simula 10 lanzamientos de una moneda y calcula las frecuencias absolutas y relativas del número de caras y número de cruces obtenido. Representa el diagrama de barras correspondiente para las frecuencias relativas. ¿Qué ocurre si realizas 100 lanzamientos? ¿Y si realizas 1000 lanzamientos? ¿Y con 10000 lanzamientos? Fíjate en la Figura 2

Parece ser que, cuanto mayor es el número de intentos, más se acerca la frecuencia relativa del número de caras a 0.5. Este valor corresponde a lo que llamamos probabilidad de obtener cara. Claro está, la probabilidad de obtener cruz es 0.5. Por lo tanto, podemos decir que la variable aleatoria X ="*Resultado de lanzar una moneda al aire*" toma el valor "C" (cara) con probabilidad 0.5 y el valor "X" (cruz) con probabilidad 0.5. Se define entonces la función de probabilidad de la variable X como $P(X=0)=0{,}5$ y $P(X=1)=0{,}5$. Se dice que las variables aleatorias que tienen función de probabilidad $P(X=1)=p$ y $P(X=0)=1-p$, para un cierto valor de p entre 0 y 1, tienen distribución de Bernoulli de parámetro p.

3 Distribución binomial

Como hemos visto, una variable aleatoria Bernoulli toma dos posibles valores con probabilidades p y $1-p$, respectivamente. A continuación, consideramos la variable aleatoria binomial que se obtiene a partir de la variable aleatoria Bernoulli. Para ello, consideramos el siguiente ejemplo. Suponemos que una empresa se dedica a la fabricación de condensadores. Cada condensador consta de 60 tubos metálicos que deben soportar la circulación de agua a 310K. Se sabe que la probabilidad de que un tubo sea defectuosos es 0.04. ¿Cuál es la probabilidad de que un condensador no contenga ningún tubo defectuoso? Recuerda que para una distribución

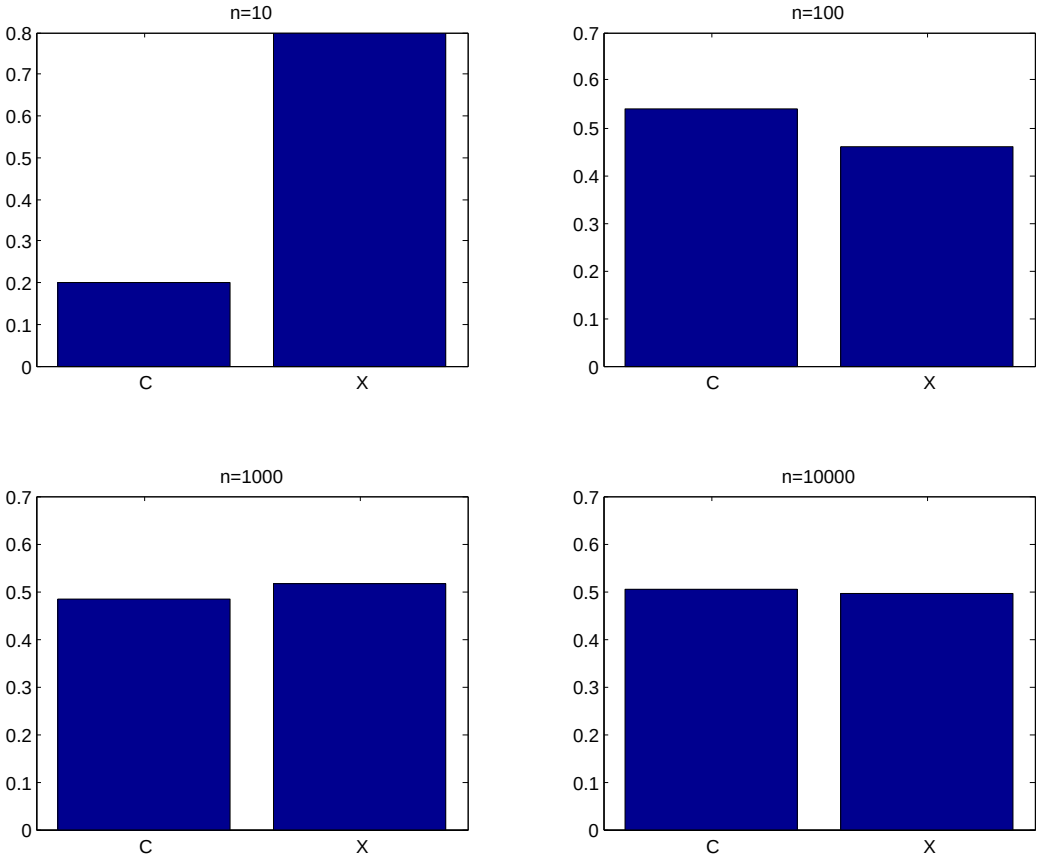

Figura 1: *Diagrama de barras para las frecuencias relativas.*

binomial de parámetros n y p, se tiene:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, 1, \ldots, n$$

Coeficientes binomiales: La función `nchoosek(n,k)` calcula 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). Las diferentes subconjuntos de k elementos se pueden obtener con la función `combnk`

Así, en el ejemplo anterior, la probabilidad de que un condensador no contenga ningún tubo defectuoso es:

```
>> nchoosek(60,0)*0.04^0*(1-0.04)^60
ans =
    0.0864
```

Supongamos ahora que la empresa consigue mejorar el sistema de fabricación de tubos reduciendo la probabilidad

de que un tubo sea defectuoso a 0.01. Calcula la probabilidad de que un condensador no tenga tubos defectuosos tras esta mejora.
La empresa decide sacar a mercado una segunda gama de condensadores más baratos (aquellos en los que el número de tubos defectuosos es mayor que 0 y menor o igual que 10). ¿Cuál es la probabilidad de que un condensador pertenezca a esta segunda gama?

> La función de masa binomial: La función `binopdf(k,n,p)` calcula la función de masa de una variable binomial $Bin(n,p)$ en k, es decir $P(X=k)$. Además, k puede ser un vector. Esta función se incluye dentro de la Statistics Toolbox de MATLAB.

4 El Quincunx o tablero de Galton

Vemos ahora qué ocurre cuando el número de intentos en una Binomial $Bin(n,p)$ va creciendo. El Quincunx o tablero de Galton (ver Figura 2) es un curioso aparato diseñado por el científico inglés Sir Francis Galton en el que una colección de bolitas van bajando de manera aleatoria. Cada vez que una bola llega a un piso "lanza" una moneda y si sale cara se desvía hacia la derecha y si sale cruz hacia la izquierda. Las bolitas se van acumulando en la parte inferior en unas cajas que representarían el número de caras obtenidas. El proceso da lugar, por lo tanto, a una distribución binomial donde el número de intentos n viene determinado por el número de filas del tablero.

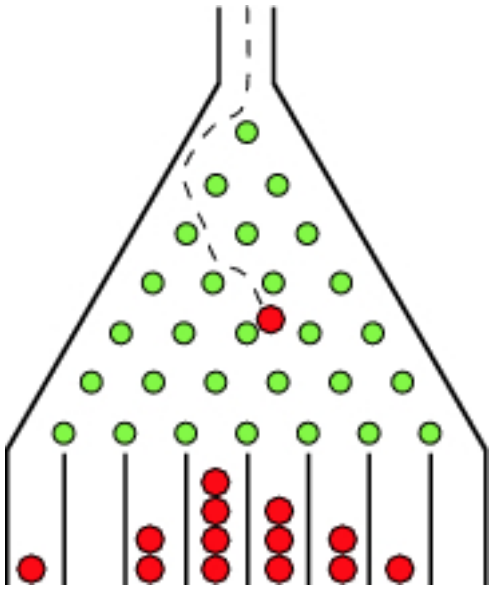

Figura 2: *Quincunx o tablero de Galton con* $n = 7$.

Existen varias páginas en internet con aplicaciones que simulan el funcionamiento de un tablero de Galton. Puedes consultar, por ejemplo, la página http://www.math.psu.edu/dlittle/java/probability/plinko/index.html
Programa una función que simule un tablero de Galton. Los argumentos de entrada serán el número de filas del tablero y el número de bolas que vamos a lanzar. Debes contar cuántas bolas caen en cada caja y representar el diagrama de barras correspondiente como se ve en la Figura 4.

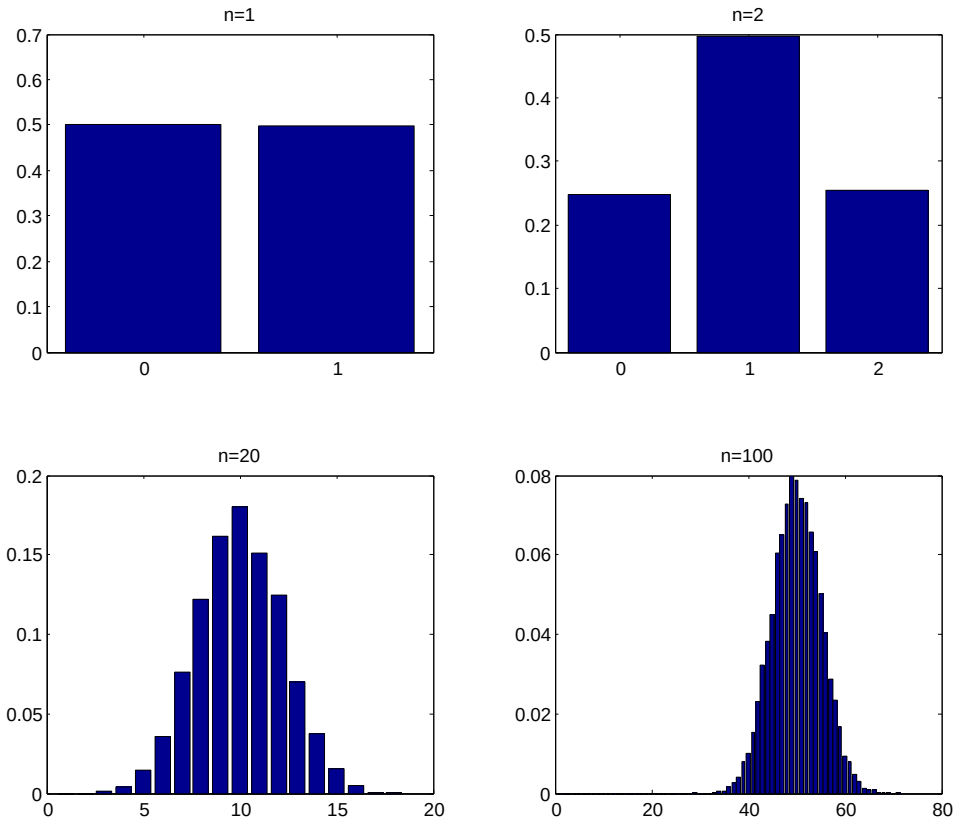

Figura 3: *Diagrama de barras para las frecuencias relativas tras lanzar 10000 bolas un tablero de Galton con 1, 2, 20 y 100 filas.*

5 Ejercicios

1. Construye una función en MATLAB que simule el siguiente experimento aleatorio. En una urna hay dos bolas negras y una bola blanca. Un jugador saca de la urna una bola al azar y gana el juego si la bola que ha sacado es blanca. Representa los correspondientes diagramas de barras para $n = 10$, $n = 100$, $n = 1000$ y $n = 10000$ repeticiones del juego. (Te puede ser de ayuda la función `unifrnd`).

2. Una mezcla contiene un 1 % de partículas de KCl y un 99 % de partículas de KNO. Si se sacan 10^4 partículas. ¿Cuál es la probabilidad de que se extraigan 10^2 partículas de KCl? ¿Cuántas partículas de KCl se espera sacar y cuál será la desviación típica si el experimento se realiza muchas veces?

3. Simula la construcción de un condensador como el del ejemplo de la Sección 3 siendo la probabilidad de fabricar un tubo defectuoso $p = 0{,}01$. Repite el proceso 1000 veces simulando la fabricación de 1000 condensadores. ¿Que proporción de condensadores están pefectos (ningún tubo averiado)?

Estadística
Práctica 5: VARIABLES ALEATORIAS UNIDIMENSIONALES II

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Objetivos de la práctica

El objetivo de esta práctica es repasar, con la ayuda de MATLAB, algunos modelos de probabilidad conocidos para variables aleatorias continuas.

2 Distribución uniforme continua

La distribución uniforme es una distribución muy simple cuya función de densidad es simplemente un tramo de línea recta horizontal, denominada densidad uniforme. Una variable aleatoria se dice **uniforme en el intervalo [a,b]**, y lo denotamos $X \in \text{Uniforme}[a, b]$, si su función de densidad es

$$f(x) = \frac{1}{b-a} \qquad \text{si} \quad x \in [a, b]$$

La media y la varianza de una Uniforme[a,b] son:

- La media será el punto medio del intervalo: $\mu = \frac{a+b}{2}$.
- La varianza es: $\sigma^2 = \frac{(b-a)^2}{12}$.

La densidad uniforme en MATLAB:La función `unifpdf` devuelve el valor de la función de densidad de una variable uniforme continua. Consulta la ayuda de la función y utilízala para representar la densidad de una Uniforme[5,10] como se muestra en la Figura 1.

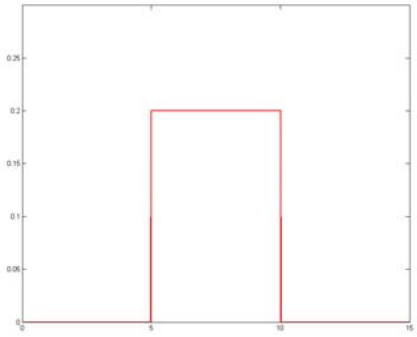

Figura 1: Función de densidad de una Uniforme[5,10].

3 Distribución exponencial

La distribución exponencial tiene especial utilidad para representar tiempos de vida: duración de una pieza hasta que se avería, longevidad de una persona, etc. Por ello, es una variable continua que toma valores en el intervalo $[0, +\infty)$. La definimos a través de su función de densidad.
Una variable aleatoria X tiene **distribución exponencial de parámetro** λ, $\lambda \in (0, +\infty)$, y lo denotamos $X \in$ Exponencial(λ), si su función de densidad viene dada por:

$$f(x) = \lambda e^{-\lambda x} \quad \text{si} \quad x \in [0, +\infty)$$

La media y la varianza de una Exponencial(λ) son:

- $\mu = \frac{1}{\lambda}$.
- $\sigma^2 = \frac{1}{\lambda^2}$.

La densidad exponencial en MATLAB:La función `exppdf` devuelve el valor de la función de densidad de una variable exponencial. Consulta la ayuda de la función y utilízala para representar las densidades de variables exponenciales para diferentes valores de λ, como se muestra en la Figura 2.

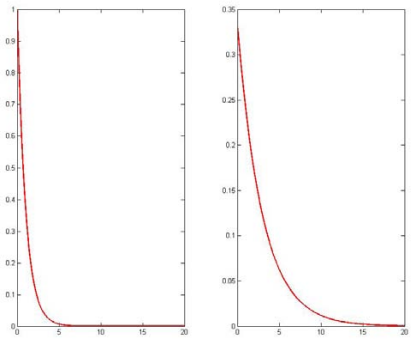

Figura 2: En la izquierda, función de densidad de una Exponencial(1). En la derecha, función de densidad de una Exponencial(1/3)

4 Distribución normal

La distribución normal es la más importante y de mayor uso de todas las distribuciones continuas de probabilidad. Por múltiples razones se viene considerando la más idónea para modelizar una gran diversidad de mediciones de la 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.
Una variable aleatoria continua Z se dice que tiene **distribución normal estándar**, y lo denotamos $Z \in N(0,1)$, si su función de densidad viene dada por:

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

La densidad normal en MATLAB:La función `normpdf` devuelve el valor de la función de densidad de una variable normal. Consulta la ayuda de la función y utilízala para representar la densidad de una normal estándar como se muestra en la Figura 3.

La distribución normal en MATLAB:La función `normcdf` devuelve el valor de la función de distribución de una variable normal. Consulta la ayuda de la función y utilízala para resolver el Ejemplo 1.

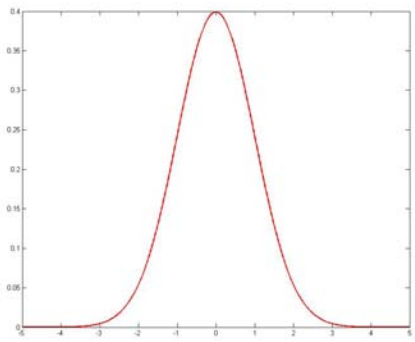

Figura 3: Función de densidad $f(z)$ para $Z \in N(0,1)$.

Ejemplo 1: Supongamos que $Z \in N(0,1)$. Calcula:

- $P(Z \leq 1{,}64)$.
- $P(Z > 1)$.
- $P(Z > -1{,}23)$.
- $P(Z \leq -0{,}53)$.
- $P(-1{,}96 \leq Z \leq 1{,}96)$.
- $P(-1 \leq Z \leq 2)$.
- $P(Z > 4{,}2)$

Cuantiles de la normal en MATLAB:La función `norminv` devuelve la inversa de la función de densidad de una variable normal. Consulta la ayuda de la función y utilízala para para resolver el Ejemplo 2.

Ejemplo 2: Supongamos que $Z \in N(0,1)$. Calcula los valores de z para los cuales:

- $P(Z \leq z) = 0{,}5$.
- $P(Z < z) = 0{,}95$.
- $P(Z > z) = 0{,}95$
- $P(Z \leq z) = 0{,}775$.
- $P(Z \leq z) = 0$.
- $P(Z \leq z) = 1$.

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

$$X = \mu + \sigma Z \in N(\mu, \sigma^2)$$

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

La función de densidad de una $N(\mu, \sigma^2)$ es

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

Utiliza de nuevo la función `normpdf` para representar la función de densidad de variables normales con diferentes valores de μ y σ, como en la Figura 4.

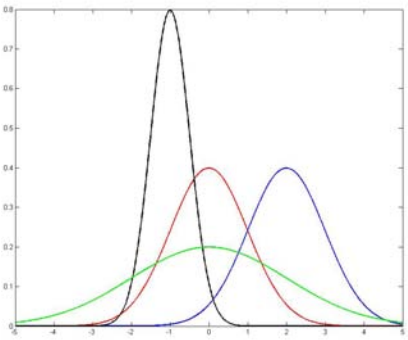

Figura 4: Funciones de densidad de variables normales con distintas medias y varianzas. En rojo densidad de una $N(0, 1)$.

5 Aproximación de otras distribuciones por la distribución normal

Utiliza las funciones `binopdf`, `poisspdf` y `normpdf` para justificar gráficamente las siguientes aproximaciones.

- Si $n \geq 30$, $np \geq 5$ y $nq \geq 5$ entonces la Binomial de parámetros n y p puede ser aproximada por una normal de media $\mu = np$ y varianza $\sigma^2 = np(1-p)$.
- Si $\lambda \geq 10$ entonces la Poisson de parámetro λ puede ser aproximada por una normal de media $\mu = \lambda$ y varianza $\sigma^2 = \lambda$.

También habíamos visto que:

- Si $n > 50$ y $p < 0{,}1$ entonces la Binomial de parámetros n y p puede ser aproximada por una Poisson de parámetro $\lambda = np$.

Estadística
Práctica 6: INFERENCIA ESTADÍSTICA: ESTIMACIÓN PUNTUAL E INTERVALOS DE CONFIANZA

Curso 2008/2009

Beatriz Pateiro López

Índice

1 Objetivos de la práctica

El objetivo de esta práctica es repasar, con la ayuda de MATLAB, algunos conceptos básicos de la inferencia estadística. Veremos cómo estimar puntualmente un parámetro desconocido. También repasaremos el concepto de intervalo de confianza y veremos como construirlos dependiendo del parámetro a estimar y de la información de la que dispongamos.

2 Planteamiento general de un problema de inferencia paramétrica

Consideramos un experimento aleatorio sobre el cual medimos una cierta variable aleatoria, que denotaremos por X. El objetivo es estudiar la variable aleatoria X, cuya función de distribución F es en mayor o menor grado desconocida.
Suponemos que la distribución de X, aún siendo desconocida, sigue un modelo como los vistos en temas anteriores. Para hacer inferencia, repetimos el experimento n veces en idénticas condiciones y de forma independiente.

- Una **muestra aleatoria simple** de tamaño n está formada por n variables

$$X_1, X_2, \cdots, X_n$$

  independientes y con la misma distribución que X.

- Llamamos **realización muestral** a los valores concretos que tomaron las n variables aleatorias después de la obtención de la muestra.

- Un **estadístico** es una función de la muestra aleatoria, y por tanto nace como resultado de cualquier operación efectuada sobre la muestra. Es también una variable aleatoria y por ello tendrá una cierta distribución, que se denomina **distribución del estadístico en el muestreo**.

- Para resolver el problema de estimación puntual, esto es, para aventurar un valor del parámetro poblacional desconocido, escogemos el valor que ha tomado un estadístico calculado sobre nuestra realización muestral. Al estadístico escogido para tal fin le llamamos **estimador** del parámetro. Al valor obtenido con una realización muestral concreta se le llama **estimación**.

El problema radica en elegir un "buen estimador", es decir, una función de la muestra con buenas propiedades.

3 Estimación puntual e intervalo de confianza para la proporción

Resolvemos ahora un problema práctico de inferencia a través de un ejemplo clásico: **la paradoja de Mèré**.

> *Existe una vieja historia sobre el Caballero de Mèré, un famoso jugador francés del siglo XVII. El Caballero de Mèré iba de camino al estado de Poitou cuando conoció a Blaise Pascal, uno de los matemáticos más famosos del siglo. De Mère le planteó dos problemas a Pascal, ambos relacionados con juegos de azar. En 1654 Pascal le propuso estas paradojas a Pierre Fermat, otro gran científico de la época con quien mantenía contacto por correspondencia. Ambos llegaron a la misma conclusión, lo cual alegró a Pascal, quien escribe en una de sus cartas: "Ya veo que la verdad es la misma en Toulouse y en Paris".*
> *La primera paradoja está relacionada con un juego de dados. Aunque no está claro cuanto hay de cierto en la historia, se cree que el Caballero de Mèré era muy aficionado al juego y que, basándose en su propia experiencia, proponía la siguiente apuesta:* **Él ganaba si al tirar cuatro dados salía al menos un 6**.

1. ¿Crees que era un juego "rentable" para el Caballero de Mèré?
2. ¿Cómo simularías una apuesta?
3. ¿Cómo estimarías la probabilidad de ganar el juego?
4. ¿Cuál es la distribución en el muestreo del estimador?
5. ¿Sabrías calcular la probabilidad de ganar el juego?

Empezaremos por plantear el problema. El parámetro desconocido para el Caballero de Mèré es p, donde

$$p = \text{Probabilidad de ganar el juego.}$$

Estimamos p mediante un estimador $\hat{p}$, que se obtiene simulando n partidas como:

$$\hat{p} = \frac{n_G}{n} = \frac{\text{Número de partidas ganadas}}{\text{Número de partidas jugadas}}.$$

Jugando a ser Mèré con MATLAB: Simula con MATLAB una apuesta del Caballero de Mèré y decide si has ganado la apuesta o no. Simula 1000 apuestas, ¿cuántas veces has ganado? ¿cuál es la probabilidad estimada de ganar $\hat{p}$ a partir de esas 1000 apuestas? Juega otras 1000 partidas. ¿Cuánto vale ahora $\hat{p}$?

Habíamos visto en clase de teoría que

- La media de $\hat{p}$ es $\mathbb{E}(\hat{p}) = p$
- La varianza de $\hat{p}$ es $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$
- Para n grande $\hat{p}$ es aproximadamente normal.

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

Además de la estimación puntual, podemos calcular un intervalo de confianza para p. Hemos visto que

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

es el intervalo de confianza para p con nivel de confianza $1-\alpha$. En la expresión anterior, $z_{\alpha/2}$ denota el número real tal que $P(Z > z_{\alpha/2}) = \alpha/2$, siendo $Z \in N(0,1)$, ver Figura 1.

Intervalo de confianza para p con MATLAB: Calcula el intervalo de confianza para p con nivel de confianza 0.95 a partir de $\hat{p}$. ¿Cuál es le intervalo de confianza al 90 %?

Distribución en el muestreo de $\hat{p}$: Cada día el caballero de Mèré realiza 1000 apuestas y apunta en una libreta (de la época) la proporción de partidas ganadas. En los últimos 5 años tiene 1500 anotaciones. Simula la situación descrita. ¿Cómo se distribuyen los valores de $\hat{p}$?

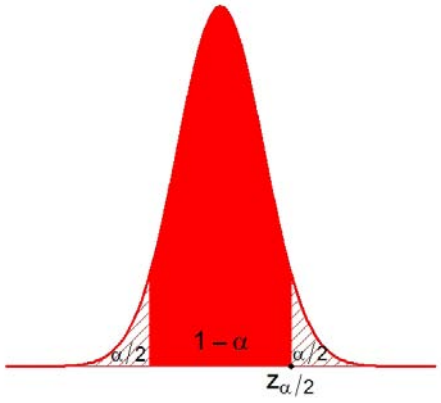

Figura 1: $z_{\alpha/2}$ denota el número real tal que $P(Z > z_{\alpha/2}) = \alpha/2$, siendo $Z \in N(0,1)$.

Y con la teoría de la probabilidad llegó la solución al problema. ¿Sabrías calcular exactamente el valor de p? A la vista del resultado, ¿hacía bien el caballero de de Mèré en apostar su dinero a este juego?

Ahora que sabemos cuál es el verdadero valor de p podemos entender el significado que tiene el nivel de confianza.

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

Interpretación del nivel de confianza. Construye los diferentes intervalos de confianza construidos en base a las 1500 anotaciones de $\hat{p}$ que has generado. ¿Cuántos de esos intervalos contienen al verdadero valor de p?

Cansado de este juego, el Caballero de Mèré propuso una nueva apuesta: **el caballero de Mèré gana si al tirar 24 veces 2 dados le sale al menos un 6 doble.** *¿Qué dirías de esta nueva apuesta?*

4 Ejercicios

- Importa de nuevo a MATLAB los datos del fichero **IQ0809.csv** utilizado en la Práctica 2. Podemos considerar dichos datos como una muestra representativa de los estudiantes universitarios de primer curso. ¿Cómo estimarías la media y la varianza de la variable altura?
- Suponiendo que la variable altura sigue una distribución normal, construye un intervalo de confianza para la altura media basándote en los datos de la muestra con un nivel de confianza del 95 %. Consulta la ayuda de las funciones `tpdf`, `tcdf`, `tinv`.
- De igual modo, construye el intervalo de confianza para la varianza de la altura basándote en los datos de la muestra. Consulta la ayuda de las funciones `chi2pdf`, `chi2cdf`, `chi2inv`.