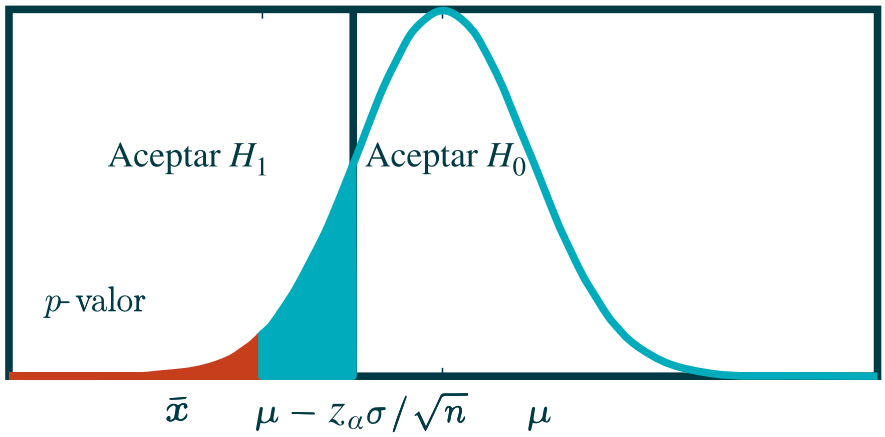
Mondragon Unibertsitatea

Escuela Politécnica Superior

PROBABILIDAD Y ESTADÍSTICA

$$p-\text{valor} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}} e^{-\frac{1}{2}x^2}\, dx$$



DANIEL SOLER MALLOL

2022

Edita: Mondragon Unibertsitateko Zerbitzu Editoriala
Loramendi kalea, 4 - (23 p.k.)
20500 ARRASATE-MONDRAGON (Gipuzkoa)
Diseño y maquetación: AZK Taldea
ISBN: 978-84-09-46048-9
Depósito Legal: 01162-2022
1ªEdición, 2022

PROBABILIDAD Y ESTADÍSTICA

Apuntes de la asignatura de Estadística presentados en

MONDRAGON UNIBERTSITATEA

Autor

DANIEL SOLER MALLOL

En ARRASATE a 2022

AGRADECIMIENTOS

Agradezco a la familia por las hora que les he robado para escribir este libro. Agradezco también a Antonio Lopez de Lacalle, Xabi Artetxe, Txus Martínez, Ainhoa Iturrsape, Jatsu Lizarribar, Germán Albistegui, Zigor Oruna, Javi Arrasate, Mikel Agirre, Josune Urien, Itziar Fraile y Julián Elorza, profesores de Mondragon Unibertsitatea que han compartido conmigo la impartición de esta asignatura. Con ellos he discutido y aprendido conceptos aquí expuesto, además con muchos de ellos nos hemos inventado buena parte de los ejercicios que se proponen en el texto.

También se agradece a la propia institución Mondragon Unibertsitatea por la implicación en la publicación de este libro.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

PREFACIO

El objetivo de este libro es acompañar a los alumnos de ingeniería en la asimilación de conceptos básicos de probabilidad y estadística. No pretende ser muy riguroso en la descripción matemática por eso son pocos los casos en los que se explicitan las demostraciones se los teoremas enunciados. La idea es explicar brevemente un concepto nuevo y a continuación utilizarlo resolviendo algunos ejemplos, un poco siguiendo el esquema que se sigue cuando se imparte esta asignatura en Mondragon Unibertsitatea. Al final de cada capítulo, además, hay un listado de ejercicios adicionales con la correspondiente solución para que el estudiante pueda probar el grado de asimilación del tema, estos ejercicios se han venido planteando durante distintos cursos en dicha universidad y se han ampliado con ejercicios planteados en exámenes.

Hay algunos temas, que están un poco más desarrollados aquí de lo que se suele hacer en clase, pero en general se puede decir que el nivel de los ejemplos, ejercicios y explicaciones son suficientes para comprender las bases de la probabilidad y la estadística al nivel requerido por un ingeniero. Sin embargo, hay que recalcar que en este libro falta una parte importante que los alumnos de Mondragon Unibertsitatea adquieren mediante las prácticas multidisciplinarias y sobre todo durante el desarrollo de proyecto semestral. Durante estas actividades formativas, los alumnos usan los conceptos y las herramientas que en este texto se les da, para aplicarlos en un caso más o menos real, pero, además, se les exige que se autoformen en un tema muy importante en para cualquier profesional, *la regresión lineal*. No se descarta añadir un capítulo en futuras ediciones que incluya este tema, pero de momento, se ha optado por no incluirlo para forzar a los alumnos a buscar en textos alternativos.

La herramienta informática por la que se ha optado en este libro es Matlab ®, porque sin ser un software específico para trabajar en estadística y probabilidad como podría ser R o Minitab ®, es un software con el que los alumnos de Mondragon Unibertsitatea están familiarizados. Probablemente sería interesante explicar cómo usar otras herramientas alternativas a Matlab como podría ser Microsoft Excell ®, o alternativas de software libre como pueden ser Octave, Python o Julia. No descarto que en un futuro se complemente o modifique este texto para dar cabida a esta inquietud.

1

INTRODUCCIÓN

La estadística es la ciencia que estudia la recopilación, presentación, análisis y uso de datos para tomar decisiones y resolver problemas.

Estos datos pueden ser numéricos, *variables* o datos cualitativos *atributos*, por ejemplo, si estamos interesado en la temperatura alcanzada en la herramienta de corte durante un el fresado de metales, la temperatura es un dato numérico y el tipo de material un atributo.

En general, lo que interesará como ingeniero es la toma de decisiones racionales basadas en una serie de datos experimentales. El objeto de estudio es la *población*. Por ejemplo, si se quiere mejorar la resistencia al impacto de unas impresoras, la población estaría compuesta por todas las impresoras fabricadas por una determinada compañía o una planta concreta. Para ello deberían hacerse test como, por ejemplo, los que se muestran en el siguiente video, https://www.youtube.com/watch?v=WX5BliU5adg .

Sin embargo, para recopilar datos, obviamente, solo se analiza un subconjunto de la *población*. este subconjunto es la *muestra*. La *estadística inferencial* engloba todos los métodos que permiten tomar decisiones a partir de los datos muestrales, para ello es necesario recopilar datos organizarlos y analizarlos, de esto se encara la *estadística descriptiva*. La base matemática que permite hacer el salto de los datos muestrales a la población es la *probabilidad*.

El siguiente ejemplo sirve para ilustrar la relación entre estadística inferencial y teoría de probabilidades:

> **Ejemplo 1.0.1** Suponga que un ingeniero tiene que analizar la resistencia al impacto de unas impresoras. Se espera y se anticipa que ocasionalmente habrá artículos defectuosos, sin embargo, se determina que, a largo plazo, la empresa sólo puede tolerar 5 % de impresora defectuosas en el proceso. Para el análisis se realizan una serie de pruebas a la resistencia al choque de 100 impresoras y se encuentra con que 10 son defectuosas.
>
> En este caso la *población* representa conceptualmente todos las impresoras posibles en el proceso de producción y obviamente estas 100 impresoras representan la *muestra*. Los elementos de probabilidad permiten al ingeniero determinar qué tan concluyente es la información muestral respecto de la naturaleza del proceso.
>
> Suponga que el proceso es aceptable, es decir, que su producción no excede un 5 % de artículos defectuosos, en tal caso la teoría de probabilidades permite al ingeniero determinar que hay una probabilidad de 0.0282 de obtener 10 o más artículos defectuosos en una muestra aleatoria de 100 artículos del proceso. Esta pequeña probabilidad sugiere que, en realidad, a largo plazo el proceso tiene un porcentaje de artículos defectuosos mayor al 5 %. En otras palabras, en las condiciones de un proceso aceptable casi nunca se obtendría la información muestral que se obtuvo. Sin embargo, ¡se obtuvo!, como es

evidente que la probabilidad de que se obtuviera sería mucho mayor si la tasa de artículos defectuosos del proceso fuera mayor que 5 %, el ingeniero tiene cierta confianza en que porcentaje de defectuosos es mayor que el 5 % y por tanto que el proceso efectivamente es defectuoso.

En el siguiente capítulo se dan las nociones básicas de la estadística descriptiva, en los capítulos del 4 hasta 7 se explican los fundamentos de la probabilidad y se introducen las nociones fundamentales de variable aleatoria, esperanza matemática y se explican las principales distribuciones de probabilidad tanto discretas como continuas. El capítulo 8 se sale un poco del nivel del conjunto del texto, y el lector puede omitirlo si lo desea puesto que no es necesaria su lectura para la comprensión de los siguientes capítulos, pero se ha decidido incluirlo para el lector que quiera profundizar un poco más. Los últimos capítulos (9 y 10) se reservan para introducir los entresijos de la estadística inferencial. En los capítulos 2 y 11 se dan algunas herramientas matemáticas e informáticas para desarrollar los temas propuestos. Además, junto con este ejemplar se pone a disposición del lector unos "live-scrips" de Matlab donde se muestra, con ejemplos concretos, el contenido del último capítulo.

A este volumen le faltaría un capítulo dedicado a la regresión lineal para cubrir la totalidad de los conceptos exigibles en una asignatura de "Probabilidad y estadística" a nivel de 2º de Grado de una carrera técnica. Como se ha indicado en el prefacio, esto se ha hecho para forzar al lector a buscar la información en textos alternativos, que usarán otra nomenclatura y le exigirán un esfuerzo extra para auto-formarse.

2

ESTADÍSTICA DESCRIPTIVA

La estadística descriptiva es la parte de la estadística que se dedica organizar, representar gráficamente y analizar los datos experimentales, además, la estadística descriptiva incluye las técnicas que permiten usar este estudio para inferir características de la *población* estudiada.

Las definiciones que se dan continuación son imprescindibles para manejar y comprender cualquier estudio con datos experimentales, por tanto, para cualquier disciplina científico-técnica. Además, muchas de estas nociones se van a usar en los capítulos posteriores por lo que, a pesar de tratarse de un tema sencillo es muy importante.

2.1 INTRODUCCIÓN

Definición 2.1 *Llamaremos* población *estadística al conjunto de referencia sobre el cual van a recaer las observaciones.*

Las poblaciones podrán ser finitas o infinitas, dependiendo del número de elementos que las forman.

Definición 2.2 *Se llama unidad estadística o* individuo *a cada uno de los elementos que componen la población estadística.*

Definición 2.3 *Se llama* muestra *a cualquier subconjunto de elementos de la población. El número de elementos de la misma se llama tamaño de la muestra.*

La observación del individuo la describimos mediante uno o más caracteres. Hay caracteres que son cuantificables, como, por ejemplo, la edad, el peso y la estatura de las personas, pero hay otros que no, como, por ejemplo, el color de los ojos, el sexo, etc. A los primeros se los llama *caracteres cuantitativos*, y a los segundos, *caracteres cualitativos.*

Definición 2.4 *Se llaman* variables estadísticas *a los valores numéricos que adoptan los caracteres cuantitativos.*

Las variables estadísticas las podemos clasificar de la forma siguiente:

a) Variables estadísticas discretas. Son aquellas que toman valores aislados, por tanto, una cantidad numerable, y que no pueden tomar ningún valor entre dos consecutivos fijados.

b) Variables estadísticas continuas. Son aquellas que pueden tomar infinitos valores en un intervalo dado. Es decir, pueden tomar valores entre dos consecutivos, por muy próximos que los fijemos.

2.2 TABLAS ESTADÍSTICAS

Vamos a estructurar y ordenar los datos obtenidos en la observación de una muestra de N elementos de la correspondiente población. Centrémonos en el caso de una variable estadística X que puede tomar distintos valores $x_1, x_2, \ldots, x_k$, está claro que cada uno de éstos puede aparecer repetido más de una vez. Por eso se definen las siguientes frecuencias:

Definición 2.5 *Llamaremos* frecuencia absoluta (n_i) *de un valor* x_i *de la variable estadística* X *al número de veces que aparece repetido dicho valor en el conjunto de las observaciones realizadas.*

Definición 2.6 *Llamaremos* frecuencia relativa (f_i) *de un valor* x_i *de la variable estadística* X *al cociente entre la frecuencia absoluta y el número de observaciones realizadas* (N)

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

Definición 2.7 *Llamamos* frecuencia absoluta acumulada (N_i) *en el valor* x_i *a la suma de las frecuencias absolutas de los valores inferiores o iguales a él.*

Cuando los valores x_i son numéricos, esto se ordenan de forma creciente de modo que la frecuencia absoluta acumulada del último valor será N. Es decir, que

$$N_k = N$$

Definición 2.8 *Llamamos* frecuencia relativa acumulada (F_i) *en el punto* x_i *al cociente entre la frecuencia absoluta acumulada y el número de observaciones realizadas* (N) :

$$F_i = \frac{N_i}{N} = \frac{\sum_{j=1}^{i} n_j}{N} = \sum_{j=1}^{i} \frac{n_j}{N} = \sum_{j=1}^{i} f_j$$

Propiedades de las frecuencias

Sea N el número total de observaciones realizadas. Podemos destacar las siguientes propiedades:

a) $n_1 + n_2 + \cdots + n_k = N \qquad \sum_{i=1}^{k} n_i = N$

b) $f_1 + f_2 + \cdots + f_k = \frac{n_1}{N} + \frac{n_2}{N} + \cdots + \frac{n_k}{N} = 1$

c) $N_k = N$

d) $F_k = 1$

e) $0 \leq n_i \leq N$

f) $0 \leq f_i \leq 1$

g) El porcentaje $(\,\%)$ correspondiente a un valor x_i de la variable se obtiene multiplicando la frecuencia relativa por 100

$$(\,\%)_{x_i} = f_i \cdot 100$$

2.2.1 *Tabla de frecuencias de una variable discreta*

La confección de una tabla de frecuencias para una variable discreta se consigue ordenando los distintos valores de la misma de menor a mayor y anotando las distintas frecuencias n_i, n_f, N_i, F_i

x_i	n_i	f_i	N_i	F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	N_k	F_k

2.2.2 *Tabla de frecuencias de una variable continua*

En el caso de una variable continua, la estrategia es similar, a modo de ejemplo supongamos que se han medido estaturas de sesenta personas, obteniéndose los siguientes resultados en metros:

1.62	1.71	1.57	1.61	1.80	1.91	1.58	1.63	1.62	1.70
1.75	1.68	1.54	1.79	1.72	1.68	1.90	1.69	1.73	1.85
1.60	1.60	1.62	1.77	1.71	1.89	1.92	1.65	1.99	2.05
1.41	1.67	1.93	1.55	2.04	1.73	1.80	1.83	1.75	1.66
1.93	1.85	1.84	1.68	1.63	1.75	1.77	1.84	1.85	1.90
2.00	1.83	2.01	1.82	1.65	1.72	1.68	1.73	1.54	1.65

A la hora de analizar estos datos es aconsejable, agrupar los datos en intervalos y hacer un recuento de las observaciones que caen dentro de cada uno de ellos. En el caso de las estaturas se podría agrupar los datos según el siguiente criterio.

Estaturas				
$[1.40 - 1.60]$	$]1.60 - 1.70]$	$]1.70 - 1.80]$	$]1.80 - 2,00]$	$]2,00 - 2,10]$

No cabe duda de que tomar como unidad de estudio el intervalo y no cada uno de los valores de la variable nos va a suponer una simplificación en nuestro trabajo, pero desgraciadamente también una pérdida de información. Hemos de elegir un número de intervalos que equilibre estos dos aspectos.

Definición 2.9 *Definimos la* marca de clase *como el punto medio de cada intervalo. Es decir, el valor que nos representa la información que contiene un intervalo.*

A la diferencia entre el extremo superior y el extremo inferior de cada intervalo la llamaremos *amplitud del intervalo.*

Se entiende que cuando hacemos una agrupación en intervalos, para nosotros solamente cuenta el número de observaciones que caen dentro del mismo y no la distribución (o colocación) en su interior. En otras palabras, nosotros suponemos que la distribución de

estos valores en el intervalo es *homogénea.* De ahí la pérdida de información a la que antes se aludía.
La elección de intervalos, tanto en número como en amplitud, es algo subjetivo del investigador y no se encuentra sometido a ninguna norma rígida, sino que habrá de analizarse en cada caso.

Tabla de frecuencias de una variable agrupada en intervalos

Intervalos	Marcas de clase x_i	n_i	f_i	N_i	F_i
$a_0 - a_1$	x_1	n_1	f_1	N_1	F_1
$a_1 - a_2$	x_2	n_2	f_2	N_2	F_2
$a_2 - a_3$	x_3	n_3	f_3	N_3	F_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_{l-1} - a_l$	x_l	n_l	f_l	N_l	F_l

2.3 REPRESENTACIONES GRÁFICAS DE CARACTERES CUANTITATIVOS

En esta sección se describen los gráficos típicos que se usan para la representación de caracteres cuantitativos, además, se explica cómo realizar estos gráficos. Sin embargo, se recomienda usar algún software informático para generarlos. Todos lo gráficos que aparecen en este capítulo se han realizado con *Matlab* ©. En el capítulo 11 se explica cómo hacerlos.

- **Diagrama de barras**

  Esta representación es válida para las frecuencias de una variable discreta.
  Se colocan en abscisas los distintos valores de una variable y sobre cada uno de ellos se levanta una línea perpendicular, cuya altura es la frecuencia (absoluta o relativa) de dicho valor. Así obtenemos un conjunto de barras verticales cuya suma de longitudes debe ser N o 1, dependiendo de si las frecuencias representadas son absolutas o relativas. *(Figura 2.1)*

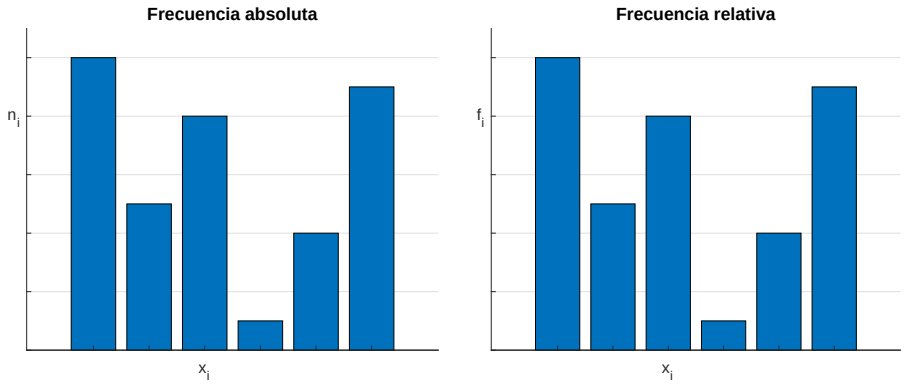


Figura 2.1: Diagramas de barras.

- **Histograma**
  Para las variables estadísticas agrupadas en intervalos vamos a representar las frecuencias mediante áreas.

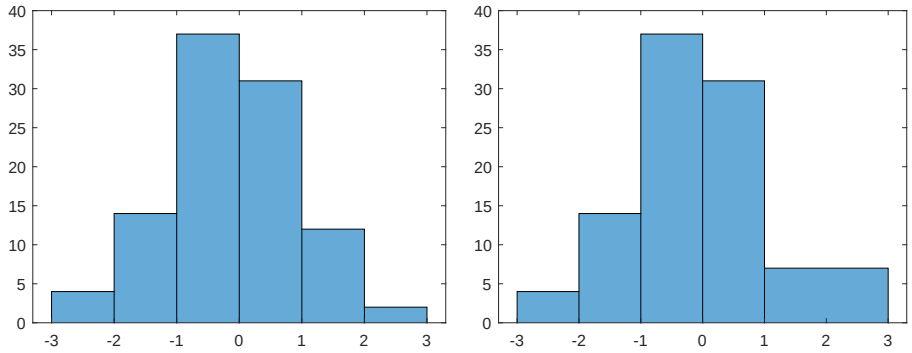


Figura 2.2: Histogramas. En de la izquierda todas las clases tienen la misma amplitud, la altura de cada clase es proporcional a la frecuencia absoluta. En el de la derecha, el área es proporcional a la frecuencia.

Un histograma se obtiene levantando sobre cada intervalo de clase un rectángulo cuya área sea igual a la frecuencia del mismo. En general cada rectángulo tiene la misma anchura, por lo que la altura de cada barra es proporcional a la frecuencia correspondiente, ya sea relativa o absoluta. En algunos casos las clases son de anchura no constante, en este caso la altura correspondiente a cada rectángulo que habrá que levantar sobre el eje de ordenadas será el cociente entre el área y la base del mismo (amplitud del intervalo).

En la figura 2.2 se pueden observar los histogramas realizados sobre un conjunto de 100 datos, en de la izquierda todas las clases tienen la misma amplitud, mientras que en el de la derecha la última clase tiene amplitud doble, por lo que la altura es la mitad de la frecuencia absoluta de la clase.

Tanto los diagramas de barras como los histogramas pueden usarse también para representar las frecuencias acumuladas. En la figura 2.3 se puede ver un diagrama de barras de frecuencias acumuladas.

- **Polígono de frecuencias**

  a) Si la variable es discreta, el polígono de frecuencias se obtiene uniendo los extremos superiores de las barras en el diagrama de barras. En la figura 2.3) se muestran los polígonos de frecuencia relativa y acumulada correspondiente a una muestra de 20 elementos.

  b) Si la variable está agrupada en intervalos, el polígono de frecuencias se obtiene uniendo los puntos medios de las bases superiores de cada rectángulo en el histograma, ver Figura 2.7.

- **Diagrama de cajas y bigotes o Box plot**

  Otra presentación útil para reflejar propiedades de una muestra es la gráfica de caja y extensión o bigotes (Montgomery y Runger, (Mexico, 2000)). Para generar

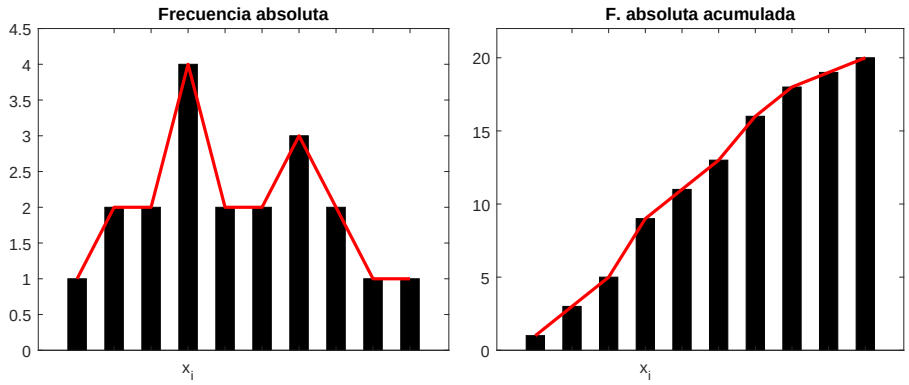


Figura 2.3: Polígono de frecuencias relativas y acumuladas de una muestra de 20 individuos.

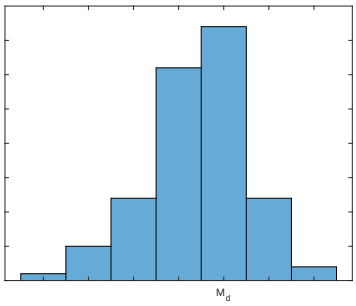


Figura 2.4: Polígono de frecuencias usando un histograma.

este tipo de diagramas se necesita de ciertas definiciones que se van a dar en las secciones 2.5 y 2.6

En concreto se representa una caja que tiene una arista sobre el primer cuartil (percentil 25) y la otra arista sobre el tercer cuartil (percentil 75), es decir, la longitud de la caja es el rango intercuartílico. Dentro de la caja se dibuja una linea, situado sobre el segundo cuartil o mediana, que divide el rectángulo es dos partes. Además de la caja se prolongan unas lineas o bigotes hasta los valores extremos, siempre estas observaciones se encuentren entre el 0 y 1.5 veces el rango intercuartil.

Las observaciones que se encuentran entre 1.5 y 3 veces el rango intercuartil a partir de las aristas del rectángulo, se denominan *valores atípicos* y se representan mediante puntos. A los que distan más de 3 veces del rango intercuartil a partir de las aristas de la caja son *valores atípicos extremos* y se suele usar un símbolo distinto para representarlos.

A modo de ejemplo, la figura 2.5 representa el contenido en nicotina de los 40 cigarrillos de una muestra aleatoria recogidos en al siguiente tabla:

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97	0.85	1.24
1.58	2.03	1.70	2.17	2.55	2.11	1.86	1.90	1.68	1.51
1.64	0.72	1.69	1.85	1.82	1.79	2.46	1.88	2.08	1.67
1.37	1.93	1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

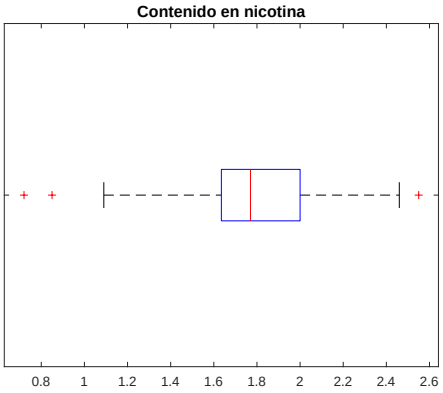


Figura 2.5: Diagrama de caja y extensiones en el que se pude observar la presencia de valores atípicos.

2.4 REPRESENTACIONES GRÁFICAS DE CARACTERES CUALITATIVOS

Para representar datos cualitativos además de los ya mencionados diagramas de barras se suelen usar los diagramas de sectores.

A modo de ejemplo se usara el recuento de las cifras de alumnos nuevos por grado de Mondragon Unibertsitatea que se especifican en la tabla 2.1.

Tabla 2.1: Alumnos nuevos matriculados en grados técnicos de Mondragon Unibertsitatea en el año 2021.

Grado	*Núm. alumnos nuevos*
Diseño	131
Electrónica	50
Informática	99
Mecánica	132
Organización	96
Energía	48
Biomecánica	103
Ecotecnología	12
Mecatrónica	188

- **Diagrama de barras**

  Se representan en abscisas los distintos caracteres cualitativos y se levantan sobre ellos rectángulos de bases iguales que no tienen por qué estar solapados y cuya altura será la correspondiente a la frecuencia absoluta de cada carácter.

- **Diagramas de sectores**

  En un círculo se asigna un sector circular a cada uno de los caracteres cualitativos, siendo la amplitud del sector proporcional a la frecuencia del carácter. Esto se

consigue haciendo corresponder 360° a la suma de todas las frecuencias (N) de los caracteres y hallando la correspondiente proporcionalidad.

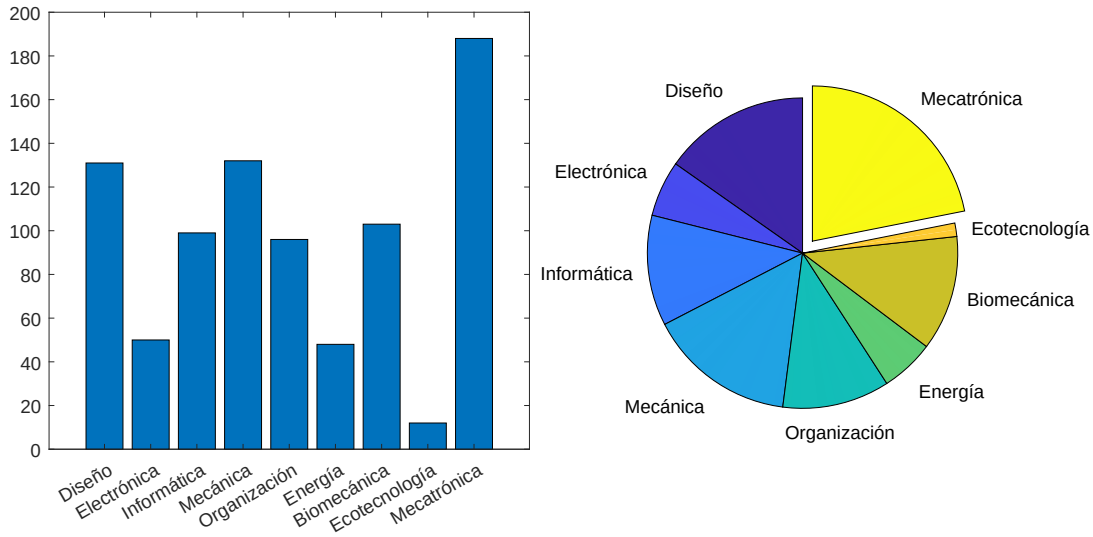


Figura 2.6: Diagrama de barras y de sectores correspondientes a los datos de la tabla 2.1.

2.5 MEDIDAS DE CENTRALIZACIÓN

A veces es conveniente reducir la información obtenida a un solo valor o a un número pequeño de valores para facilitar la comparación entre distintas muestras o poblaciones. Estos valores, que de alguna forma centralizan la información, reciben el nombre de *medidas de tendencia central.*

En todas las definiciones que daremos supondremos que la variable estadística X toma los valores $x_1, x_2 \ldots, x_k$ con las frecuencias $n_1, n_2, \ldots n_k$, respectivamente.

Definición 2.10 *Se llama* media aritmética $(\bar{X})$

$$\bar{X} = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_k n_k}{N} = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \sum_{i=1}^{k} x_i f_i$$

Naturalmente, si cada valor de la variable aparece una sola vez, $x_1, x_2, \ldots, x_k$, entonces el denominador será k

$$\bar{X} = \frac{\sum_{i=1}^{k} x_i}{k}$$

Cuando los datos corresponden a una población, se suele denotar el valor de la media aritmética con la letra griega μ.

Definición 2.11 *Se llama* media geométrica $(\bar{X}_G)$

$$\bar{X}_G = \sqrt[N]{x_1^{n_1} \cdot x_2^{n_2} \cdot \dots \cdot x_k^{n_k}}$$

Definición 2.12 *Se llama* media armónica $(\bar{X}_H)$

$$\bar{X}_H = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$$

Este parámetro no está definido si el valor 0 está dentro de la muestra.

Para ver otras medias posibles y algunas de sus aplicaciones se puede consultar por ejemplo Narvaiza et al., 2001a.

Definición 2.13 *Se llama* mediana (M_e) *a la medida central que, puestos los valores de la variable ordenados de forma creciente, deja igual número de observaciones inferiores que superiores a ella, es decir, es el valor de la muestra que cumple que por lo menos el 50 % de los valores son menores o iguales que él y mayores iguales que él, en caso de que haya dos números que cumplan esta condición se suele hacer la media aritmética entre ellos.*

Por ejemplo, si $X = \{1, 3, 7, 10, 15, 22, 36\}$, entonces la mediana es el valor 10.
Si hubiese un número par de valores, como, por ejemplo, $X = \{1, 3, 5, 10, 21, 27, 36, 42\}$, entonces tomaríamos como valor mediano la media aritmética de los dos centrales

$$M_e = \frac{10 + 21}{2} = 15,5$$

Definición 2.14 *La* moda (M_d) *es el valor de la variable que tiene máxima frecuencia absoluta.*

La moda no tiene por qué ser única; así, si hay dos modas, la distribución se llama bimodal, si tres, trimodal, etc.

Cuando la variable viene agrupada en intervalos de clase, hablaremos de *intervalo modal.* Este será el intervalo que en su histograma le corresponda el rectángulo de mayor área. En la figura 2.7 se indica la clase modal.

Definición 2.15 *Se definen los* cuartiles *como tres valores de la variable que dividen las observaciones en cuatro partes iguales*[1] *(Figura 2.8)*

Por tanto la mediana es el segundo cuartil.

[1] Dada la definición dada para la mediana, ésta se generaliza a los cuartiles. Por ejemplo, para el primer cuartil, tiene que haber el 25 y 75 % respectivamente en ambos lados de la lista, con el candidato incluido. Montgomery y Runger, (Mexico, 2000)

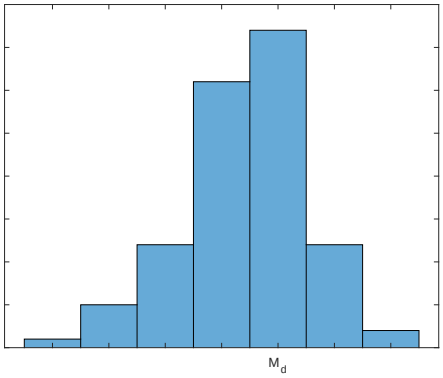


Figura 2.7: Determinación gráfica del intervalo modal.

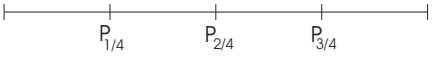


Figura 2.8: Los cuartiles separan los datos en 4 intervalos con el mismo numero de elementos.

2.6 MEDIDAS DE DISPERSIÓN

Las medidas de tendencia central reducen la información de una muestra a un solo valor, pero, en algunos casos, éste estará más próximo a la realidad de las observaciones que en otros. Por ejemplo, consideremos las variables X e Y con sus respectivas frecuencias:

X	n_i
0	1
500	1
100	1

$\bar{x} = \frac{0+500+1000}{3} = 500$

Y	n_i
499	1
501	1

$\bar{y} = \frac{499+501}{2} = 500$

En ambos casos la media aritmética es 500; pero la variable X está mucho más dispersa que la Y, por lo que parece lógico pensar que la representatividad de $\bar{y}$ es mayor que la de $\bar{x}$. Las medidas de dispersión o concentración nos van a cuantificar esa representatividad de los valores centrales.

Definición 2.16 *Se llama* recorrido *o* rango *a la diferencia entre el mayor y el menor de los valores que toma la variable.*

$$R = máx(X) - mín(X)$$

Definición 2.17 *Si* $x_1, x_2, x_3, ..., x_n$ *representa una muestra aleatoria de tamaño* N*, entonces la* varianza de la muestra *se define*[2]

$$S^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} \tag{1}$$

[2] Quesada, Isidoro y Lopez, 2000, Spiegel, (Mexico, 1998)

A veces, ver por ejemplo, Walpole et al., 2012 o Troconiz, 1993, en el denominador de (1), en lugar de N se pone $N-1$, para N grandes esta diferencia es despreciable.

$$S_{N-1}^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1} \tag{2}$$

El motivo de escoger S_{N-1}^2 es que éste es un estimador no sesgado de la varianza de una variable aleatoria. Por otro lado con la definición aquí dada, el método de cálculo de la varianza muestral y la de una variable aleatoria discreta coinciden (ver pág. 47) .

Es evidente que al ser S^2 una suma de cuadrados tomará siempre valores positivos. En el caso en que $S^2 = 0$ entendemos que todos los x_i coinciden con la media $\bar{X}$, es decir, todas las observaciones están concentradas en un mismo punto, por lo que la dispersión es mínima (nula).

Definición 2.18 *Llamaremos* desviación estándar *de la muestra (S) a la raíz cuadrada positiva de la varianza de la muestra*

$$S = +\sqrt{S^2} = \sqrt{\frac{\sum_i (x_i - \bar{x})^2 n_i}{N}}$$

Es importante notar que las unidades de S son las mismas que las de la variable estadística X. En el caso que la muestra X corresponda a una medida experimental se suele usar S o $2S$ como estimador de la incertidumbre en esa medida.

Definición 2.19 *Llamaremos* Coeficiente de variación de Pearson

$$C.V. = \frac{S}{\bar{x}} \qquad \bar{x} \neq 0$$

Este parámetro permite medir la dispersión de X en relación al valor medio de X, lo que en el caso de la interpreatción de X como medida experimental da una idea de la incertidumbre relativa.

Otra medida de dispersión, algo menos usada que la anteriores es

Definición 2.20 Desviación media *o desviación estándar de una muestra*

$$DM = \frac{\sum_{i=1}^{N} |X_i - \bar{x}|}{N} \tag{3}$$

Definición 2.21 Rango Intercuartílico *es la distancia entre el cuartil 3 y el 1 (ver definición 2.15), corresponde a la longitud de la caja del diagrama de caja y extensiones (figura 2.5)*

$$IQR = q_3 - q_1 \tag{4}$$

2.7 OTROS PARÁMETROS ESTADÍSTICOS

2.7.1 *Momentos*

Definición 2.22 *Definimos el* momento de orden r *respecto al parámetro c,a*

$$M_r(c) = \frac{\sum_i (x_i - c)^r n_i}{N}$$

En particular, nos interesarán dos casos:

- *Momentos respecto al origen*
  Cuando $c = 0$ se tienen los momentos respecto al origen

$$a_r = \frac{\sum_i (x_i - 0)^r n_i}{N} = \frac{\sum_i x_i^r n_i}{N}$$

- *Momentos respecto a la media*
  En el caso en que $c = \bar{x}$ tenemos los momentos respecto a la media o momentos centrales

$$m_r = \frac{\sum_i (x_i - \bar{x})^r n_i}{N}$$

  Citaremos algunos momentos particulares

$$\begin{aligned} a_0 &= \frac{\sum_i x_i^0 n_i}{N} = 1 & m_0 &= \frac{\sum_i (x_i - \bar{x})^0 n_i}{N} = 1 \\ a_1 &= \frac{\sum_i x_i n_i}{N} = \bar{x} & m_1 &= \frac{\sum_i (x_i - \bar{x}) n_i}{N} = \frac{0}{1} = 0 \\ a_2 &= \frac{\sum_i x_i^2 n_i}{N} & m_2 &= \frac{\sum_i (x_i - \bar{x})^2 n_i}{N} = S^2 \end{aligned} \tag{5}$$

2.7.2 *Medidas de asimetría y apuntamiento*

Definición 2.23 *Diremos que una distribución de frecuencias es* simétrica *cuando valores de la variable equidistantes de un valor central tienen las mismas frecuencias.*

Es importante destacar en este caso que $\bar{x} = M_d = M_e$. (Ver Figura 2.9)

Definición 2.24 *Se llaman* Distribuciones asimétricas *a aquellas distribuciones que no son simétricas.*

La asimetría puede presentarse a la derecha o a la izquierda.

- *Asimetría a la derecha o positiva*
  Se caracteriza porque la gráfica de las frecuencias presenta cola a la derecha, es

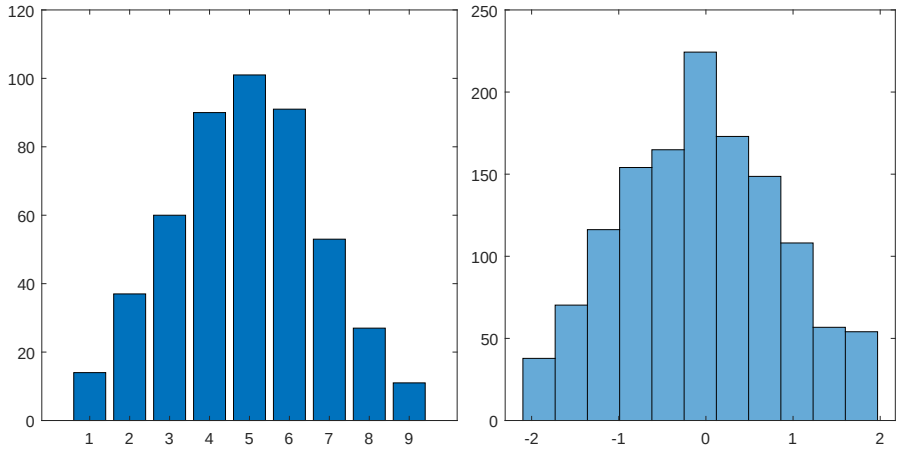


Figura 2.9: Diagrama y histograma de un par de muestras con alta simetría.

decir, éstas descienden más lentamente por la derecha que por la izquierda, como se puede ver en la Figura 2.10. Se verifica en este caso que $\bar{x} \geq M_e \geq M_d$.

- *Asimetría a la izquierda o negativa*
  Se caracteriza porque la gráfica presenta cola a la izquierda, es decir, las frecuencias descienden más lentamente por la izquierda que por la derecha.(ver Figura 2.10) Se verifica que $\bar{x} \leq M_e \leq M_d$.

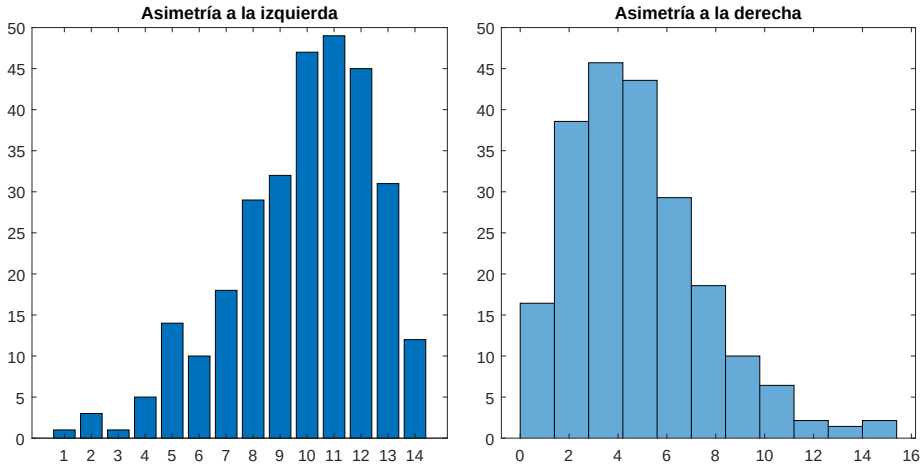


Figura 2.10: Ejemplos de muestras asimétricas, a la izquierda variable discreta, a la derecha variable continua.

Para cuantificar la asimetría observada en una gráfica se pueden definir una serie de coeficientes.

Definición 2.25 *El* Coeficiente de asimetría de Pearson *está definido por*

$$A_p = \frac{\bar{x} - M_d}{\sigma} \begin{cases} A_p > 0 & \textit{Asimetría a la derecha o positiva} \\ A_p = 0 & \textit{Simetría} \\ A_p < 0 & \textit{Asimetría a la izquierda o negativa} \end{cases}$$

Definición 2.26 *El* Coeficiente de asimetría de Fisher *se define de la forma*

$$A_F = \frac{m_3}{\sigma^3} \begin{cases} A_F > 0 & \textit{Asimetría a la derecha o positiva} \\ A_F = 0 & \textit{Simetría} \\ A_F < 0 & \textit{Asimetría a la izquierda o negativa} \end{cases}$$

Por otro lado, si observamos las gráficas correspondientes a tres distribuciones X, Y, Z, (ver Figura 2.11) vemos que, a pesar de que el área encerrada por ellas es la unidad, presentan un "apuntamiento" distinto. En general,

"Bajo apuntamiento $\iff$ Gran aplastamiento."

Para cuantificar esta noción se define la curtosis.

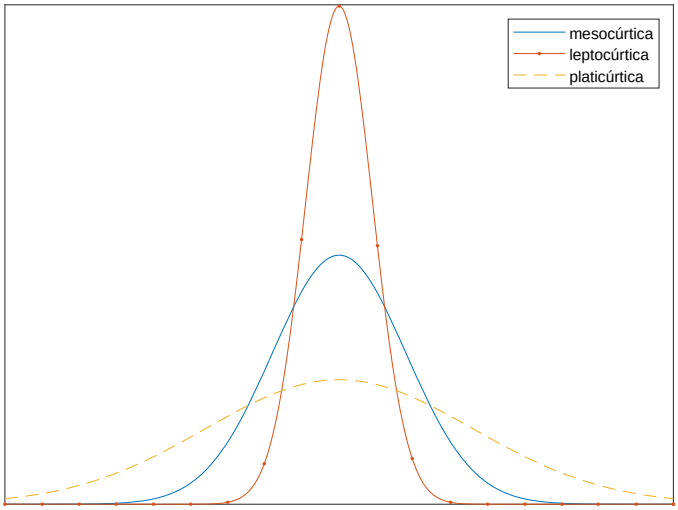


Figura 2.11: Gráficas de funciones con distinta curtosis.

Definición 2.27 Coeficiente de apuntamiento o curtosis *nos indica cuál es el apuntamiento de forma que la distribución comparándola con la normal (o campana de Gauss)*

$$g_2 = \frac{m_4}{\sigma^4}$$

ya que

$$\begin{array}{ll} g_2 > 3 & \textit{Más apuntamiento que la normal: leptocúrtica} \\ g_2 = 3 & \textit{Igual apuntamiento que la normal: mesocúrtica} \\ g_2 < 3 & \textit{Menor apuntamiento que la normal: platicúrtica.} \end{array}$$

EJERCICIOS Y PROBLEMAS

2.1 En la siguiente tabla se presentan la duración, en minutos, de las ultimas interrupciones en el flujo eléctrico de una empresa.

22 18 135 15 90 78 69 98 102
83 55 28 121 120 13 22 124 112
40 6 74 89 103 24 21 112 21
40 98 87 132 115 21 28 43 37
50 96 118 158 74 78 83 93 95

a) Realice un histograma con 7 columnas.

b) Determina la distribución de frecuencias relativas.

c) Realice un diagrama de caja y bigotes.

d) Indique la media, la mediana y la moda.

e) Calcule el rango, la desviación típica, el coeficiente de variación de Pearson y el IQR (Rango Intercuartílico).

2.2 Los siguientes datos representan la duración de la vida, en segundos, de 50 moscas de la fruta que se someten a un nuevo aerosol en un experimento de laboratorio controlado.

17 20 10 9 23 13 12 19 18 24
12 14 6 9 13 6 7 10 13 7
16 18 8 3 3 32 9 7 10 11
13 7 18 7 10 4 27 19 16 8
7 10 5 14 15 10 9 6 7 15

a) Realice un histograma con 8 columnas.

b) Determina la distribución de frecuencias relativas.

c) Realice un diagrama de caja y bigotes.

d) Indique la media, la mediana y la moda.

e) Calcule el rango, la desviación típica, el coeficiente de variación de Pearson y el IQR.

2.3 Use los datos proporcionados con este ejemplar del fichero “motozaleak.xlsx” donde se indica la edad de los motoristas que has sufrido heridas grabes en accidentes de moto.

a) Importe los datos a Matlab.

b) Realice un diagrama de caja y bigotes.

c) Indique la media, la mediana, la moda, la varianza y la desviación típica.

d) ¿Para analizar si la edad es un factor importante a la hora de padecer heridas grabes qué parámetro es más adecuado la media o la mediana?, ¿porque?

2.4 Los datos proporcionados en el fichero "EstadisticaDescriptiva1.xlsx" indican en número de ciclos transcurridos en un ensayo de tracción con probetas de aluminio. El ensayo consistió en someter a un esfuerzo alternante repetido de $21\dot{1}0^3$ psi a 18 Hz.

a) Importe los datos a Matlab.

b) Divida los dato en clases y realice un histograma.

c) Indique la media, la mediana y la moda.

d) Calcule la desviación típica, el coeficiente de variación de Pearson y el IQR.

e) ¿Qué porcentaje de piezas se rompen antes de los 1700 ciclos?

f) ¿Existe alguna evidencia de que una pieza? ¿sobrevivirá más allá de los 2000 ciclos? Justifique la respuesta.

2.5 Cargue los datos de los ficheros "datos1.xlsx" y "datos2.xlsx" y represente lo datos en una misma ventana de Matlab. Comente las diferencias más relevantes entre los conjuntos de datos, fíjese en la simetría de las muestras.

3

TEORÍA DE CONTEO

3.1 ANÁLISIS COMBINATORIO

Como veremos en el capítulo 4, cuando se calculan probabilidades, algunas veces se necesita determinar el número de resultados posibles de un experimento aleatorio. En esta sección se describirán diversos métodos con ese propósito.

La regla básica, que se conoce como principio fundamental de conteo (ver por ejemplo Navidi, 2006):

El principio fundamental del conteo:

Suponga que se pueden realizar k operaciones. Si hay n_1 maneras de realizar la primera operación y si para cada una de esas maneras hay n_2 maneras de realizar la segunda operación y si para cada una de esas elecciones de esas maneras de realizar las dos primeras operaciones hay n_3 maneras de realizar la tercera operación y así sucesivamente, entonces el número total de maneras de realizar la secuencia de las k operaciones es $n_1 n_2 \cdots n_k$.

Usando este principio es fácil demostrar la veracidad de las siguientes definiciones.

3.1.1 *Variaciones*

Definición 3.1 *Dado un conjunto de m elementos distintos, se llaman* variaciones de orden *n de m elementos, a las agrupaciones de n elementos que se puedan formar, conviniendo que dos grupos son distintos si difieren en algún elemento, o si teniendo los mismos difieren en su orden de colocación. El número de variaciones es*

$$V_{m,n} = \frac{m!}{(m-n)!} \tag{6}$$

Ejercicio 3.1.1

En una carrera de 5000 m toman parte 12 corredores. ¿De cuántas formas se pueden repartir las medallas de oro, plata y bronce? **Solución:** $V_{12,3}$

Ejercicio 3.1.2

Con los dígitos 1, 3, 5, 7 y 9:

a) ¿Cuantos números hay de tres cifras?

b) ¿Cuantos son múltiplos de cinco?

Solución: a) $V_{5,3}$, b) como debe terminar con 5, $V_{4,2} + 2 \cdot 4 + 1$.

3.2 VARIACIONES CON REPETICIÓN

Definición 3.2 *Dado un conjunto de m elementos distintos, se llaman* variaciones con repetición de orden *n de m elementos, a las agrupaciones de n elementos que se puedan*

formar, conviniendo en considerar que cada elemento se puede repetir dentro de un grupo y que dos grupos son distintos si difieren en algún elemento, o si teniendo los mismos difieren en su orden de colocación. El número de variaciones con repetición es

$$VR_{m,n} = m^n \tag{7}$$

Ejercicio 3.2.1

¿Cuantos números diferentes de cuatro cifras se pueden obtener con los dígitos 3 y 6? Entre estos ¿Cuántos tienen dos cifras diferentes? ¿Cuantos son mayores que 5000?

Solución: $VR_{2,4}$; $VR_{2,4} - 2$; $VR_{2,3}$

Ejercicio 3.2.2

Con las cifras 3, 5, y 8, ¿Cuántos números de 5 cifras se pueden obtener? ¿Cuántos son más pequeños que 60000? ¿Y cuántos pares?

Solución: $VR_{3,5}$; $2VR_{3,4}$; $VR_{3,4}$

3.3 PERMUTACIONES CON Y SIN REPETICIÓN

Definición 3.3 *Las* permutaciones *son variaciones de n elementos tomadas de n en n*

y en total hay

$$P_n = n! \tag{8}$$

permutaciones.

Las permutaciones sin repetición son por tanto variaciones de n en grupos de n en n, i.e.

$$P_n = V_{n,n} \tag{9}$$

Definición 3.4 *Dado un conjunto de n elementos entre los cuales hay n_i iguales entre si, de modo que $\sum_{i=1}^{p} n_i = n$, se llaman* Permutaciones con repetición *las distintas ordenaciones que se pueden dar, y su número es*

$$P_n^{n_1 n_2 \ldots n_p} = \frac{n!}{\prod_{i=1}^{p} n_i!} \tag{10}$$

Ejercicio 3.3.1

¿De cuántas formas se pueden introducir 6 postales diferentes en 6 sobres de distintos colores, si en cada sobre sólo se puede introducir una postal?

Solución: P_6

Ejercicio 3.3.2

¿Cuántos números distintos se pueden obtener con las cifras del número 111446? ¿Cuántas empiezan por 61?

Solución: $P_6^{3,2,1} = \frac{6!}{3!2'!1!}$; $P_4^{2,2}$

3.4 COMBINACIONES

Definición 3.5 *Dado un conjunto de m elementos distintos, se llaman* combinaciones de orden n *de m elementos, a las agrupaciones de n elementos que se puedan formar, conviniendo que dos grupos son distintos si difieren en algún elemento. El número de combinaciones es*

$$C_{m,n} = \left(\begin{array}{c} m \\ n \end{array} \right) = \frac{m!}{(m-n)!\, n!} \tag{11}$$

Propiedades de los números combinatorios

- $\left(\begin{array}{c} m \\ n \end{array} \right) = \left(\begin{array}{c} m \\ m-n \end{array} \right)$
- $\left(\begin{array}{c} m \\ n \end{array} \right) + \left(\begin{array}{c} m \\ n+1 \end{array} \right) = \left(\begin{array}{c} m+1 \\ n+1 \end{array} \right)$

Ejercicio 3.4.1

Si se unen 5 vértices de un heptágono, ¿cuántos pentágonos se pueden obtener?

Solución: $C_{7,5}$

Ejercicio 3.4.2

¿Cuántas diagonales hay en un hexágono?

Solución: $C_{6,2} - 6$

3.5 COMBINACIONES CON REPETICIÓN

Definición 3.6 *Las* combinaciones con repetición *son agrupaciones de n elementos, iguales o distintos, que difieren uno de otro por contener al menos un elemento distinto.*

En otras palabras, dados m elementos $\{a_1, a_2, a_3, ...\}$ distintos, debemos escoger grupos de n elementos, teniendo en cuenta que un elemento de puede escoger más de una vez.

Para determinar de cuantas formas se puede hacer nos fijamos que podríamos describir cada grupo de la siguiente forma: escribir los m elemento ordenadamente y escribir al lado de cada elemento un asterisco por cada vez aparece en el grupo así el grupo de 5 elementos $\{a_1, a_2, a_2, a_4, a_2\}$ correspondería a

$$a_1 * a_2 * * * a_3 a_4 * a_5 a_6 ...$$

de modo que podemos contar de cuantas maneras distintas podemos ordenar las $a_i's$ y los $*$, como la a_1 siempre ocupa el mismo lugar no sería necesario colocarla de modo que tenemos que ordenar un total de $m - 1 + n$ elementos. Como el orden de las $a_i's$ es irrelevante y el de los * también concluimos que podemos realizar

$$CR_{m,n} = \left(\begin{array}{c} m-1+n \\ n \end{array} \right) = \frac{(m-1+n)!}{(m-1)!\, n!} \tag{12}$$

Dada la argumentación es evidente que

$$CR_m^n = PR_{n,m-1}$$

Ejercicio 3.5.1

¿De cuántas formas se pueden repartir 15 caramelos iguales entre 20 niños, si como máximo damos un caramelo a cada niño? ¿y si se reparten de cualquier forma?

Solución: $C_{20,15}$; $CR_{20,15}$

Ejercicio 3.5.2

¿Cuántos números menores de 1000 son tales que la suma de sus cifras es 9?

Solución: $CR_{2,9}$ es equivalente a repartir 9 caramelos a tres chavales

Esquema resumen

Teniendo en cuenta que hay m elementos que se agrupan de n en n

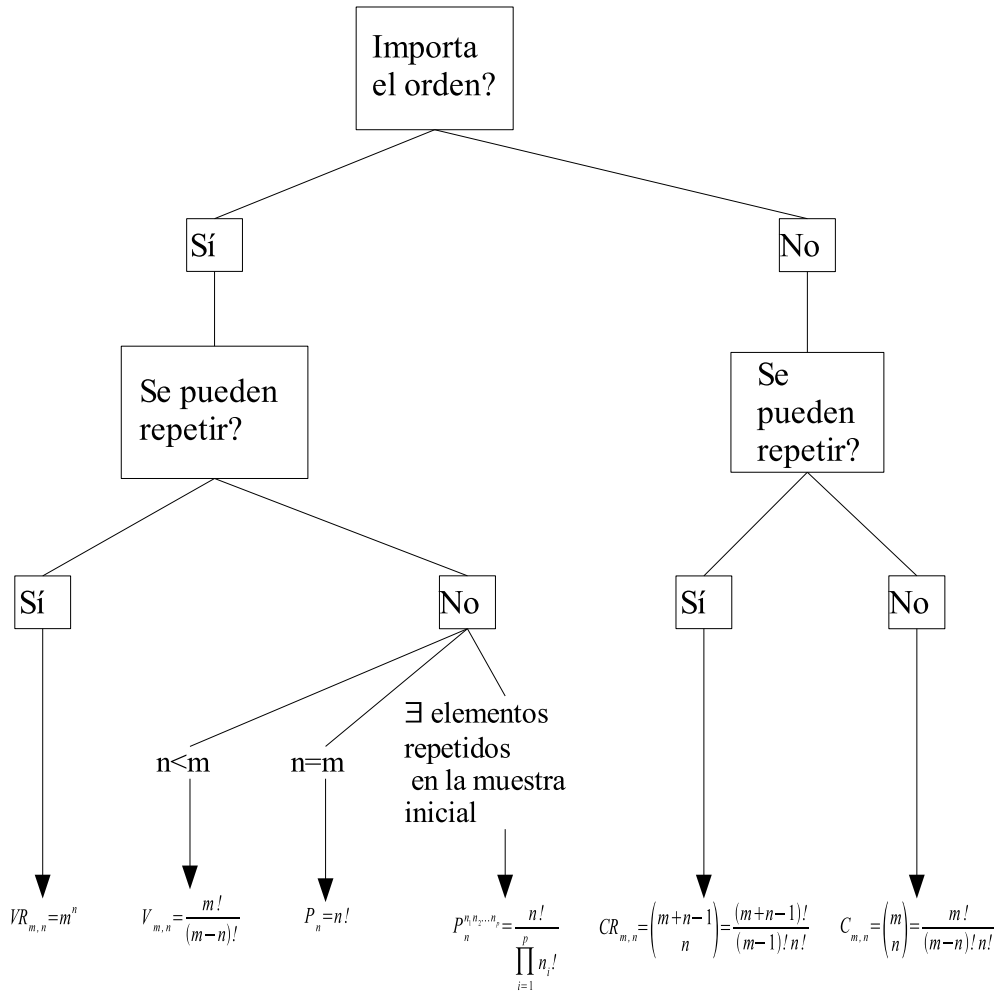


Figura 3.1: Se puede tener en cuenta la idea intuitiva que las combinaciones sirven para escoger elementos (hacer subgrupos), la variaciones para contar subconjuntos y las permutaciones para contar ordenaciones

EJERCICIOS Y PROBLEMAS

3.1 Considere un código binario con 4 bits (0 ó 1) en cada palabra del código. (p.e. 0110)

a) ¿Cuántas palabras distintas hay? $VR_{2,4}$

b) ¿Cuántas con exactamente 2 ceros? $P^{2,2} = C_{2,4}$

c) ¿Cuántas empiezan por cero? $VR_{2,3}$

d) Un código binario de N bits se dice que es de relación constante si en cada palabra hay M de los N bits que son 1 y el resto son 0. ¿Cuántas palabras distintas se pueden hacer con N= 8 y M= 3? $P^{M,N-M} = C_{M,N}$

3.2 Entre los primeros 1000 números, ¿en cuántos es 9 la suma de sus cifras?

SoluciónTengo tres cajas y tengo que rellenarlas con 9 bolas idénticas con lo que el total de números sería: $CR_{3,9} = \begin{pmatrix} 12 \\ 9 \end{pmatrix} = 55$

Otro método sería codificar los números del modo siguiente:

$$441 \rightarrow \ 11110111101$$

de modo que lo que nos piden es el conjunto de números que se pueden formar ordenando 9 unos y 2 ceros $P_{9,2} = 55$

3.3 ¿Cuántos números de cuatro cifras distintas se pueden formar con los números 1,2,3,4,5,6? Calcular la suma de todos ellos.

Solución $V_{6,4} = 360$ cada número ocupa una posición $V_{5,3} = 60$ de modo que $60(1+2+3+4+5+6) = 1260$ y por tanto $1260(1+10+100+1000) = 1399860$

3.4 Hallar la suma de los números enteros de seis cifras que pueden formarse utilizando las cifras 2,5,7 de modo que cada número contenga una vez la cifra 2, tres veces la cifra 5 y dos veces la cifra 7.

Solución En total hay $P_{1,3,2} = 60$ números distintos el número 2 ocupa una posición $P_{2,3} = 10$ veces, el 7 $P_{1,3,1} = 20$ y el 5 $P_{1,2,2} = 30$ por tanto hay $20 \cdot 7 + 10 \cdot 2 + 30 \cdot 5 = 310$ de modo que la suma total es $310(1+10+100+1000+10000) = 34444410$

3.5 Una red rectangular consta de 9×6 nodos, se quiere mandar un mensaje des de el primer nodo (1,1) al último (9,6). ¿De cuántas maneras puede recorrer la red utilizando un recorrido mínimo?

Solución Para contar las posibles rutas basta con codificar el camino, por ejemplo anotando con una A cuando hay que ir hacia arriba y D hacia la derecha, de modo que una ruta seria ($AAAAAAAADDDDDD$) y otra podría ser ($AAAADDDAAADDAD$) de modo que cada ruta es una lista de 8A y 5D, así que habrá $P_{8,5}$ rutas posibles.

Otro modo de codificar una ruta es indicando el número de fila en que está el nodo donde gira a la derecha, así la primera ruta del ejemplo anterior seria (11111) y la segunda (55577) y otra podría ser (99999) así que se trata de $CR_{9,5}$ rutas posibles

4

PROBABILIDAD

4.1 INTRODUCCIÓN. CONCEPTOS BÁSICOS

En probabilidad se llama experimento aleatorio a aquél cuya idéntica repetición genera un conjunto de datos, sean cualitativos o cuantitativos, que puedan ser distintos.

Las leyes de la física clásica son deterministas, pero incluso en este caso, el resultado de una experiencia o experimento puede dar lugar a distintos resultados debido a imprecisiones en el proceso de medida, o a imprecisiones en la determinación del estado inicial. En el caso de la física cuántica se ha visto que sencillamente la naturaleza física del experimento es tal, que el resultado del mismo es aleatorio. Por tanto, el estudio de la leyes de la probabilidad incumbe a la totalidad de los experimentos físicos. Por ejemplo, es un experimento, lanzar una moneda al aire, o preguntar a unos votantes su intención de voto, o medir la resistencia de un conjunto de cables o ...

Definición 4.1 *Dado un experimento aleatorio, al conjunto de todos los resultados posibles mutuamente excluyentes lo llamaremos* espacio muestral Ω

Hay acontecimientos de los que podemos decir si se han producido o no, a estos acontecimientos los llamaremos sucesos aleatorios o sucesos.

Hay sucesos simples (que contienen un solo elemento de Ω), imposibles, seguros, ... De otro modo,

Definición 4.2 *Un* suceso *es un subconjunto de* Ω *vacío o no.*

Dados dos sucesos $A, B \subset \Omega$

- El suceso unión $A \cup B$ es el conjunto de todos los elementos que están en A o en B, osea, $x \in A \cup B$, si y sólo si, $x \in A$ o $x \in B$.
- El suceso intersección $A \cap B$ es el conjunto de todos los elementos que están en A y en B, osea, $x \in A \cap B$, si y sólo si, $x \in A$ y $x \in B$.
- El suceso complementario $\bar{A}$ es el conjunto de todos los elementos que no están en A, osea, $x \in \bar{A}$, si y sólo si, $x \notin A$.
- Se dice que dos sucesos A, B son mutuamente excluyentes o disjuntos o incompatibles, si y sólo si, $A \cap B = \varnothing$

Propiedades

- Distributiva

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$
$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

- Conmutativa

$$(A \cap B) = (B \cap A)$$
$$(A \cup B) = (B \cup A)$$

Teorema 4.1 *Leyes de Morgan*

$$\overline{A \cup B} = \bar{A} \cap \bar{B} \tag{13}$$
$$\overline{A \cap B} = \bar{A} \cup \bar{B} \tag{14}$$

4.2 DEFINICIÓN CLÁSICA DE PROBABILIDAD

La probabilidad mide el grado de certeza de un evento al realizar un experimento aleatorio. En generla este grado de certeza se expresa con un número entre 0 y 1, o en algunos casos entre 0 % y 100 %, siendo 0 el grado de certeza de un suceso imposible y 1 el de un suceso seguro.

En términos más matemáticos, una probabilidad es una aplicación $P : \mathcal{P}(\Omega) \to [0, 1]$ donde $\mathcal{P}(\Omega)$ es el conjunto de las partes de Ω. Para definir esta aplicación hay varios métodos:

- El primero debido, a Laplace (s.XVIII-XIX): La probabilidad asociada a un suceso A es el cociente entre los casos favorables y los posibles, siempre que todos los casos sean equiprobables.
- Definición estadística: $P(A) = \lim_{n\to\infty} \frac{n(a)}{n}$
- Probabilidad geométrica: Dado $B \subset A \Rightarrow P(B) = \frac{S(B)}{S(A)}$

En los siguientes ejercicios se usarán las distintas definiciones para determinar la probabilidad de un suceso, pero las tres definiciones tienen sus pros y sus contras y ninguna es aplicable en todos los casos.

La definición de Laplace es una definición tramposa en el sentido que incluye en la misma la palabra equiprobable, y a priori, no podemos decir si dos resultados tienen la misma probabilidad sin haberla definido antes.

Ejercicio 4.2.1

Usa la definición de Laplace para calcular la probabilidad de los siguientes sucesos:

1. Que al tirar un dado equilibrado el resultado sea par.

   Solución: C.P. = $C_{6,1}$ C.F. $C_{3,1} \Rightarrow P = \frac{1}{2}$

2. Que al tirar dos dados equilibrados
   a) El resultado de la suma sea 2.
   b) El resultado de la suma sea 7.

   Solución: C.P. = $VR_{6,2}$ a) C.F.=1; $P = \frac{1}{36}$ b) C.F.= 6 $P = \frac{6}{36}$

Ejercicio 4.2.2

Usa la definición de Laplace para calcular la probabilidad de los siguientes sucesos: Que en una partida de Mus te toquen tres reyes y un As

a) Si se juega a 4 reyes.

b) Se juega a 8 reyes.

Solución: C.P. $= C_{40,4}$ a) C.F.$= C_{4,3} \cdot C_{4,1}$ $P = \frac{16}{91390}$ b) C.F. $= C_{8,3} \cdot C_{8,1}$

La para aplicar la definición estadística habría que realizar el experimento aleatorio infinitas veces, lo que lo hace inviable, sin embargo, si el experimento se realiza un número "suficiente" de veces se suele asumir que la frecuencia relativa es la probabilidad.

Ejercicio 4.2.3

Realice experimento aleatorio lanzar una moneda al aire varias veces y vaya apuntando la frecuencia relativa del suceso cara. ¿Tiende el resultado a 0.5 que es lo que daría con la definición de Laplace?

Ejercicio 4.2.4

Realice experimento aleatorio lanzar una chincheta al aire varias veces y vaya apuntando la frecuencia relativa del suceso la punta en el aire. ¿tiene sentido en este caso usar la definición de Laplace?

A pesar de que esta definición pueda parecer un poco extraña, como veremos en el capítulo 5, esta definición es mucho más usada de lo que cabe esperar.

Ejercicio 4.2.5

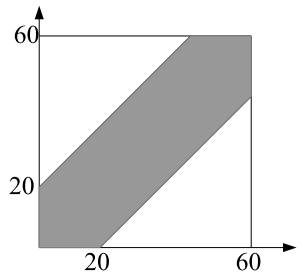
[Problema del encuentro] Dos personas han decidido encontrase entre las 12 y la 13, pero el primero que llegue sólo se va a esperar 20 min. Cual es la probabilidad de encontrarse?

Solución

Si x_i hora de llegada de A_i, enonces $\Rightarrow$ $|x_1 - x_2| \leq 20$

$$0 \geq x_i \geq 60$$
$$x_2 \leq x_1 + 20$$
$$x_2 \geq x_1 - 20$$

$$P(B) = \frac{200}{360} = \frac{5}{9}$$



4.3 DEFINICIÓN AXIOMÁTICA DE LA PROBABILIDAD

Debida a Kolomogorov (1933):

Una familia de conjuntos de Ω, $\mathcal{P}(\Omega)$ finita, infinita numerable o no, es una σ-álgebra, si y sólo si, satisface las propiedades:

i) $\varnothing \in \mathcal{P}(\Omega)$

ii) Si $A \in \mathcal{P}(\Omega) \Rightarrow \bar{A} \in \mathcal{P}(\Omega)$

iii) Si para toda sucesión $A_1, A_2, ...$ numerable de conjuntos de $\mathcal{P}(\Omega)$ $\Rightarrow$

$$\bigcup_i A_i \in \mathcal{P}(\Omega)$$

De *i* y *ii* se demuestra que $\Omega \in \mathcal{P}(\Omega)$. De *ii* y *iii* se demuestra la σ-álgebra es también cerrada bajo una iteración numerable de intersecciones de subconjuntos de $\mathcal{P}(\Omega)$, i.e.,

$$\bigcap_i A_i \in \mathcal{P}(\Omega)$$

Comentario 1 *Si el espacio muestral Ω consta de n elementos entonces $\mathcal{P}(\Omega)$ puede contener a lo sumo 2^n elementos*

DEMOSTRACIÓN:

$$\mathrm{Card}(\mathcal{P}(\Omega)) = \sum_{\mathrm{i=0}}^{\mathrm{n}} \mathrm{C_{n,i}} = (1+1)^{\mathrm{n}}$$

□

Ejercicio 4.3.1

Con el conjunto $\Omega = \{a, e, i, o, u\}$. Compruébese

a) Que el conjunto de partes

$$\mathcal{A} = \{\varnothing,\ (e),\ (a,i),\ (o,u),\ (a,e,i,o,u),\ (a,i,o,u),\ (e,o,u),\ (a,e,i)\}$$

es σ-algebra

b) Que el conjunto de partes

$$\mathcal{B} = \{\varnothing,\ (e),\ (a,i),\ (o,u),\ (a,e,i,o,u),\ (e,o,u),\ (a,e,i)\}$$

no es σ-algebra

Definición 4.3 *Al par $(\Omega, \mathcal{P}(\Omega))$, donde Ω es el conjunto de resultados posibles y $\mathcal{P}(\Omega)$ es una σ-álgebra sobre Ω, se llama* espacio medible *(measurable en inglés), o* espacio probabilizable.

Definición 4.4 *Dada una σ-algebra $\mathcal{P}(\Omega)$, la función $P : \mathcal{P}(\Omega) \to [0,1]$ es una* probabilidad *si cumple*

1 $\forall \omega \in \mathcal{P}(\Omega) \quad P(\omega) \geq 0$

2 $P(\Omega) = 1$

3 Si $A_1, A_2, ...$ es un conjunto de sucesos mutuamente excluyentes dos a dos, entonces

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i)$$

Propiedades:

- $P(\varnothing) = 0$
- $P(\bar{A}) = 1 - P(A)$
- Si $A \subset B \;\Rightarrow\; P(A) \leq P(B)$
- $P(\cup_i A_i) \leq \sum_i P(A_i)$, en particular

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

Como se ha comentado en la sección 4.2, para asignar el valor a P(A) a menudo se sigue el criterio de Laplace[1], o sea, si Ω está formada por n sucesos elementales se asume que son equiprobables de modo que $P(A_i) = \frac{1}{n}$. Otras veces, sin embargo, esta asignación se hace en virtud al conocimiento histórico de los resultados del experimento aleatorio, es decir, según la definición estadística[2].

Otras muchas veces, el criterio que se usa es el geométrico, ya que en definitiva esto es lo que se hace cuando se define función distribución de probabilidad de una variable aleatoria continua (Definición 5.4)

Ejercicio 4.3.2

Se lanza dos veces una moneda, ¿cuál es la probabilidad de que ocurra al menos una cara?

$\Omega = \{cc, cx, xc, xx\}$ que son un total de $VR_{2,2} = 4$ sucesos posibles, éstos son equiprobables con $P = \frac{1}{4}$, de modo que $P\{cc \cup cx \cup xc\} = 3 \cdot \frac{1}{4}$, o de otro modo como de casos favorables hay 3 $P(A) = \frac{3}{4}$

Ejercicio 4.3.3

Las consultas que se reciben en el servicio de atención al cliente de una compañía de telefonía, se pueden clasificar según sean habladas (H) si alguien ha llamado al teléfono de atención al cliente, o de datos (D), en caso de consultas por email. También se pueden clasificar según sean largas(L), si dar una respuesta cuesta más de 3 minutos, o cortas (C). La compañía nos da el siguiente modelo de probabilidades: $P(H) = 0.7$, $P(L) = 0.6$, $P(HL) = 0.35$. Calcular las siguientes probabilidades:

$P(DL); P(D \cup L); P(HC); P(H \cup L); P(H \cup D); P(LC)$

Solución: $0.25, 0.65, 0.35, 0.95, 1, 0$

Ejercicio 4.3.4

En una urna hay 8 bolas negras y 7 bolas blancas.

1. Si extraemos una bola, ¿Cuál es la probabilidad de que sea blanca?
2. Si extraemos dos bolas, ¿Cuál es la probabilidad de que sean blancas?
3. Si extraemos dos bolas, ¿Cuál es la probabilidad de al menos una sea blanca?
4. Si extraemos dos bolas con reposición, ¿Cuál es la probabilidad de que sean blancas?

$\frac{7}{15}, \frac{C_{7,2}}{C_{15,2}}, 1 - P(\bar{A}) = 1 - \frac{C_{8,2}}{C_{15,2}}, \frac{VR_{8,2}}{VR_{15,2}}$

[1] Como en los ejemplos 4.3.2 o 4.3.5

[2] Como en los ejemplos 4.3.3 o 4.3.6

Ejercicio 4.3.5

Se tiran dos dados.

1. ¿Cuál es la probabilidad de que el resultado sean dos unos?
2. ¿Cuál es la probabilidad de que el resultado sean un uno y un dos?
3. ¿Cuál es la probabilidad de que el resultado sea par?
4. ¿Cuál es la probabilidad de que siendo par el resultado, la suma sea 8?

$\frac{1}{VR_{6,2}}, \frac{2}{VR_{6,2}}, \frac{1}{2}, \frac{5}{18}$

Ejercicio 4.3.6

Una empresa desea lanzar un producto nuevo al mercado, pero antes desea que sean evaluados por dos potenciales clientes, con base en datos históricos, y teniendo en cuenta que los dos evalúan de manera independiente completar las probabilidades asociadas a los eventos elementales y calcular la probabilidad de que el segundo cliente apruebe el producto.

Cliente 1	**Cliente 2**	**Probabilidad**
Aprobado	Aprobado	0.04
Aprobado	Modificar	0.16
Modificar	Aprobado	0.16
Modificar	Modificar	?

Solución: : $0.64, 0.20$

Ejercicio 4.3.7

Dos bolsas, contienen 4 bola blancas y 6 negras. Se extraen 3 bolas de cada bolsa, pero de la primera con reposición y de la segunda sin reposición.

Calcular la probabilidad de obtener 2 bolas blancas y una negra en alguna de las bolsas.

$P(2b1n \text{ con reposición} = A) = \frac{PR_{2,1}VR_{4,2}V_{6,1}}{VR_{10,3}} = 3\frac{4}{10}\frac{4}{10}\frac{6}{10}$.

Para el caso sin reposición se puede hacer de distintas formas:

$P(2b1n \text{ sin reposición} = B) = \frac{PR_{2,1}V_{4,2}V_{6,1}}{V_{10,3}} = 3\frac{4}{10}\frac{3}{9}\frac{6}{8} = \frac{C_{4,2}C_{6,1}}{C_{10,3}} = \frac{3}{10}$.

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B) - P(A)P(B) = \frac{627}{1250}$

Ejercicio 4.3.8

De una urna que contiene 2 bolas negras y una blanca se saca una bola con reposición. El juego se detiene si el jugador saca la bola blanca. Si saca negra se introduce otra bola negra en la urna.

Calcular la probabilidad de sacar la bola blanca en el n-ésimo intento.

$P(1N \cap 2N \cap \cdots \cap (n-1)N \cap nB) = \frac{2}{(n+1)(n+2)}$

Ejercicio 4.3.9

Calcular la probabilidad que de un grupo de 10 personas, se dé el suceso A:{al menos 2 cumplan años el mismo día}. ¿Y de un grupo de N personas?

$P(A) = 1 - P(\bar{A})$, $P(\bar{A}) = \frac{V_{365,N}}{VR_{365,N}}$, $N = 10 \Rightarrow P = 0.117 \qquad N = 40 \Rightarrow P = 0.891$

4.4 PROBABILIDAD CONDICIONAL

Definición 4.5 *Se llama* probabilidad *de un suceso* A condicionada *por* B*, suceso de probabilidad no nula, al número*

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{15}$$

Definición 4.6 *Dos sucesos* A *y* B *son* independientes *si* $P(A \cap B) = P(A)P(B)$

Ejercicio 4.4.1

Las consultas que se reciben un operario del servicio de atención al cliente de una compañía de telefonía se analizan en grupos de tres, se pueden clasificar según sean habladas (H) si alguien ha llamado al teléfono de atención al cliente, o de datos (D), en caso de consultas por email. Cada grupo está formado por tres letras (así HHH, corresponde a tres llamadas habladas).

Se comprueba que los grupos HHH y DDD se dan con una probabilidad 0.2, mientras que el resto de posibilidades son equiprobables. Considerar los sucesos que cuenta el número llamadas habladas en cada grupo ($N = i, \quad i = 1 \dots 3$).

Calcular:

$$P(HDH); \quad P(N = 2); \quad P(N \geq 1); \quad P(HHD \mid N = 2)$$
$$P(DDH \mid N = 2); \quad P(N = 2 \mid N \geq 1); \quad P(N \geq 1 \mid N = 2)$$

Solución: $0.1, 0.3, 0.8, \frac{1}{3}, 0, \frac{3}{8}, 1$

Definición 4.7 *Una colección* $\{A_i\}$ *de sucesos no nulos disjuntos, es una* partición *de* Ω *si* $\cup_i A_i = \Omega$.

Teorema 4.2 (Probabilidad total) *Dada una colección* $\{A_i\}$ *de sucesos disjuntos arbitrarios tales que*

$$P(A_i) \neq 0 \qquad y \qquad B \subset \cup_i A_i$$

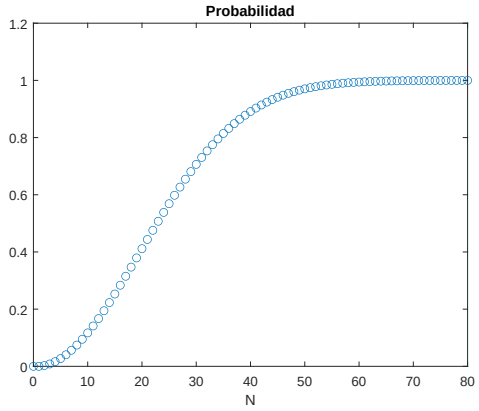


Figura 4.1: Probabilidad de que al menos dos personas de un grupo de N cumplan años el mismo día.

Entonces

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i)P(A_i)$$

Ejercicio 4.4.2

En una cadena de producción se tienen tres máquinas A_1, A_2, A_3 que fabrican un mismo artículo H a un ritmo de 64, 56, 40 por hora respectivamente. El 80 % de A_1, 90 % de A_2 y 96 % de las piezas producidas por A_3 responden satisfactoriamente en un banco de pruebas. Si al término de la jornada ponemos a prueba un artículo elegido al azar ¿cuál es la probabilidad de que responda satisfactoriamente?

Solución: $P(A_1) = 0.4 \quad P(A_2) = 0.35 \quad P(A_3) = 0.25$

$P(S) = \sum_{i=1}^{3} P(S \mid A_i)P(A_i) = 0.875$

Ejercicio 4.4.3

Se tienen dos bolsas V, W conteniendo la bolsa V 5 bolas blancas y 3 bolas negras, mientras que la bolsa W contiene 3 bolas blancas y 4 bolas negras. Se extrae en primer lugar una bola al azar de la bolsa V y tras anotar su color se le introduce en la otra bolsa W. Luego se extrae al azar una segunda bola pero esta vez puede ser de la bolsa V (si la primera bola extraída fue negra). ¿Cuál es la probabilidad de que la segunda bola extraída sea blanca?

Solución: $P(2B) = P(2B \mid 1B)P(1B) + P(2B \mid 1N)P(1N) = \frac{223}{448}$

Ejercicio 4.4.4

Se tienen dos bolsas V, W conteniendo la bolsa V 9 bolas blancas y una negra mientras que la bolsa W contiene 8 bolas blancas. Se efectúan dos operaciones consecutivas consistiendo la primera en tomar al azar 4 bolas de la bolsa V y pasarlas a la W, y la segunda en tomar al azar 4 bolas de la W y pasarlas a la V. ¿Cuál es la probabilidad de que tras la segunda operación la bola negra figure en la bolsa V?

Solución: $P = \frac{11}{15}$

Teorema 4.3 (de Bayes) *Dada una colección* $\{A_i\}$ *de sucesos disjuntos arbitrarios tales que*

$$P(A_i) \neq 0 \qquad y \qquad B \subset \cup_i A_i$$

Entonces

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_{i=1}^{n} P(B \mid A_i)P(A_i)} = \frac{P(B \mid A_i)P(A_i)}{P(B)}$$

Ejercicio 4.4.5

Dado el enunciado del ejercicio 4.4.2 calcular la probabilidad que siendo H defectuoso éste haya sido confeccionado por la máquina A_1.

Solución: 0.64

Ejercicio 4.4.6

En cierta población el 70 % de los habitantes son blancos, el 25 % negros, y el 5 % amarillos. El 70 % de los blancos son católicos, el 60 % de los negros también

y el 10 % de los amarillos también son católicos. Calcular la probabilidad de que:

a) una persona de esa población tomada al azar sea católica.

b) un católico sea blanco.

Solución: a)0.645 b)0.7597

Ejercicio 4.4.7

Una compañía de transporte cubre tres líneas, de forma que el 50 %, 30 % y 20 %, respectivamente de sus camiones trabajan en la línea 1, en la 2 y en la 3. Se sabe que la probabilidad de que un camión esté averiado es del 3 % en la primera línea, del 4 % en la segunda y del 1 % en la tercera. Calcular:

a) La probabilidad de que un día un camión esté averiado.

b) Sabiendo que un camión está averiado, la probabilidad de que sea de la segunda línea.

Solución: a)0.029 b) 0.4138

Ejercicio 4.4.8

Un fabricante realiza una encuesta entre 1000 televidentes con los siguientes resultados: número de los que ven su anuncio 200; número de los que ven su anuncio y compran el producto 50; número de los que compran el producto sin ver el anuncio 20. ¿Cuál es la probabilidad de que una persona que vea el anuncio compre el producto? ¿Cuál es la probabilidad de que una persona que compre el producto haya visto el anuncio?

Solución: $P(v) = \frac{1}{20},\ P(C \mid V) = \frac{1}{4},\ P(C \mid \bar{V}) = \frac{1}{10},\ P(C \mid V) = \frac{1}{4},\ P(V \mid C) = \frac{5}{7}$

Ejercicio 4.4.9

Un tirador con arco suele dar en el centro de la diana 1 de cada 3 flechas. En el próximo campeonato tiene 6 oportunidades para acertar 2 veces y así clasificarse. Cuál es la probabilidad de que se clasifique. Si se ha clasificado, ¿cuál es la probabilidad de que lo haga con la 3ª flecha?

EJERCICIOS Y PROBLEMAS

4.1 Lanzamos dos dados al aire. Calcular las probabilidades de los siguientes sucesos:

A: La suma de las cifras es 7

B: La suma de las cifras es 4 o 5.

C: La suma de las cifras es ≤ 10.

D: Las dos cifras sean iguales

E: La suma de las cifras es numero primo o par.

F: A $\cup$ B

G: B $\cup$ D

Solución: 1/6, 7/36, 11/12, 1/6, 19/36, 13/36, 1/3

4.2 En el juego de la bonoloto se seleccionan 6 números del 1 al 49

a) ¿Cuál es la probabilidad de acertar los 6 números de la bonoloto?

b) ¿Cuál es la probabilidad de acertar 5 de los 6?

c) ¿Qué es más fácil acertar 2, acertar 1 o no acertar ninguno?

Solución: $7.15 \cdot 10^{-8}; 1.84 \cdot 10^{-5}; 0.132; 0.413; 0.436$

4.3 El 12 % de las mujeres que acuden a un centro médico con molestias en los pechos tiene tumor de mama. Uno de los análisis que realizan los médicos para diagnosticarlo da positivo en el 95 % de las mujeres enfermas, pero también en el 5 % de las sanas.

a) Se realiza el análisis a una mujer con dolor de pecho y el resultado da positivo. ¿Cuál es la probabilidad de que tenga un tumor?

b) A otra mujer con dolor de pecho el análisis le da negativo. ¿Cuál es la probabilidad de que no padezca la enfermedad?

Solución: : 0.7215 ; 0.99287

4.4 Le pedimos a una persona que extraiga una bola de una urna que contiene 15 bolas negras y 15 blancas. ¿Cuál es la probabilidad de que saque una blanca?

Las bolas del anterior apartado las repartimos en 3 urnas diferentes. En la primera metemos 2 negras y 4 blancas, en la segunda 3 negras y 6 blancas y en la tercera las bolas que faltan. Después le pedimos a una persona que saque una bola de la urna que quiera. ¿Cuál es la probabilidad de que sea blanca?

Solución: $\frac{1}{2}; 5/9$

4.5 -Lanzando dos dados al aire queremos obtener 8 puntos la primera vez y 9 la segunda. Calcular:

a) La probabilidad de acertar las dos.

b) La probabilidad de acertar alguna de las dos.

Solución: 5/324; 19/81

4.6 En una bolsa tenemos 4 monedas normales y 5 especiales (los dos lados son caras). Sacamos una moneda de la bolsa, la lanzamos 3 veces y obtenemos 3 caras. ¿Cuál es la probabilidad de que la moneda sea especial?

Solución: 10/11

4.7 En una cadena de producción tenemos tres máquinas A, B y C que fabrican 60, 50 y 42 artículos por hora, respectivamente. Sabemos que el 70 % de los artículos producidos por A, el 80 % de B y el 85 % de C son aceptables. Al finalizar un día de trabajo se escoge un artículo al azar. ¿Cuál es la probabilidad de que sea aceptable?

Solución: 0.77434

4.8 Un fabricante realiza una encuesta entre 1000 televidentes con los siguientes resultados: 200 ven su anuncio; 50 ven su anuncio y compran el producto 50; 20 compran el producto sin ver su anuncio 20. ¿Cuál es la probabilidad de que una persona que vea el anuncio compre el producto? ¿Cuál es la probabilidad de que una persona que compre el producto haya visto el anuncio?

Solución: $\frac{1}{4}; 5/7$

4.9 En una determinada carretera pasan 100 turismos y 20 camiones por hora de promedio. Para los próximos 5 vehículos que pasarán, ¿cuánto valen las probabilidades de los siguientes sucesos?

E1: Los 5 sean turismos.

E2: Por lo menos 3 sean turismos.

E3: 3 sean turismos.

E4: $=$ E2 $\cap$ E3

E5: $=$ E1 $\cap$ E3

Solución: 0.4019; 0.9645; 0.16075; 0.16075; 0

4.10 Para realizar una reparación se necesitan dos piezas diferentes, A y B. Las piezas de tipo A se encuentran en una caja, siendo el 10 % de ellas defectuosas, y las de tipo B en otra caja con el 5 % de defectuosas. Si se elige una pieza de cada tipo, ¿cuál es la probabilidad de que aparezca alguna defectuosa?

Solución: 0.145

4.11 Para realizar una reparación se necesitan 5 piezas diferentes, A, B, C, D y E. De cada tipo el 10, 9, 8, 7 y 6 %, respectivamente, son defectuosas. Si se elige una pieza de cada tipo, ¿cuál es la probabilidad de que aparezca alguna defectuosa?

Solución: 0.3413

4.12 Una urna contiene 5 bolas negras y 5 blancas y otra urna 7 negras y 3 blancas.

a) Si sacamos dos bolas de cada urna, con reemplazamiento, ¿cuál es la probabilidad de obtener alguna blanca?

b) Si sacamos dos bolas de cada urna, sin reemplazamiento, ¿cuál es la probabilidad de obtener alguna blanca?

c) Si sacamos una bola de la primera urna y dos de la segunda, sin reemplazamiento, ¿cuál es la probabilidad de obtener una sola blanca?

Solución: 351/400; 121/135;7/5

4.13 Un tirador con arco da en el centro de la diana una de cada cinco flechas. En el próximo campeonato tendrá 5 oportunidades para dar en el centro y así clasificarse para la siguiente fase.

a) ¿Cuál es la probabilidad de que se clasifique para la siguiente fase?

b) En el caso de que se clasifique, ¿cuál es la probabilidad de que lo consiga con la primera flecha?

Solución: 0.6732; 0.297

4.14 Tenemos un lote de 10 piezas, 7 aceptables y 3 defectuosas. Elegimos 4 al azar.

a) ¿Cuál es la probabilidad de que 3 sean defectuosas?

b) ¿Cuál es la probabilidad de que solo haya una defectuosa?

c) ¿Cuál es la probabilidad de que haya menos de tres defectuosas?

d) Si las dos primeras son aceptables, ¿cuál es la probabilidad de que no haya ninguna defectuosa?

Solución: 1/30; ½; 29/30

4.15 En un sorteo de lotería se utilizan 5 bombos. Cada bombo tiene 10 bolas con una cifra del 0 al 9, de modo que el menor número que puede salir es el 00000 y el mayor el 99999.

a) Calcular la probabilidad de que el número premiado tenga las 5 cifras iguales.

b) Calcular la probabilidad de que el número premiado tenga las 5 cifras distintas.

c) La persona A compra un billete que tiene las 5 cifras iguales, la persona B uno con las 5 cifras distintas y la persona C uno con dos cifras iguales y las otras tres diferentes. ¿Quién tiene más posibilidades de llevarse el premio?

Solución: $10^{-4}; 0.3024; 10^{-5}$

4.16 Si elegimos 5 personas al azar, ¿cuál es la probabilidad de que haya alguna coincidencia en sus días de cumpleaños (que dos al menos cumplan los años el mismo día)?

Solución: 0,02714

4.17 El 30 % de las piezas de un paquete y el 70 % de otro son defectuosas. Elegimos un paquete al azar y extraemos 10 piezas con reemplazamiento, obteniendo el siguiente resultado: +-++-+-+-+ (+: aceptable, -: defectuosa). ¿Cuál es la probabilidad de que las piezas hayan sido extraídas del primer paquete?

Solución: 49/58

4.18 Una impresora es capaz de producir 18 impulsos eléctricos diferentes (I1, I2,...I18) relacionados con 18 letras o símbolos gráficos (L1, L2,... L18), de tal modo que el impulso Ik provoca la impresión del símbolo Lk siempre que no haya fallo, o, en caso de fallo, cualquiera de los otros 17 símbolos y todos con la misma probabilidad. La probabilidad de que un impulso falle es 0,12. Se pulsa el mismo impulso dos veces seguidas y la máquina imprime dos veces el símbolo L5. ¿Cuál es la probabilidad de que el impulso haya sido el I5?

Solución: 0.9989073

4.19 Un lote de 40 piezas contiene 8 defectuosas. Si se analizan todas las piezas una a una, calcular la probabilidad del siguiente suceso: "la 13^{a} pieza analizada es la última defectuosa".

Solución: $1.02 \cdot 10^{-5}$

5 VARIABLES ALEATORIAS

5.1 MOTIVACIÓN

En este capítulo y en los siguientes se van a definir conceptos básicos de la teoría de la probabilidad y la estadística, ideas fundamentales como el concepto de variable aleatoria, función densidad de probabilidad, valor esperado o esperanza matemática. Es cierto que estos conceptos requieren de cierta capacidad de abstracción y que para desarrollarlos y comprenderlos correctamente también son necesarios manejar con soltura elementos básicos del cálculo diferencial e integral, así como nociones del cálculo de series numérica.

Estos conceptos son esenciales para el desarrollo de un sinfín de disciplinas científicas, no en vano, es imprescindible manejarlos para poder introducirse en el mundo de la física cuántica, disciplina fundamental para el desarrollo de casi toda la tecnología del siglo XX y XXI, Rovelli, 2021, pero también lo son para poder comprender la base de las nuevas disciplinas en manejo de datos como lo son la inteligencia artificial y en concreto el "deep learning", Mitchell et al., 1997. Por supuesto lo son también para el desarrollo de disciplinas propias de la telecomunicación como lo es la teoría de los procesos estocásticos y candenas de Markov, Ross, 1996. En un plano más ingenieril, se pude decir que es imprescindible conocer los conceptos que se introducirán en este capítulo y las principales distribuciones de probabilidad descritas someramente en los capítulos 6 y 7 para poder calcular y reportar correctamente las incertidumbres de cualquier proceso de medida experimental, Taylor, 1997. En los capítulos 9 y 10 se usan estas mismas nociones para introducirse en el mundo de la toma de decisiones y en el de la comunicación de datos estadísticos, pero esto no es más que un primer paso en el mundo del diseño de experimentos, tan útil en el mundo de la investigación y desarrollo como en el el mundo empresarial, Gutiérrez Pulido y Vara Salazar, 2008.

5.2 VARIABLE ALEATORIA

Cuando se estudian experimentos aleatorios, a menudo, el lugar de interesarnos el resultado obtenido se prefiere obtener un resultado numérico que lo identifique. Por ejemplo, en el caso de un juego de azar, a menudo, en lugar de interesar el resultado del mismo nos interesan las pérdidas o las ganancias obtenidas. Si se estudia tanto estadísticamente como probabilísticamente el experimento aleatorio lanzar dos monedas al aire, puede interesarnos identificar cada uno de los resultados posibles con un número.

Una manera de sistematizar las posibles salidas de un experimento aleatorio consiste definir sobre él una variable aleatoria:

Definición 5.1 *Una función real definida en* Ω

$$\begin{aligned} X: \ \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) = x \end{aligned} \tag{16}$$

es una variable aleatoria[1] *si*

- *El conjunto* $\{X \leq x\}$ *es un evento para cualquier valor de x:*

$$\forall x \in \mathbb{R} \quad \exists \quad A = \{\omega \in \Omega | X(\omega) \leq x\} \in \mathcal{P}(\Omega)$$

- *La probabilidad de los eventos* $\{X = \infty\}$ *y* $\{X = -\infty\}$ *es* 0

La segunda condición no implica que X no pueda tomar los valores ∞ o $-\infty$, si no que sencillamente que la probabilidad que los tome es nula.

Ejemplos

- Un sistema de comunicación por voz de una empresa tiene 48 líneas externas. En un determinado momento, se observa el sistema y algunas están ocupadas. Una posible variable aleatoria sería X que denota el número de líneas que están en uso. Rango=$(0 \ldots 48)$
- Se tira una moneda al aire tres veces, definimos la variable aleatoria X que indica el número de caras. Rango = $(0 \ldots 3)$
- Una variable aleatoria mide el número de ciclos del reloj de un ordenador necesarios para que realice un determinado cálculo aritmético. Rango = $(1 \ldots \infty)$

Las variables aleatorias cuyos únicos valores posibles son 1 y 0 son indicadores de un suceso y a veces reciben el nombre de variables de Bernoulli. Evidentemente sobre un mismo espacio muestral se pueden definir varias variables aleatorias.

5.3 FUNCIÓN DE DISTRIBUCIÓN DE PROBABILIDAD

Como A es un suceso tiene definida una probabilidad, de modo que podemos obtener la probabilidad que una variable tome valores inferiores a x.

Definición 5.2 *Llamaremos* función de distribución de probabilidad *de una variable aleatoria* $X(\omega)$ *a*

$$F_X(x) = P\{X \leq x\} \tag{17}$$

Propiedades:

- $P\{x_1 < X \leq x_2\} = F_X(x_2) - F_X(x_1)$
- $0 \leq F_X(x) \leq 1$
- $F_X(x)$ es monótona creciente.
- $F_X(x)$ es continua por la derecha.
- $$\lim_{x \to \infty} F_X(x) = 1 \qquad \lim_{x \to -\infty} F_X(x) = 0$$

[1] Ver por ejemplo Papoulis, 1991, Peebles, 2001

5.4 FUNCIÓN DENSIDAD DE PROBABILIDAD

Definición 5.3 *Una* variable aleatoria es discreta *si toma valores discretos, o sea, si toma valores* $\{x_1, x_2, ...x_n, ...\}$ *finitos o infinito numerable.*

A cada valor de X se le puede asignar un suceso ω y por lo tanto su correspondiente probabilidad.

Se llama *función densidad de probabilidad*[2]

$$f_X(x_i) = P(X = x_i)$$

De modo que en este caso

$$F_X(x) = \sum_{x_i \leq x} f_X(x_i)$$

Propiedades

- $0 \leq f_X(x_i) \leq 1$
- $\sum_i f_X(x_i) = 1$

A modo de ejemplo consideraremos el siguiente experimento aleatorio. Se tira una moneda trucada tal que $P(c) = \frac{2}{3}$ y $P(+) = \frac{1}{3}$ al aire tres veces. Definimos la variable aleatoria X que indica el número de caras.

a) Encuentra y dibuja la función densidad de probabilidad.

b) Encuentra y dibuja la función distribución de probabilidad.

Está claro que los valores relevantes de la variable aleatoria son $\{0, 1, 2, 3\}$. La probabilidad de cada una de estas posibilidades viene dada por la siguiente expresión

$$P(X = i) = \begin{pmatrix} 3 \\ i \end{pmatrix} \left(\frac{2}{3}\right)^i \left(\frac{1}{3}\right)^{3-i} \qquad i = 0:3$$

donde siguiendo la notación empleado por Matlab, 0:3 significa {0, 1, 2, 3}.

En la figura 5.1, podemos observar la representación gráfica de la función densidad de probabilidad de X y la correspondiente función de distribución de probabilidad.

Se puede apreciar que la función distribución de probabilidad tiene 4 discontinuidades de salto, siendo continua por la derecha en todos los puntos. Además, se puede observar que efectivamente se trata de una función monótona creciente.

Ejercicio 5.4.1

Un aparato electrónico contiene tres componentes que funcionan de manera independiente. La probabilidad de que el primer componente sea defectuoso es de 0.1, 0.2 de que lo sea el segundo y la probabilidad que sea defectuoso el tercero es de 0.1. Sea X el número de componentes defectuosos de uno de estos dispositivos.

[2] Algunos textos, por ejemplo Walpole et al., 2012, llaman a esta función función de masa de probabilidad, y reservan el nombre de función densidad de probabilidad para las variables aleatorias continuas, en este texto se ha optado por usar un único nombre para los dos tipos de variables aleatorias.

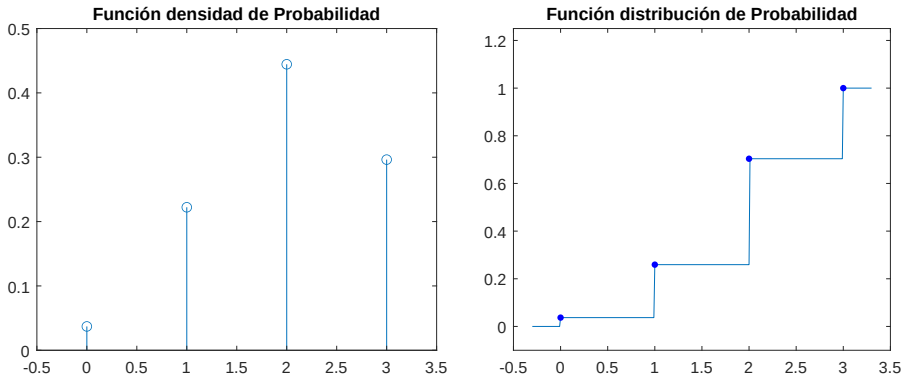


Figura 5.1: Representación gráfica de las funciones densidad y distribución de probabilidad de la variable discreta.

a) ¿Cuáles son los valores posibles de X?
b) Encuentra la función densidad de probabilidad de X.
c) Encuentra la función de distribución de probabilidad de X.
d) ¿Cuál es la probabilidad de que al menos un componente sea defectuoso?
e) ¿Cuál es la probabilidad de que un dispositivo tenga menos de 2 componentes defectuosos?
f) ¿Cuánto es P$(1.2 < X \leq 2.5)$?
g) Encuentra la media y la desviación estándar de X (ver sección 5.6)

solución: a)$X = \{0, 1, 2, 3\}$ b) $P_X(0) = 0.68,\ P_X(1) = 0.306,\ P_X(2) = 0.044,\ P_X(3) = 2 \cdot 10^{-3}$ c) $P(X \geq 1) = 1 - F(0) = 0.352$ d) $F_X(1) = 0.954$ e) $F(2) - F(1) = 0.044$ f) $\bar{x} = 0.4\ \bar{x^2} = 0.5\ \sigma^2 = 0.34$

Ejercicio 5.4.2

Un operador del servicio de atención al cliente de una empresa contesta tanto llamadas (l) como emails (e). Se sabe que la probabilidad contestar correctamente una u otro es la misma (P(l)=P(e)). Nos fijamos en tres respuestas satisfactorias. X denota el número de las que se han dado a llamadas, Y el número de las que se han dado a emails y R = XY.

a) Indica cual es el espacio muestral
b) Indica el Rango de las 3 v.a.
c) dibuja la función densidad de probabilidad y la función distribución de cada una de las variables

solución: a)$\Omega = \{lll, lle, lel, ell, lee, ele, eel, eee\}$; b) $X = 0 \ldots 3, Y = 0 \ldots 3, R = \{0, 2\}$; c) $P_X(0) = \frac{1}{8} = P_Y(3),\ P_X(1) = \frac{3}{8} = P_Y(2),\ P_X(2) = \frac{3}{8} = P_Y(1),\ P_X(3) = \frac{1}{8} = P_Y(0),\ P_R(0) = \frac{2}{8},\ P_R(2) = \frac{6}{8}$

Ejercicio 5.4.3

La variable aleatoria N tiene función densidad

$$f_N(x_i) = \begin{cases} \frac{c}{x_i} & x_i = 1, 2, 3 \\ 0 & \text{en caso contrario} \end{cases}$$

a) Cuanto vale c?

b) $P(N = 1)$

c) $P(N \geq 2)$

d) $P(N > 3)$

Solución: a)$c = \frac{6}{11}$ b) $P(N = 1) = c$ c)$P(N \geq 2) = \frac{5}{11}$ d) $P(N > 3) = 0$

Ejercicio 5.4.4

Una llamada de teléfono ocurre en un momento dado entre $(0, 1)$ min. En este experimento el espacio muestral incluye todos los valores posibles entre 0 y 1, i supondremos que la probabilidad que la llamada se de entre t_1 y t_2 viene dada por

$$P\{t_1 \leq t \leq t_2\} = t_2 - t_1$$

Sea X la v.a. que mide el momento en que se produce la llamada.

a) Calcular F(x), y representarla gráficamente

b) Calcular la probabilidad que la llamada se produzca en el último segundo.

c) calcular que la llamada se produzca exactamente en t = 0,5 min

solución: a)$F(x) = x\ 0 \leq x < 1; \quad F(x) = 0\ x \leq 0; \quad F(x) = 1\ x \geq 1$ b) $P = \frac{1}{60} = F(1) - F(\frac{59}{60})$ c)$P(x = 0,5) = 0 = \lim_{\epsilon \to\ 0} F(0,5 + \epsilon) - F(0,5)$

Definición 5.4 *Se dice que la* variable aleatoria es continua *si lo es también su función de distribución y existe una función $f(x)$, llamada densidad de probabilidad, tal que para todo número real*

$$F_X(x) = \int_{\infty}^{x} f_X(t)dt \tag{18}$$

En particular se dice que $F_X(x)$ es absolutamente continua cuando es derivable, en cuyo caso

$$f_X(x) = \frac{dF_X(x)}{dx}$$

Propiedades

- $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$

Dada la definición 5.2 de la función distribución de probabilidad y la ecuación (18) es evidente que

$$P(a < X \leq b) = \int_{a}^{b} f_X(x)dx \tag{19}$$

Es importante recalcar que, en el caso de variables continuas, la probabilidad de que X tome un valor concreto dentro del dominio de definición de la misma es 0, de modo que, en general, para variables aleatorias continuas los símbolos $\leq$ y $\geq$ se pueden cambiar por $<$ y $>$ con cierta ligereza.

Definición 5.5 *Se llaman* variables aleatorias mixtas[3]*, las variables aleatorias cuya función de distribución es la suma de dos funciones, una discontinua tipo escalones y otra continua.*

Como en el caso de la variables aleatorias discretas, a modo de ejemplo vamos a considerar que el error, en grados, de la temperatura de una reacción química en un experimento es una variable aleatoria continua X con función densidad

$$f_X(x) = \begin{cases} \frac{x^2}{c} & -1 \leq x \leq 2 \\ 0 & \text{en caso contrario} \end{cases}$$

a) Encuentra c.

b) Calcula $F_X(x)$.

c) Calcula $P(0 < X \leq 1)$.

Efectivamente la función dada está definida positiva, de modo que para determinar el valor de c se usa la segunda propiedad:

$$1 = \int_{-\infty}^{\infty} f_X(x)dx = \int_{-1}^{2} \frac{x^2}{c}dx = \frac{3}{c} \;\Rightarrow\; c = 3$$

Para calcular $F_X(x)$ hay que rehacer la integral anterior, pero cambiando los límites, como

$$\int_{-1}^{x} \frac{t^2}{3}dx = \frac{x^3+1}{9}$$

es evidente que

$$F_X(x) = \begin{cases} 0 & x \leq -1 \\ \frac{x^3+1}{9} & -1 \leq x < 2 \\ 1 & x \geq 1 \end{cases}$$

Finalmente, para calcular la probabilidad de que el error en la medida de temperatura sea entre 0 y 1 °C

$$P(0 < X \leq 1) = F(1) - F(0) = \int_0^1 f(x)dx = \frac{1}{9}$$

En la figura 5.2 se muestran las representaciones gráficas de $f_X(x)$ y de $F_X(x)$, así como la probabilidad del apartado c).

Se puede observar que en el caso de una variable continua la función densidad de probabilidad puede tomar valores superiores a 1, pero a pesar de ello, la función $F_X(x)$ sigue estando acotada por 1, y es creciente.

Ejercicio 5.4.5

Calcula la función densidad de probabilidad del ejercicio 5.4.4.

solución: $f(x) = 1\ x \in (0,1) \quad f(x) = 0$ else

[3] Ver por ejemplo Yates y Goodman, 1999, Peebles, 2001

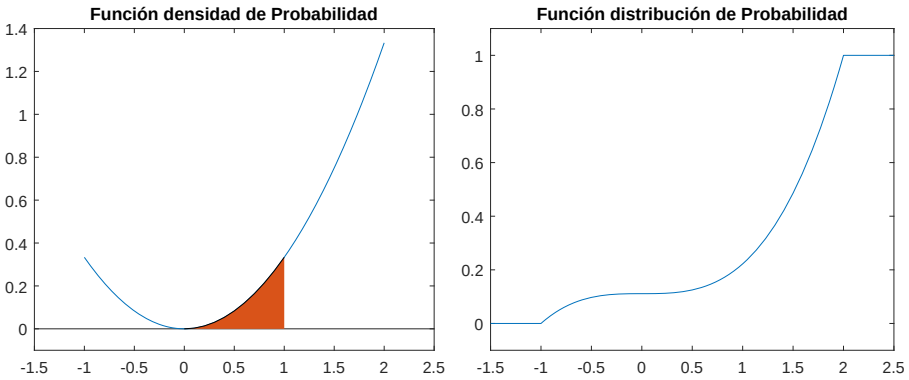


Figura 5.2: Representación gráfica de las funciones densidad y distribución de probabilidad de la variable continua.

Ejercicio 5.4.6

La vida útil, en días, para frascos de cierta medicina es una variable aleatoria que tiene función densidad

$$f(x) = \begin{cases} \frac{20000}{(x+100)^3} & x > 0 \\ 0 & \text{en caso contrario} \end{cases}$$

Encuentre la probabilidad de que un frasco de esta medicina tenga una vida útil de

a) al menos 200 días.

b) cualquier duración entre 50 y 200 días.

c) Calcular la función distribución.

solución: a)$\frac{1}{9}$ b)$\frac{1}{3}$ c) $F(x) = \frac{x(x+200)}{(x+100)^2}$ $x > 0$;

Ejercicio 5.4.7

Sea f_X la función de densidad de una variable aleatoria continua X de recorrido $C = [0.5, b]$, con $b > 0.5$, y $f(x) = \diagdown(2x^2 - 5x - 12)$ para todo $X \in C$, siendo $\diagdown \equiv cte$.

¿Qué condiciones deben de cumplir las constantes $\diagdown$, b?

¿Cuál es la función de distribución de la variable aleatoria X?

solución: a)como $f(x) \geq 0 \Rightarrow b \leq 4$ y $\lambda < 0$, b) $F(x) = \lambda(\frac{2x^3}{3} - \frac{5x^2}{2} - 12x - \frac{157}{24})$

5.5 DISTRIBUCIONES CONDICIONADAS

Teniendo en cuenta la definición 4.5

Definición 5.6 *Dada una v.a. X, la* distribución condicional de X dado B *es*

$$F_X(x|B) := P\{X \leq x|B\} = \frac{P\{(X \leq x) \cap B\}}{P(B)} \qquad P(B) \neq 0$$

Propiedades

- $F_X(-\infty|B) = 0 \qquad F_X(\infty|B) = 1$
- $0 \leq F_X(x|B) \leq 1$
- $F_X(x_1|B) \leq F_X(x_2|B) \quad$ si $\quad x_1 < x_2$
- $P\{x_1 < X \leq x_2|B)\} = F_X(x_2|B) - F_X(x_1|B)$
- $\lim_{x \to\ x_0^+} F_X(x|B) = F_X(x_0|B)$

Si $F(x|B)$ es diferenciable entonces se pude definir la función densidad de probabilidad condicionada

$$f(x|B) = \frac{dF(x|B)}{dx}$$

Ejercicio 5.5.1

En el experimento aleatorio tirar un dado, sea X la v.a. que indica el valor de la tirada.

a) Calcula la función densidad de probabilidad (distribución uniforme discreta).

b) Calcula $F(x|x \leq 2)$.

solución: a)$f(x) = \frac{1}{6}\ x = 1, \ldots, 6 \quad f(x) = 0$ else b)$F(x|x \leq 2) = \frac{P(x \cap x<2)}{F(2)} = \frac{x/6)}{2/6}\ x = 1, 2 \quad F(x|x \leq 2) = 0$ else

5.6 ESPERANZA Y VARIANZA DE UNA V.A.

Definición 5.7 *Sea X una variable aleatoria con función densidad de probabilidad $f_X(x)$ la* esperanza[4] *de X es*

$$\mu_X = E(x) = \sum_x x f_X(x)$$

si X es discreta, y en el caso de que sea continua, como

$$\mu_X = E(x) = \int_{-\infty}^{\infty} x f_X(x) dx$$

Esta definición se puede generalizar:

Definición 5.8 *Se llama valor esperado,* valor medio o esperanza *de una función $g(X)$ a la cantidad*[5]

[4] a menudo nos referimos a la esperanza de X como media, o valor esperado de X

[5] Éste cálculo puede hacerse de otra forma: podría calcularse primero la función densidad de probabilidad de la nueva v.a. $Y = g(X)$ y entonces usar la definición 5.7.

Para ello habría que tener en cuenta que dada una variable aleatoria discreta X con función densidad de probabilidad $f(x)$, la variable Y, es tal que $g(x)$ define una relación biyectiva, (para todo valor y existe un solo valor x tal que $x = w(y)$) y entonces la densidad de probabilidad $(g(y))$ de Y viene dada por

$$g(y) = f(w(y)) \qquad \text{si X discreta}$$

$$g(y) = p(w(y))|J| \qquad \text{si X continua}$$

donde $|J| = \frac{dw}{dy}$ es el jacobiano de la transformación.

$$< g(X) > \equiv E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx \qquad si\ X\ continua \tag{20}$$

$$= \sum_i f(x_i)p(x_i) \qquad si\ X\ discreta \tag{21}$$

Propiedades

El valor esperado es un operador lineal sobre una función de una variable aleatoria, ya que es fácil de comprobar que

- $E(aX + b) = aE(X) + b$

Definición 5.9 *Sea X una v.a. con función densidad de probabilidad $f_X(x)$ y media μ_X, la* varianza *de X es*

$$\sigma_X^2 = E((x - \mu_X)^2) = \sum_x (x - \mu_X)^2 f(x)$$

si X es discreta y

$$\sigma_X^2 = E((x - \mu_X)^2) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)dx$$

para X continua

Propiedades

- $\sigma_X^2 = E(x^2) - \mu_X^2$.
- $\sigma_{aX+b}^2 = a^2\sigma_X^2$

De modo que la varianza no es un operador lineal.

Definición 5.10 *Se llama* desviación estándar *a σ_X*

El siguiente teorema y el correspondiente corolario sirven para ilustrar el significado de la varianza y la desviación estándar como herramienta para medir que tan parecidos son los valores de X al de su media μ_X

Teorema 5.1 (Txebyshev) *Si la varianza existe y es finita, se verifica que para $k > 0$*

$$P(|X - \mu_X| \geq k) \leq \frac{\sigma_X^2}{k^2}$$

Corolario 5.1.1 *La probabilidad que una v.a. tome un valor que se aleje de la media en n veces la desviación típica puede acotarse por $\frac{1}{n^2}$*

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

En cualquier proceso de medida experimental hay una cierta incertidumbre, de ahí que a menudo se usen variables aleatorias para describir el resultado de una media. En este contexto, los ingenieros y físicos suelen usar el valor de la desviación estándar de

la medida como una primera aproximación de la incertidumbre de la medida realizada. Es más en muchos casos, con el afán de ser más exigentes con los resultados mostrados, se suele multiplicar la desviación estándar por un coeficiente, "cover factor" en inglés, típicamente 2 para ser usado como medida de incertidumbre expandida. Para más detalles consultar BIPM, LFCC y IUPAC, 1995

Los conceptos de media y varianza de una variable aleatoria se pueden generalizar.

Definición 5.11 *Se llaman* momentos *de X a los valores esperados de la potencias de X*

En particular la esperanza (μ) es el primer momento de X

Definición 5.12 *Se llaman* momentos centrados *de X a los valores esperados de las potencias de $(X-\mu)$*

En particular la varianza (σ^2)(ver la definición 5.9) es el primer momento centrado de X no nulo.

Ejercicio 5.6.1

Un individuo va a un casino y escoge dos fichas al azar de entre un conjunto de 3 fichas de 1 €, 2 de 10 € y 3 de 20 €. Sea X la v.a. que indica los euros que ha escogido

a) Calcular la función densidad de probabilidad.

b) Calcular la media.

solución: a) Sin Repetición: $f(x) = \frac{3}{28}$ $x = 2, 40$ $f(x) = \frac{12}{28}$ $x = 11, 30$ $f(x) = \frac{1}{28}$ $x = 20$ $f(x) = 0$ en cualquier otro punto. b)$E(x) = 22.78$

Ejercicio 5.6.2

Calcular la vida media del medicamento del ejercicio 5.4.6.

solución: a)$\mu = 100$

Ejercicio 5.6.3

La longitud de ciertos tornillos en centímetros se distribuye según la función de densidad

$$f(x) = \begin{cases} \frac{3}{4}(x-1)(3-x), \text{si } x \in [1,3] \\ 0, \text{en otro caso} \end{cases}$$

a) Calcular la media y la varianza.

b) Si los tornillos son válidos solo cuando su longitud está entre 1'7 y 2'4, calcular la probabilidad de que un tornillo sea válido.

solución: a)$\mu = 2$, $\sigma^2 = \frac{1}{5}$, b)$P(1.7 < X \leq 2.4) = 0.50$

5.7 FUNCIÓN DE UNA VARIABLE ALEATORIA

Con frecuencia se hacen operaciones con varias variables aleatorias, o se definen funciones cuyas variables son variables aleatorias. En general cuando esto pasa, las matemáticas necesarias se escapan un poco del nivel exigido en este texto, por lo que se ha optado por incluirlos de manera muy somera en el capítulo 8, opcional. Sin embargo, en esta sección

se dan algunas nociones y resultados para comprender ejemplos prácticos interesantes de manera más o menos intuitiva. Algunos de los resultados aquí expuestos serán retomados en el capítulo 9.

Empecemos indicando que $Y = g(X)$, es una función de la variable aleatoria X, es ella misma una variable aleatoria.

5.7.1 *Funciones lineales de variables aleatorias*

Un caso particular de función lineal es la función que resulta de sumar a una v. a. una constante. La nueva v. a. será $Y = x + \lambda$. El valor esperado de esta variable es $E(Y) = E(X) + \lambda$ y su varianza es $\sigma_Y^2 = \sigma_X^2$.

Si se multiplica una v.a. por una constante, la v. a. $Y = \lambda X$ tiene media y varianza $\mu_Y = \lambda\mu_x$, $\sigma_Y^2 = \lambda^2\sigma_X^2$ respectivamente.

De modo que en general

$$\begin{aligned} Y &= aX + b && (22)\\ \mu_Y &= a\mu_X + b && (23)\\ \sigma_Y^2 &= a^2\sigma_X^2 \qquad \sigma_Y = |a|\sigma_X && (24) \end{aligned}$$

Ejercicio 5.7.1

La longitud de ciertos tornillos en centímetros viene dada por la variable X de media μ y desviación típica σ. Se define la variable $Z = (X - \mu)/\sigma$. Calcular la media y la varianza de Z.

solución: $\mu = 0$, $\sigma^2 = 1$

5.7.2 *Combinaciones lineales de variables aleatorias*

Si $Y = \sum_i^n c_i X_i$ es una combinación lineal de variables aleatorias entonces se pude demostrar que

$$\begin{aligned} Y &= \sum_i^n c_i X_i && (25)\\ \mu_Y &= \sum_i^n c_i \mu_{X_i} && (26) \end{aligned}$$

si además las n variables aleatorias son *independientes*, entonces

$$\sigma_Y^2 = \sum_i^n c_i^2 \sigma_{X_i}^2 \qquad (27)$$

Es interesante considerar $Z = X - Y$, donde X e Y son variables aleatorias independientes, en este caso

$$\begin{aligned} \mu_{X-Y} &= \mu_X - \mu_Y && (28)\\ \sigma_{X-Y}^2 &= \sigma_X^2 + \sigma_Y^2 && (29) \end{aligned}$$

Otro caso interesante es cuando Y es la suma de dos o más variables aleatorias idénticas e independientes, es decir, $Y = X + X$. En este caso, usando las ecuaciones (26) y (27), es evidente que $\mu_Y = 2\mu_X$ y que $\sigma_Y^2 = 2\sigma_X^2$, sin embargo, no hay que confundir la variable $Y = X + X$ con la variable $Y_2 = 2X$, puesto que, a pesar de que ambas variables tiene la misa media ($2\mu_X$) tienen diferente varianza, pues $\sigma_{Y_2}^2 = 4\sigma_X^2$. O sea, en el caso de las variables aleatorias $X + X \neq 2X$

Ejercicio 5.7.2

La suela de un tipo de zapatos se fabrica pegando cinco capas de distintos colores pero de idénticos espesores. El espesor medio es de 2 mm y la desviación estándar de 0.5 mm.

a) Determine el espesor medio de la suela de los zapatos.

b) Determine la desviación estándar de este espesor.

Solución: $Y = X + X + X + X + X$ por tanto $\mu_Y = 5\mu_X = 10$ mm y $\sigma_Y = \sqrt{\sigma_Y^2} = \sqrt{5}\sigma_X \approx 1.1$ mm)

Ejercicio 5.7.3

Un carnicero decide determinar el precio de los filetes multiplicando por 5 su peso en gramos, de modo que, si un filete pesa 0.5 kg lo vende a 2.5 €. Los filetes tienen un peso medio de 0.5 kg y una desviación estándar de 0.05 kg.

a) Determine el valor medio de un filete.

b) Determine la desviación estándar del mismo.

Solución: $Y = 5X$ por tanto $\mu_Y = 5\mu_X = 2.5$ €y $\sigma_Y = \sqrt{\sigma_Y^2} = 5\sigma_X \approx 0.25$ €)

5.7.3 *Media y varianza de la media de una muestra*

Como se verá con más detalle en la sección 9.2, cuando se toma una muestra aleatoria de una población dada, si esta es suficientemente extensa comparada con la muestra, cada elemento de la muestra se puede tomar una variable aleatoria independiente y de igual distribución. En este caso

$$\bar{X} = \frac{1}{n}\sum_{i}^{n} X_i \tag{30}$$

de modo que usando (26)

$$\mu_{\bar{X}} = \mu_X \tag{31}$$

pero de la ecuación (27),

$$\sigma_{\bar{X}}^2 = \frac{1}{n}\sigma_X^2 \qquad \sigma_{\bar{X}} = \frac{1}{\sqrt{n}}\sigma_X \tag{32}$$

5.7.4 *Error en una medición*

Las ecuaciones (31) y (32) son de vital importancia en teoría de errores y en la interpretación de los resultados de los procesos de medida realizados por lo científicos e ingenieros.

En general cuando se quiere determinar una propiedad física es necesario realizar una medida experimental, sin embargo, el resultada de esta media puede no darnos el resultado exacto de la propiedad física medida.

Definición 5.13 *Se llama* error absoluto de la medida*, al valor absoluto de la diferencia entre el valor medido y el valor real, y* error relativo*, al cociente entre el error absoluto y el valor real.*

En general se dice hay dos fuentes de errores experimentales, lo que se llama error sistemático o de **sesgo**, y el **error aleatorio**.

El primero de ellos representa un error que se puede producir en todas las medidas, normalmente causadas por algún problema en el método de medición y está relacionado con la exactitud de la medida. Por ejemplo, imaginemos que queremos determinar experimentalmente la temperatura de fusión del plomo que es de 327.5 °C, y que el sistema de medida está mal calibrado y sistemáticamente da valores 5 °C por encima del valor real, en este caso la sería poco exacta por tener un sesgo de 5 °C.

El segundo error, el aleatorio, es distinto en cada medición y en promedio tiende a anularse. Este error está ligado a la falta de precisión en la medida, también llamado incertidumbre de la medida.

Evidentemente cuando uno realiza una medida, en general, no sabe el valor de la real que se desea medir, por lo que es muy difícil estimar el sesgo de una medida, normalmente se hace, cuando es posible, utilizando distintos métodos de medida. Sin embargo, el error aleatorio es fácilmente tratable.

Para determinar una propiedad física hay que realizar de forma sistemática varias medidas, para poder estimar el error aleatorio. En general se considera que cada medida es una realización de una variable aleatoria, de modo que la media de la muestra (31) coincide con el valor real si la medida es no sesgada, en el caso de conocerse el valor real,

$$\text{Sesgo} = \mu_{\bar{X}} - \text{Valor real}$$

Definición 5.14 *Por otro lado, la desviación estándar de muestra se usa para medir la incertidumbre de la media, y recibe el nombre de* error estándar*. De ahí que en general se indique el resultado de una medición como*

$$\mu_{\bar{X}} \pm \sigma_{\bar{X}}$$

La ecuación (32) demuestra que la incertidumbre en la medida decrece rápidamente con el número de medidas, de modo que el promedio de varias medidas tiene la misma exactitud que cada una de las medidas realizadas pero es mucho más preciso, pues su incertidumbre se reduce en un factor $\sqrt{n}$.

5.7.5 *Funciones no lineales (opcional)*

La función distribución de la variable aleatoria $Y = g(X)$, es

$$F_Y(y) = P\{\omega \in \Omega | g(X(\omega)) \leq y)$$

En el caso particular que para todo y la ecuación $g(x) = y$ tenga una solución numerable $\{x_i\}$, y siempre que para cada una de estas soluciones $g'(x_i) \neq 0$, la función densidad de probabilidad para Y de puede obtener, ver por ejemplo Proakis y Salehi, 2002, con

$$f_Y(y) = \sum_i \frac{f_X(g^{-1}(y_i))}{|J|} \qquad |J| = \frac{dg^{-1}(y)}{dy} \tag{33}$$

Ejercicio 5.7.4

En teoría de la información se suele definir función $I(X) = -\log P(x)$ como la auto-información del suceso $X = x$. El valor esperado de esta función recibe el nombre de entropía de Shannon.

$$H(x) = E(I(X))$$

Calcular la entropía de Shannon de una variable de Bernoulli de probabilidad p.

solución: $E(I(X)) = (p-1)\log(1-p) - p\log p$

5.8 FUNCIÓN CARACTERÍSTICA DE UNA VARIABLE ALEATORIA (OPCIONAL)

Definición 5.15 *Dada una variable aleatoria X se llama* Función característica[6] *$(\varphi_X(t))$ de X a la función*

$$\varphi_X(t) = E(e^{ixt}) = \begin{cases} \int_{-\infty}^{\infty} e^{ixt} p(x)dx & \text{si } X \text{ continua} \\ \sum_k e^{ix_k t} p(x_k) & \text{si } X \text{ discreta} \end{cases} \tag{34}$$

Ésta función es muy útil debido al teorema

Teorema 5.2 *Dada una v.a. X con función característica $\varphi_X(t)$ entonces*

$$E(X^r) = \frac{1}{i^r} \frac{d^r \varphi_X(t)}{dt^r}\bigg|_{t=0} \tag{35}$$

Propiedades de $\varphi_X(t)$

- La función característica "caracteriza" la variable X en el sentido que dos variables con la misma $\varphi_X(t)$ tienen la misma distribución,
- La función característica de la suma de dos variables aleatorias independientes es el producto de las funciones características.

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$$

- Para cualquier variable aleatoria la función característica existe, es continua y está acotada.

$$|\varphi_X(t)| \leq \varphi_X(0) = 1$$

[6] para el caso continuo, $\varphi_X(t)$ es la transformada de Fourier de $p(x)$.

EJERCICIOS Y PROBLEMAS

5.1 Una máquina fabrica ejes cuyos radios se distribuyen según la siguiente función de densidad: $f_X(x) = k(x-1)(3-x)$ si $1 \leq x \leq 3$ y 0 en el resto de los casos. La variable X se mide en metros. Si el radio está comprendido en el intervalo [1.20 , 2.80] el eje se considera aceptable.

a) Determinar el valor de k.

b) Calcular la longitud media de los ejes fabricados y su desviación estándar.

c) Calcular el tanto por ciento de ejes que serán rechazados.

Solución: a) k = 4/3; b) 2, $\sqrt{5}/5$; c) 5.6 %

5.2 Sea la función

$$f_X(x) = \begin{cases} \alpha \frac{3^\alpha}{x^{\alpha+1}} & x \geq 3 \\ 0 & x < 3 \end{cases}$$

a) ¿Para qué valores de α es función de densidad f(x)?

b) Calcular α para que la media de la variable aleatoria sea 6.

Solución: a) $\alpha > 0$; b) 2

5.3 Cierto tipo de componente está empaquetado en lotes de cuatro. Sea X el número de componentes que funcionan de modo adecuado en un lote elegido de manera aleatoria. Suponga que la probabilidad de que exactamente x componentes funcionen es proporcional a x; en otras palabras, suponga que la función de probabilidad de X es dada por $f_X(x)$ donde k es una constante.

$$f_X(x) = \begin{cases} kx, & x = 1, 2, 3, 4 \\ 0, & \text{en otro caso} \end{cases}$$

a) Determine el valor de la constante k para que $f_X(x)$ sea una función de densidad de probabilidad.

b) Determine $P(X = 2)$.

c) Determine la media del número de componentes que funcionan adecuadamente.

d) Determine la varianza del número de componentes que funcionan adecuadamente.

e) Determine la probabilidad de que 4 funcionen correctamente si más de 2 lo hacen.

Solución: a) $k = \frac{1}{10}$; b)1/5; c) 3; d) 1; $\frac{4}{7}$

5.4 El grosor (en mm) de las tablas de madera producidas en una serrería tiene la siguiente función de probabilidad:

$$f_X(x) = \begin{cases} k(1-(x-5)^2), & 4 \leq x \leq 6 \\ 0, & \text{En otro caso} \end{cases}$$

Calcula:

a) El grosor medio de las tablas

b) La varianza

c) Cogemos 3 tablas de madera de manera aleatoria y se colocan una encima de la otra. ¿Cuál será la media y la varianza de las 3 tablas apiladas?

Solución: a) 4.62; b) 0.059; c) 3*a) 3*b)

5.5 El tiempo (en días) de estancia en la UCI se puede aproximar por una variable aleatoria con función de probabilidad:

$$f_X(x) = \begin{cases} \frac{k}{x}, & x = 1, 2, 3 \\ 0, & \text{en caso contrario} \end{cases}$$

a) Calcular el valor de la constante k.

b) Calcular la probabilidad de estar uno o dos días en la UCI.

c) Calcular el tiempo medio de estancia en la UCI.

d) Calcular la varianza de esta variable aleatoria.

Solución: a) $k = 0.54$; b) 1.64; c) 1.64; d) 0.59

5.6 10 bola negras y 5 blancas están en un recipiente del que se extraen aleatoriamente con reemplazamiento. Determinar las funciones densidad de probabilidad de las siguientes variables aleatorias

a) $X1$: número de bolas blancas de 10 extracciones

b) X_2 número de extracciones hasta que sale la primera bola blanca

c) X_4 número de bolas negras antes de que salga la tercera blanca

d) X_5 número de bola blancas extraídas en 10 extracciones, pero ahora sin reemplazamiento.

Solución: a) $f_{X_1}(x) = \binom{10}{x} \left(\frac{1}{3}\right)^x \left(\frac{2}{3}\right)^{10-x}$; b) $f_{X_2}(x) = \left(\frac{2}{3}\right)^{x-1} \frac{1}{3}$;c) $f_{X_3}(x) = \binom{x+2}{x} \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^3$;d) $f_{X_4}(x) = \dfrac{\binom{10}{10-x}\binom{5}{x}}{\binom{15}{10}}$

5.7 En una caja hay 10 piezas, 4 de ellas defectuosas y el resto buenas. Se realiza el siguiente experimento aleatorio: sacar 5 piezas con reemplazamiento y seguidamente sacar 5 piezas sin reemplazamiento. Sea X la variable aleatoria que mide el número de piezas defectuosas sacadas. Calcular $P(X = 1)$ y $E(X)$

Solución: 0.02469; b) 4

6

DISTRIBUCIONES DE PROBABILIDAD DISCRETA

En el capítulo anterior hemos definido y trabajado con variables aleatorias que describían experimentos aleatorios más o menos reales. En muchos casos el resultado de dichos experimentos se puede modelizar usando variables aleatorias discretas, es más, es posible usar un conjunto relativamente corto de distribuciones discretas de probabilidad para describir una gran variedad de fenómenos aleatorios. En este capítulo se describen las más relevantes.

Por ejemplo, en cualquier experimento en el que se puedan producir un conjunto finito de resultados distintos con idéntica probabilidad se usará una distribución uniforme discreta:

6.1 DISTRIBUCIÓN UNIFORME DISCRETA

Distribución usada cuando en un experimento se pueden producir n resultados distintos igualmente probables, la correspondiente función densidad de probabilidad es

$$p(x,n) = \frac{1}{n}, \qquad x = x_1, x_2, ..., x_n \tag{36}$$

y su media y varianza son,

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n} \tag{37}$$

Ejercicio 6.1.1

Se supone que el tiempo T que un individuo debe esperar el autobús es entre 1 y 20 en minutos, supongamos que T está bien descrito por una distribución discreta uniforme.

a) Calcula la función densidad de probabilidad.

b) Calcula el tiempo medio que debes esperar y la varianza.

c) Calcula la probabilidad de que se espere 15 min o más.

d) Calcula la probabilidad que se espere un tiempo x si se sabe que el autobús tarda más de 8 minutos.

solución: a)$f_T(t) = \frac{1}{20}\ x = 1, 2, \ldots, 20$ b)$\mu = \frac{21}{2}\ \sigma^2 = 33.25$ c) $F(15 \leq x \leq 20) = 0.3$ d) $P(x|t > 8) = \frac{1/20}{12/20}\ x = 9, 10, \ldots, 20$ 0 else

Sin embargo, hay una distribución discreta mucho más usada que la uniforme, se trata de la distribución binomial

6.2 DISTRIBUCIÓN BINOMIAL

Los procesos que consisten en realizar n experimentos independientes, cuyo resultado se clasifica como éxito (con una probabilidad p) o fracaso, recibe el nombre de Proceso de Bernoulli[1]. Nombrado así en honor al matemático suizo Jakob Bernoulli[2]

Figura 6.1: Imagen de Jakob Bernoulli pintada por su hermano Nicolaus en 1687

Definición 6.1 *El número X de éxitos en n experimentos de Bernoulli se denomina variable aleatoria* binomial.

Su función densidad de probabilidad viene dada por

$$B(x;n,p) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x} \tag{38}$$

Con esta función densidad se obtiene

$$\mu = np \qquad \sigma^2 = np(1-p) \tag{39}$$

Ejercicio 6.2.1

En este ejercicio se propone comprobar algunas de las propiedades dadas en la definición precedente.

a) Comprobar que $\sum_x B(x;n,p) = 1$

b) Comprobar que $\mu = np$

c) Comprobar que $\sigma^2 = npq$ con $q = 1-p$

Solución: a)Basta con usar el binomio de Newton $(p+q)^n = \sum_{i=0}^{n} \left(\begin{array}{c} n \\ i \end{array} \right) p^i q^{n-i}$

b) Hay varios métodos, el más sencillo es tener en cuenta que la suma de n variables de Bernoulli independientes es una v binomial, y el hecho que calcular el valor esperado de una variable es un operador lineal con lo cual

$$E(X) = E(\sum X_i) = \sum E(X_i) = \sum p = np.$$

[1] Ver por ejemplo Cao et al., 2001

[2] Ver https://historia-biografia.com/jakob-bernoulli/

Otro método consiste en darse cuenta que

$$E(X) = p\frac{d}{dp}\sum_{i=0}^{n}\left(\begin{array}{c} n \\ i \end{array}\right)p^i q^{n-i} = p\frac{d}{dp}(p+q)^n = np$$

El tercer método es escribir sin más que

$$E(x) = \sum_{i=0}^{n} x\left(\begin{array}{c} n \\ i \end{array}\right)p^i q^{n-i}$$

hacer el cambio de variable $x-1=i$ y usar el binomio de Newton.

c)Se usa cualquiera de los métodos anteriores, en particular, para el primero hay que tener en cuenta que si X_i y X_j son independientes entonces $E(X_iX_j) = E(X_i)E(X_j)$. (ver sección 8.4)

Ejercicio 6.2.2

El 40 % de los automóviles que vende un concesionario están equipados con motor diésel. Obtener la función de densidad y de distribución del número de automóviles diésel entre los siguientes 4 vendidos.

solución: a)$B(x;4,\,0.4)$

Ejercicio 6.2.3

Los doce servidores de una empresa se cuelgan en promedio un día de cada diez.

a) ¿Cuál es la probabilidad de que un cierto día más de 3 servidores estén colgados?

b) ¿Cuál es la media de servidores colgados?

solución: a)$1-\sum_{x=0}^{3} B(x;12,\,0.1) = 0.0256,$ b) $\mu = 1.2$

Ejercicio 6.2.4

Una urna contiene 9 bolas iguales numeradas del 1 al 9 sacamos una bola, apuntamos el número y la devolvemos. ¿Cuantas bolas hay que sacar para que la probabilidad de sacar alguna bola con el 7 sea por lo menos 0,9?

solución: a)$1-P(0) \geq 0.9 \rightarrow\ 1-(\frac{8}{9})^n \geq 0.9 \rightarrow\ n \geq 19.55 \rightarrow\ n = 20$

Hay otras variables aleatorias discretas relacionadas con los procesos de Bernoulli y por tanto con la distribución binomial. En las siguientes subsecciones se describen a modo de ejemplo algunas de ellas, pero en un curso básico de probabilidad y estadística no son imprescindibles, y, además, debido a su simplicidad conceptual, es relativamente sencillo deducir las expresiones de las funciones de densidad de probabilidad correspondientes.

6.2.1 *Distribución multinomial*

Cuando en cada experimento se consideran k resultados posibles de probabilidad p_k entonces tenemos la *distribución multinomial* cuya función densidad es

$$B(x_1,x_2,...x_k;n_1,n_2,...n_k,p_1,p_2,...,p_k) = \frac{n!}{n_1!\,n_2!\,...,n_k!}\prod_{i=1}^{k} p_i^{n_i} \tag{40}$$

Esta función representa la probabilidad de que dados n experimentos el resultado x_i se de n_i veces.

Es obvio que $\sum_1^n x_i = n$ y que $\sum_1^n p_i = 1$

Ejercicio 6.2.5

Se extraen 5 cartas con devolución de una baraja española. Calcular la probabilidad de obtener 2 copas, 2 espadas y 1 oros **solución:** $B(2,2,1,0;\frac{1}{4},\frac{1}{4},\frac{1}{4},\frac{1}{4}) = \frac{5!}{2!\,2!\,1!}\frac{1}{4}^2\frac{1}{4}^2\frac{1}{4}^1\frac{1}{4}^0$

6.2.2 *Distribución hipergeométrica*

La distribución hipergeométrica se aplica cuando hay que hacer una selección aleatoria entre objetos de dos tipos distintos (como hacer un equipo de 8 de entre un grupo de 3 chicas y 12 chicos)

Definición 6.2 *La distribución* hipergeométrica *depende de* 3 *parámetros: el número total de objetos (N), el número de objetos de la primera clase (ej. chicas) (k) y el número de objetos escogidos (n).*

$$h(x;N,n,k) = \frac{\left(\begin{array}{c} k \\ x \end{array}\right)\left(\begin{array}{c} N-k \\ n-x \end{array}\right)}{\left(\begin{array}{c} N \\ n \end{array}\right)} \qquad x = max(0, n+k-N)...min(k,n) \tag{41}$$

En este caso $\mu = np$ y $\sigma^2 = np(1-p)\frac{N-n}{N-1}$ donde $p = \frac{k}{N}$

6.2.3 *Distribuciones binomial negativa y geométrica*

Definición 6.3 *Dado un proceso de Bernoulli, la variable aleatoria que indica que en la x-ésima prueba que se realiza se da el k-ésimo éxito, se llama* binomial negativa

$$B^*(x,k,p) = \left(\begin{array}{c} x-1 \\ k-1 \end{array}\right)p^k(1-p)^{x-k} \qquad x = k, k+1, ... \tag{42}$$

El caso particular en que $k = 1$, se denomina variable geométrica o de Pascal

Definición 6.4 *Dado un proceso de Bernoulli, la distribución que indica la probabilidad que en la x-ésima prueba se dé el primer éxito, se llama* distribución geométrica o de Pascal[3]

$$g(x,p) = p(1-p)^{x-1} \qquad x = 1, 2, ... \tag{43}$$

[3] A veces, Cuadras, (Barcelona, 1990), se cogen $g(x,p) = p(1-p)^x$ para $x = 0, 1, ...$
En este caso, Yates y Goodman, 1999, $\mu = \frac{1-p}{p}$ y $\sigma^2 = \frac{1-p}{p^2}$

Con esta función densidad se obtiene

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \tag{44}$$

Ejercicio 6.2.6

La probabilidad de que un establecimiento de conexión con un servidor de internet vía módem tenga éxito es igual a 0.8. Supóngase que se hacen intentos de conexión hasta que se establecen 3 conexiones.

a) ¿Cuál es la probabilidad de que sean necesarios 6 intentos?

b) ¿Cuál es la probabilidad de que sea necesarios menos de 6?

solución: X =" número de fallos hasta que ocurre el tercer éxito" a)$P(x=3) = B^*(6,3,0.8) = \frac{5!}{3!2!}0,8^3\,0.2^3 = 0.04096$ b)$P(X<3) = \sum_0^2 \frac{(3+x-1)!}{x!2!}(0.8)^3(0.2)^x$

Ejercicio 6.2.7

La variable aleatoria X que mide el número de años adicionales que vive una persona de 70 años es una variable aleatoria geométrica de parámetro $p = 0,9$ si ésta tiene una presión arterial alta (A), y $p = 0.95$ si la tiene normal.

a) Calcula la función densidad de probabilidad $f_{X|A}(x)$ y $f_{X|\bar{A}}(x)$.

b) Calcula la función densidad de probabilidad $f_X(x)$ si se sabe que el 40 % de los mayores de 70 años tienen la presión alta.

solución: a)$f_{X|A}(x) = 0.1(0.9)^{x-1}$ $x = 1,2,3\ldots$ 0 else, $f_{X|\bar{A}}(x) = 0.05(0.95)^{x-1}$ $x = 1,2,3\ldots$ 0 else; b)$f_X(x) = 0.1(0.9)^{x-1}(0.4) + 0.05(0.95)^{x-1}(0.6)$ $x = 1,2,3\ldots$ 0 else

6.3 DISTRIBUCIÓN DE POISSON

Del mismo modo que la distribución binomial está relacionada con un proceso de Bernoulli, hay otro tipo de procesos a lo que se le llama procesos o experimentos que permiten definir otra variable aleatoria discreta de vital importancia, se trata de la variable aleatoria de Poisson, llamada así en honor del matemático francés Siméon Denis Poisson [4] que

proporcionó un modelo para los experimentos que dan valores numéricos a una variable aleatoria X, que representa el número de resultados que ocurren en un determinado intervalo de tiempo o de espacio.

[4] Ver https://mathshistory.st-andrews.ac.uk/Biographies/Poisson/

La distribución de Poisson se ajusta bien a un experimento si cumple las propiedades siguientes[5]:

Si se puede dividir el intervalo en subintervalos suficientemente pequeños de modo que

- La probabilidad que ocurra más de un evento en un subintervalo (región) diferencial es despreciable.
- La probabilidad de que ocurra un evento durante un subintervalo es proporcional a la longitud del mismo (tamaño de la región)
- El número de eventos que ocurren en un intervalo o región es independiente del número que ocurre en cualquier otro intervalo o región disjunto. (Se dice que un proceso de Poisson no tiene memoria)

El número de eventos que ocurren en un experimento de Poisson se llama variable aleatoria de Poisson.

Un ejemplo de variable de Poisson podría ser el número de usuarios que se conectan a un servidor de Internet en un intervalo dado.

Definición 6.5 *La densidad de probabilidad de la variable aleatoria de* Poisson*, que representa el número de sucesos que ocurren en un intervalo dado o región específica t es*

$$p(x, \lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!} \tag{45}$$

Se demuestra que

$$\mu = \lambda t \qquad \sigma^2 = \lambda t \tag{46}$$

Ejercicio 6.3.1

El número de internautas que se conectan a un servidor es de 40 por hora. Calcular la probabilidad que en una hora se conecten 60. Y la probabilidad que en 10 minutos se conecten 10 usuarios

solución: a) $P(X = 60) = e^{-40}40^{60}/60! = 6.79 \cdot 10^{-4}$, b) $\lambda = 40/60 * 10$ P(X =10) = 0.06

Ejercicio 6.3.2

Se supone el número de defectos en los rollos de tela de cierta industria textil es una variable aleatoria Poisson con una media de 0.1 defectos por metro cuadrado

a) ¿Cuál es la probabilidad de tener dos defectos en un metro cuadrado de tela?

b) ¿Cuál es la probabilidad de tener un defecto en 10 metros cuadrados de tela?

c) ¿Cuál es la probabilidad de que no halla defectos en 20 metros cuadrados de tela?

d) ¿Cuál es la probabilidad de que existan al menos dos defectos en 10 metros cuadrados de tela?

solución: a) $4.52 \cdot 10^{-3}$, b) 0.3679, c) 0.1353 d) 0.2642

[5] En Canavos, Medal y Ramírez, 1987, página 126, se demuestra a partir de los supuestos anteriores la fórmula (45).

Ejercicio 6.3.3

Un mecanógrafo comete, en promedio, 2 errores por página. ¿Cuál es la probabilidad de que tenga más de 20 errores en un documento de 7 páginas?

solución: $\lambda = 14$ de modo que $P(X > 20) = 1 - P(X \leq 20) = 0.0479$

Ejercicio 6.3.4

El número de baches en una sección de una carretera interestatal que requieren reparación urgente, pueden modelarse con una distribución Poisson, que tiene una media de dos baches por km.

a) ¿Cuál es la probabilidad de que no haya baches que reparar en un tramo de cinco km?

b) ¿Cuál es la probabilidad de que sea necesario reparar al menos un bache en un tramo de medio km?

c) Si el número de baches está relacionado con la carga vehicular de la carretera y algunas secciones de ésta tienen una carga muy pesada mientras que otras no, ¿qué puede decirse sobre la hipótesis de que el número de baches que es necesario reparar tiene una distribución de Poisson?

solución: a) $4.53 \cdot 10^{-5}$, b) 0.63, c) no

Por un lado, interesa remarcar que

Comentario 2 *Si X_1 y X_2 son variables aleatorias de Poisson con parámetros λ_1 y λ_2 respectivamente, entonces la v.a. $X = X_1 + X_2$ es de Poisson con parámetro $\lambda = \lambda_1 + \lambda_2$*

Por el otro lado, es importante también recalcar, que las dos variables aleatorias discretas más importantes, la distribución binomial y la de Poisson, están relacionadas entre sí por el siguiente teorema:

Teorema 6.1 *Dada X una variable aleatoria binomial, si $n \to \infty$ y $p \to 0$ pero con $\mu = np = \lambda$ finito, entonces*

$$B(x; n, p) \to p(x, np) \tag{47}$$

DEMOSTRACIÓN: Se trata de calcular el límite de (38) cuando $n \to \infty$ pero teniendo en cuenta que λ es finito, para ello se usa la primera relación de (39), de modo que calcularemos

$$\begin{aligned}
\lim_{n\to\infty} & \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \left(\frac{\lambda}{n}\right)\right)^{n-x} = \\
&= \lim_{n\to\infty} \frac{n(n-1)\dots(n-x+1)}{x!} \frac{\lambda^x}{n^x} \frac{\left(1 - \left(\frac{\lambda}{n}\right)\right)^n}{\left(1 - \left(\frac{\lambda}{n}\right)\right)^x} \\
&= \frac{\lambda^x}{x!} \lim_{n\to\infty} \frac{n(n-1)\dots(n-x+1)}{n^x} \frac{\left(1 - \left(\frac{\lambda}{n}\right)\right)^n}{\left(1 - \left(\frac{\lambda}{n}\right)\right)^x} \\
&= \frac{\lambda^x}{x!} e^{-\lambda}
\end{aligned}$$

□

Comentario 3 *En la práctica* [6] *se considera que la aproximación es válida cuando* $n > 30$ *y* $p < 0.1$

Ejercicio 6.3.5

Un proveedor envía unos lotes de 80 productos. Si el defectivo es del 5 %

a) ¿Cuál es la probabilidad de que haya 1 pieza defectuosa en un lote enviado?

b) ¿Cuál es la probabilidad de que haya 2 o más piezas defectuosas? (usar la distribución de Poisson y comparar los resultados)

Si se decide como criterio de aceptación un número máximo de 4 piezas defectuosas por lote:

c) ¿Cuál es el riesgo de aceptar un lote que tuviera un defectivo del 8 %?

d) ¿Cuál es la probabilidad de rechazar un lote con un 5 % de defectuoso?

solución: a) $B(x = 1; 80, 0.05) = 0.069$ con aprox poisson $P(\lambda = 4)(x = 1) = 0.073$ c) $\lambda = np = 6.4 \Rightarrow p(x \leq 4) = 0.24$ d)$\lambda = np = 4 \Rightarrow p(x > 4) = 0.35$

Ejercicio 6.3.6

Cierta compañía aérea vende más billetes que el número de plazas que tiene disponibles, fenómeno llamado "overbooking". Lo hacen por que han detectado que, en promedio, un 1 % de los pasajeros no se presenta. Para un vuelo de 198 plazas se han vendido 200 billetes. ¿Cuál es la probabilidad que todos los pasajeros que se presenten tengan plaza?

solución: Según Poisson con $\lambda = 2$ $P(X \geq 2) = 1 - P(X = 0) + P(X = 1) = 1 - 3e^{-2} = 0.594$. Según binomial $P(X \leq 198) = 1 - P(X > 198) = 1 - (200\,199)0.99^{199}0.01^{1} - (200\,200)0.99^{200} = 0.595$

[6] Ver por ejemplo, Cuadras, (Barcelona, 1990). El criterio aquí sugerido es uno de entre tantos que se puede encontrar en la bibliografía, así por ejemplo en Cao et al., 2001 se sugiere que la aproximación es buena si $p \leq 0.1$ y $n > 50$ o bien si $np < 5$.

EJERCICIOS Y PROBLEMAS

6.1 La media mensual de accidentes laborales en determinada empresa es 0.2.

a) Calcular la probabilidad de que haya 3 accidentes en 1 año.

b) Sabiendo que en abril ha habido algún accidente, calcular la probabilidad de que haya habido menos de 3.

Solución: a)0.2090; b) 0.99366

6.2 El número de defectos superficiales de los paneles de plástico utilizados en los interiores de un tipo de automóvil es de 0.25 defectos por metro cuadrado de panel. Suponga que el interior de un automóvil contiene 5 m2 de este material.

a) ¿Cuál es la probabilidad de que no haya defectos superficiales en los interiores de un automóvil?

b) Si se venden 5 automóviles a una compañía de alquiler de coches, ¿cuál es la probabilidad de que, como máximo, uno de ellos tenga defectos superficiales en el interior?

Solución: a)0.2865; b) 0.02597

6.3 A un puerto llegan 4 barcos al día de media. El puerto tiene dos muelles, en el primero caben 6 barcos, y solo cuando este está completo se empieza a llenar el segundo muelle. Calcular la media de barcos al día del segundo muelle.

Solución: 0.1954

6.4 En una fábrica el 1 % de las piezas procesadas por el método 1 son defectuosos y el 5 % de las procesadas con el método 2 también. El 95 % de las piezas se procesan con el primer método.

a) Se coge una pieza y resulta ser defectuosa. calcular la probabilidad que se haya fabricado con el primer método?

b) Se tienen 10 piezas y de ellas sólo una es defectuosa. ¿Cuál es la probabilidad de que hayan sido procesadas por el primer método?

Solución: a)0.7917; b) 0.8463

6.5 El número medio de meteoritos que cruzan una determinada región atmosférica es de 20 por día. Josetxo vive en esta zona y le gusta salir a mirar el cielo con su telescopio de vez en cuando. Calcula:

a) La probabilidad de que entre las 7:00 y las 9:00 de un día concreto observe más de 2 meteoritos.

b) La probabilidad de que esté más de 4 horas sin observar ningún meteorito.

c) La probabilidad de que en un mes (31 días) se observen más de 550 meteoritos.

Solución: a) 0.2340; b) 0.0357; c) con ayuda de un ordenador 0.9977, otra opción es usar la normal como aproximación, eso se verá en el siguiente capítulo

6.6 Una empresa consta de 20 servidores. La probabilidad de que uno de ellos falle en un día es de 0.1.

a) ¿Cuál es la probabilidad de que en un día fallen exactamente 2 servidores?

b) ¿Cuál es la probabilidad de que en un día el número de servidores que funciona correctamente sea menor que 18?

c) ¿Cuántos servidores fallarán de media al día?

d) ¿Cuál es la desviación estándar del número de servidores que fallan?

e) ¿Cuál es la probabilidad de que en un día fallen 5 o más servidores?

Solución: a) 0.2852; b) 0.3231; c) 2; d) 1.3416, d) 0.0432

6.7 Suponga el número de visitas a una web es de cinco por minuto de media.

a) Determine la probabilidad de que haya 17 visitas en los siguientes tres minutos

b) ¿Qué probabilidad hay que entre las llamadas 5 y 6 pasen más de 20 s?

c) ¿Qué probabilidad hay que en 10 horas se conecten menos 2950 personas?

Solución: a) 0.0847; b) 0.1889; c) 0.1784

6.8 Una sustancia radiactiva emite 3.87 partículas α cada 7.5 s por término medio. Calcular la probabilidad de que se emita al menos una partícula α en un segundo.

Solución: 0.4031

6.9 La media mensual de accidentes laborales en determinada empresa es 0.2.

a) Calcular la probabilidad de que haya 3 accidentes en 1 año.

b) Sabiendo que en abril ha habido algún accidente, calcular la probabilidad de que haya habido menos de 3.

Solución: a) 0.2090, b) 0.99366

6.10 El número de defectos superficiales de los paneles de plástico utilizados en los interiores de un tipo de automóvil es de 0.25 defectos por metro cuadrado de panel. Suponga que el interior de un automóvil contiene 5 m2 de este material.

a) ¿Cuál es la probabilidad de que no haya defectos superficiales en los interiores de un automóvil?

b) Si se venden 5 automóviles a una compañía de alquiler de coches, ¿cuál es la probabilidad de que, como máximo, uno de ellos tenga defectos superficiales en el interior?

Solución: a) 0.2865; b) 0.02597

7

DISTRIBUCIONES DE PROBABILIDAD CONTINUA

En el capítulo anterior hemos visto algunas de las distribuciones de probabilidad discreta más comunes, en este capítulo volveremos a tratar con las distribuciones continuas. Como en el capítulo anterior, la selección de distribuciones tratadas no es exhaustiva, nos vamos centrar en las más comunes: la distribución normal, sin ninguna duda la más utilizada en cualquier ámbito, principalmente debido al Teorema de Limite Central; la exponencial, por su papel de complementación de las variables de Poisson; y la uniforme continua, por su simplicidad.

7.1 DISTRIBUCIÓN UNIFORME CONTINUA

En el caso de una variable aleatoria continua definida en un intervalo dado $[a, b]$, cuando no hay ninguna razón para asignar más probabilidad a ningún punto o subintervalo dado, se opta por usar una función densidad de probabilidad uniforme continua:

Definición 7.1 *La función densidad de una distribución* uniforme continua *es*

$$f(x; a, b) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \textit{en caso contrario} \end{cases} \tag{48}$$

Se demuestra que

$$\mu = \frac{a+b}{2} \qquad \sigma^2 = \frac{(b-a)^2}{12} \tag{49}$$

La probabilidad de que una variable aleatoria uniformemente distribuida se encuentre dentro de algún subintervalo de longitud finita es independiente de la ubicación del intervalo (aunque sí depende del tamaño del intervalo), es decir, probabilidad que X esté

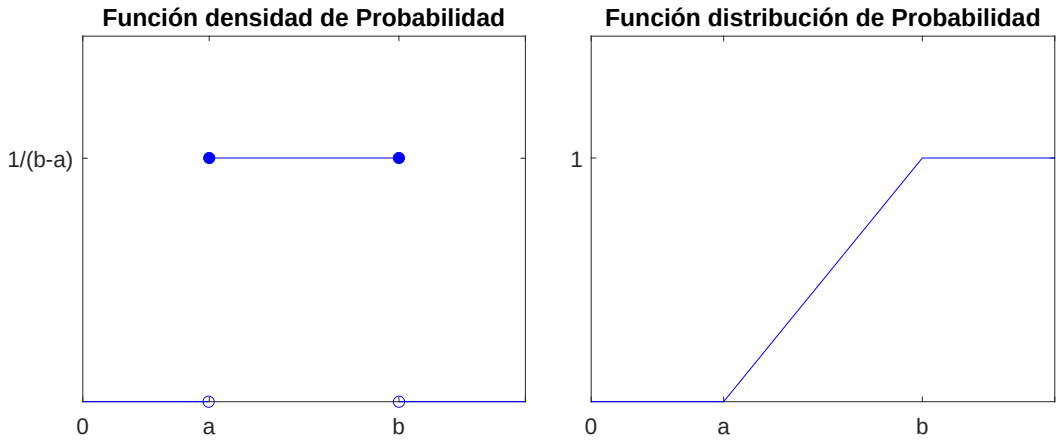


Figura 7.1: Representación gráfica de las funciones asociadas a una variable uniforme continua.

en subintervalo $[c,d] \subset [a,b]$ es proporcional a la longitud de $[c,d]$ pero no de la posición de este en relación a $[a,b]$. De ahí el nombre de distribución uniforme, ver figura 7.1.

Es importante recalcar una vez más, que, en el caso de variables continuas, la probabilidad de que X tome un valor concreto dentro del dominio de definición de la misma es 0, de modo que, en el caso de la variable aleatoria uniforme continua, los símbolos $\leq$ de la ecuación (48) se pueden cambiar por $<$con cierta ligereza.

Ejercicio 7.1.1

Un estudiante acude todos los días a clase andando. suponiendo que tarda en llegar 15 min., que sale en un instante aleatorio entre las 8 : 40 y las 8 : 50 y que la clase empieza a las 9 : 00, calcular la probabilidad de que un día llegue tarde.

solución: $P(45 < X \leq 50) = \int_{45}^{50} \frac{1}{50-40} dx = \frac{1}{2}$

7.2 DISTRIBUCIÓN NORMAL O GAUSSIANA

De todas las variable aleatorias, la más comúnmente usada, es la Normal o Gaussiana en honor al gran matemático y físico alemán Carl Friedrich Gauss cuyo retrato[1] podemos ver en la figura 7.2

Figura 7.2: Retrato de Johan Carl Friedrich Gauss.

Definición 7.2 *La función densidad de una distribución* normal o Gaussiana *de media* μ *y varianza* σ^2*, es*

$$N(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty \tag{50}$$

Con la transformación

$$Z = \frac{X-\mu}{\sigma} \tag{51}$$

se obtiene una *distribución normal estándar o tipificada*, esto es, una distribución normal con media nula y varianza unitaria.

[1] Realizado por Gottlieb Biermann y fotografiado por A. Wittmann, Dominio público

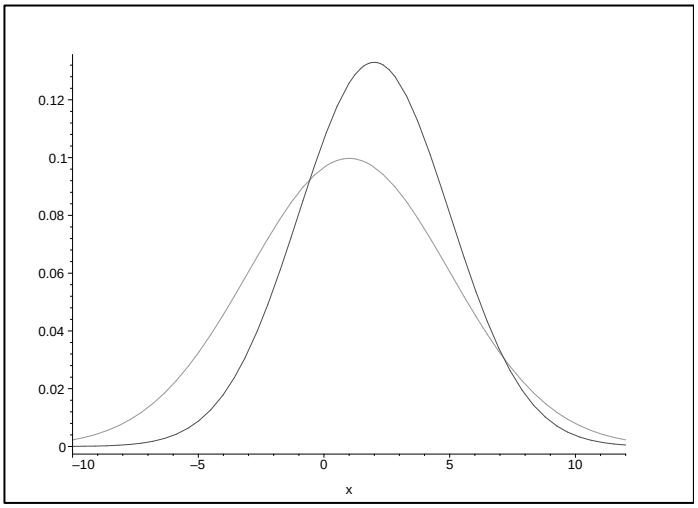


Figura 7.3: Distribuciones normales con distintas medias ($\mu = 1, 2$) y distintas varianzas ($\sigma = 3, 4$).

Como es bien sabido, la función (50) no tiene integral analítica, de modo que para calcular probabilidades[2] como las que se piden en el ejemplo siguiente se tienen que realizar integraciones numéricas. Otra alternativa es el uso de tablas como la tabla del apéndice A, donde se puede ver el resultado de la siguiente integral

$$\Phi(z_0) = \int_{-\infty}^{z_0} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \qquad \text{para} \quad z \geq 0 \tag{52}$$

que se obtiene al tipificar curva normal, es decir al aplicar el cambio de variable dado por (51) a la integral

$$P(X \leq x_0) = \int_{-\infty}^{x_0} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Ejercicio 7.2.1

El coeficiente de inteligencia es una variable aleatoria que sigue una distribución normal ($\mu = 100, \sigma = 16$)

a) Calcular la probabilidad que un individuo tenga un coeficiente inferior a 120.

b) Calcular la probabilidad que un individuo tenga un coeficiente entre 118 y 122

c) Se supone que un un tipo que termina una carrera universitaria tiene un coeficiente superior a 110. ¿Cuál es la probabilidad que sea superior a 120?

solución: a) $P(X < 120) = P(Z < 1.25) = 0.8943$, b) $P(118 < X < 122) = P(1.125 < Z < 1.375) = 0.0457$, c) $P(x > 120|x \geq 110) = \frac{P(x>120 \cap X>110)}{P(X>110)} = 0.3974$

[2] Recordemos que tal como se puede ver en la ecuación (19), para calcular una probabilidad usando una distribución continua, hay que realizar una integral.

7.2.1 *Propiedades de la normal*

- La moda de la normal y la media coinciden.
- La distribución normal es simétrica respecto la media.
- El punto de inflexión de la normal se da en los puntos $x = \mu \pm \sigma$.
- La distribución normal tiene asíntotas horizontales en $y = 0$ para $x = \pm\infty$.
- El área bajo la curva y sobre el eje horizontal es 1.
- La suma de variables aleatorias normales independientes es una variable aleatoria normal de media suma de las medias y varianza suma de las varianzas:

$$S = \sum X_i, \qquad \text{donde } \mu_S = \sum \mu_i \qquad \sigma_S^2 = \sum \sigma_i^2 \tag{53}$$

7.2.2 *Importancia de la Normal. Teorema central del Límite*

La distribución Normal tiene una importancia capital en el campo de la Probabilidad y la estadística, no sólo por el hecho que hay multitud de experimentos cuyo resultado se ajusta a la esta distribución, sino porque además existe el **Teorema central del límite**, TCL que en su versión clásica dice que

Teorema 7.1 (Central del Límite, clásico) *Dada*

$$S_n = \sum_{i=1}^{n} X_i$$

suma de variables aleatorias (X_i) independientes e idénticamente distribuidas de media $E(X_i) = \mu$ y varianza $\sigma^2_{X_i} = \sigma^2$, entonces la función densidad de S_n se aproxima a una distribución normal $N(s; \mu, \sigma)$ con $\mu = n\mu$ y $\sigma^2 = n\sigma^2$.

¿Pero qué tal es esta aproximación? En general se puede decir que la aproximación será mejor cuanto mayor sea el número de variables aleatorias sumadas. Sin embargo, hemos visto que si las variables sumadas son normales el resultado es exacto, Por otro lado, si las variables son muy distintas de una normal, es decir no son simétricas, ni continuas se suele tomar como buena la aproximación para valores de $n \geq 30$. La aproximación puede ser buena para valores $n < 30$ si las variables tienen "pinta" de normal, es decir, si son más o menos simétricas y con forma acampanada.

Existe unas versiones del TCL de Lyapunov y de Lindeberg que todavía suavizan más el enunciado del teorema para asegurar que, bajo condiciones muy generales[3], la suma de variables aleatorias independientes converge a una variable normal, es decir, que ni siguiera es necesario que las variables sumadas sean idénticamente distribuidas:

[3] La demostración y la discusión de las condiciones para la convergencia de la aproximación se escapan de los objetivos de este texto, pero se pueden encontrar en Papoulis y Pillai, 2002, o Masoliver y Wagensberg, 1996 y en sus referencias bibliográficas

Teorema 7.2 (Central del Límite) *Dada*

$$S_n = \sum_{i=1}^{n} X_i$$

suma de variables aleatorias (X_i) independientes de medias $E(X_i) = \mu_i$ y varianza $\sigma^2_{X_i} = \sigma^2_i$, entonces la función densidad de S_n se aproxima a una distribución normal $N(s; \mu, \sigma)$ con $\mu = \sum \mu_i$ y $\sigma^2 = \sum \sigma^2_i$.

Muy a menudo[4], al enunciar el teorema central del límite, en lugar de hacer referencia a la suma de variables aleatorias, se considera la media aritmética[5] de las mismas, de modo que el teorema central del límite, en su versión clásica, sería la siguiente:

Teorema 7.3 (Central de Límite, clásico) *Sea*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

La media de variables aleatorias independientes e idénticamente distribuidas de media μ y desviación típica σ, si n es suficientemente grande

$$\bar{X} \sim N(\mu, \sigma/\sqrt{n})$$

Como se verá con los ejemplos y problemas sugeridos en este capítulo, el teorema central del límite es muy importante en cualquier disciplina científico-tecnológica porque permite aproximar con rigor una gran variedad de situaciones con una distribución bien conocida. Es más, según Rovelli[6], el TCL y el hecho que la desviación estándar de la media disminuya a razón de $\sqrt{n}$ es uno de los motivos por lo que los efectos cuánticos de la naturaleza no son visibles en sistemas macroscópicos.

Un caso particular del teorema central del límite se da cuando las variables aleatorias X_i son variable de Bernuilli, en cuyo caso podemos enunciar el teorema

Teorema 7.4 *Dada X una variable binomial, con $\mu = np$ y $\sigma^2 = npq$ su densidad de probabilidad se puede aproximar por la normal $N(x; \mu, \sigma)$.*

La aproximación será mejor cuanto mayor sea n, pero en la práctica[7], se considera buena para $n > 30$ y $0.1 < p < 0.9$.

Como la suma de variable aleatorias de Poisson es a su vez una variable de Poisson

Teorema 7.5 *Las variables aleatorias de Poisson también se puede aproximar por una Gaussiana, la aproximación será mejor cuanto mayor sea la λ, pero se considera[8] que la aproximación es aceptable si $\lambda > 5$*

[4] Walpole et al., 2012; Montgomery y Runger, (Mexico, 2000); Navidi, 2006.
[5] Ver definición 2.10
[6] Rovelli, 2021.
[7] ver por ejemplo Cao et al., 2001
[8] Ver por ejemplo Cuadras, (Barcelona, 1990)

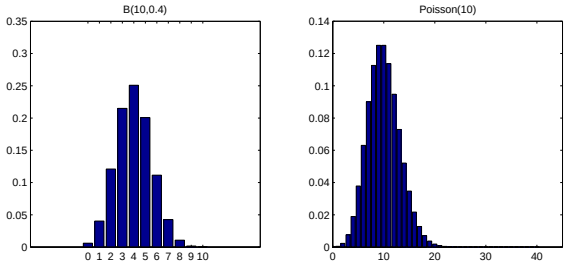


Figura 7.4: Función densidad de probabilidad de una Binomial(10,0.4) y una Poisson(10).

Corrección por continuidad

Cuando se hace uso del teorema central del límite para aproximar una variable discreta, como en el caso de las variables de Poisson o la binomial, con una normal hay que ir con cuidado con los límites de integración, porque el número x de la variable discreta corresponde al segmento $[x-0.5, x+0.5]$ de la variable continua.

Tener en cuenta este hecho como se hace en los ejemplos siguientes en general mejora la aproximación, sin embargo, cuando el área a aproximar es muy pequeña, o incluye zonas donde importa que la variable original sea claramente no simétrica, la esta corrección podría dar resultados algo peores. En este texto se recomienda hacer uso de esta corrección aun a sabiendas que en algún caso el resultado de la aproximación, de no aplicarla, podría ser ligeramente mejor.

Veamos con un ejemplo a qué no referimos:

Ejercicio 7.2.2

En una universidad grande el 25 % de los estudiantes tienen más de 21 años. En una muestra de 400 estudiantes, ¿Cuál es la probabilidad de que más de 110 supere los 21 años?

Solución: X cuenta número de alumnos con más de 21 años, es por tanto una binomial $B(400, 0.25)$, y nos piden $P(X > 110)$, para calcular dicha probabilidad se aproxima X por una normal $Y : N(400 \cdot 0.25, \sqrt{(400 \cdot 0.25 \cdot 0.75)})$ de modo que $P(X > 110) \approx P(Y > 110.5)$ normalizando para poder usar la tabla $P(Z > 1.2124) = 0.1131$, usando Matlab podemos calcular $P(Y > 110.5) = 1 - binocdf(110, 400, 0.25) = 0.1135$

Ejercicio 7.2.3

Sabemos que dos de cada 5 alumnos matriculados en estadística no acudirán al examen. ¿A cuántos alumnos debe convocarse a un aula para 120 personas para poder asegurar que todos se puedan sentar con una probabilidad de al menos 0.975.

Solución: $P(X \leq 120) \geq 0.975$ se trata da aproximar la binomial $B(n, \frac{3}{5})$ de modo que $Y : N(\mu = n\frac{3}{5}, \sigma = \sqrt{n\frac{3}{5}\frac{2}{5}})$

$$P(X \leq 120) \approx P(Y \leq 120.5) = P(Z \leq \frac{120.5 - \mu}{\sigma}) \geq 0.975$$

en la tabla vemos que $Z = 1.96 \Rightarrow n \leq 179$

Ejercicio 7.2.4

Usar la aproximación normal para resolver el ejercicio 6.3.3

Solución: $P(X > 20) = 1 - P(X \leq 20)$ usando la normal $\lambda = 14, \sigma = \sqrt{14}$ de modo que $1 - P(z \leq 1.737) = 0.0412 \cong 0.041$

7.3 DISTRIBUCIÓN EXPONENCIAL

Definición 7.3 *La función densidad de una distribución* exponencial *de parámetro β, es*

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}} & x > 0 \\ 0 & x < 0 \end{cases} \tag{54}$$

Se demuestra que

$$\mu = \beta \qquad \sigma^2 = \beta^2 \tag{55}$$

Proposición 1 *Dado un proceso de Poisson el tiempo (espacio) transcurrido entre un evento y el siguiente viene dado por una distribución exponencial*

DEMOSTRACIÓN: Sea X_t una variable de Poisson que mide el número de eventos que suceden en un tiempo t (espacio x). Entonces se puede definir la variable aleatoria Y que mide el tiempo (espacio) transcurrido entre $t = t_0$ ($x = x_0$) y el primer evento.

La probabilidad de que no se de ningún evento en antes de un tiempo t viene dado por

$$\begin{aligned} P(Y > t) &= P(X_{(t-t_0)} = 0) = e^{-\lambda(t-t_0)} \frac{(\lambda(t-t_0))^0}{0!} \\ &= 1 - P(Y \leq t) = 1 - F_Y(t) \end{aligned}$$

Derivando respecto a t se obtiene

$$p_Y(t) = \lambda e^{-\lambda(t-t_0)}$$

□

La distribución exponencial es un caso particular de una función densidad más general[9] que es la distribución Γ. Dado que, tanto para la distribución exponencial, como para la más general Γ, la variable aleatoria no puede tomar valores negativos, estas distribuciones se suelen usar para modelizar tiempos de espera hasta que se produce un evento, incluso en el caso que éstos no sean eventos de Poisson.

Ejercicio 7.3.1

En una red de ordenadores grande, el acceso de los usuarios al sistema puede modelarse como un proceso Poisson con una media de 25 accesos a la hora

a) ¿Cuál es la probabilidad de que no haya accesos en un intervalo de 6 minutos?

b) ¿Cuál es la probabilidad de que el tiempo que transcurre hasta el siguiente acceso esté entre 2 y 3 minutos?

c) Determinar el intervalo de tiempo para el que la probabilidad de que no se presenten accesos al sistema durante ese tiempo sea de 0,9

[9] Walpole et al., 2012.

d) La desviación estándar del tiempo que transcurre hasta el siguiente acceso.

Ejercicio 7.3.2

La actividad de un fármaco es de 400 días como mínimo. A partir de los 400 días el tiempo que de acción es aleatorio con distribución exponencial de media 25 días. Dado un lote de 12 unidades de este fármaco, calcular la probabilidad que al menos 8 unidades duren más de 430 días.

solución: probabilidad que uno dure más de 30 días $P(X > 30) = 0.3012$, luego es una binomial, $B(n = 12, p = 0.3012)$ de modo que $P(X \geq 8) = \sum_{x=8}^{12} B(12, 0, 3012)(x) = 0.097$

EJERCICIOS Y PROBLEMAS

7.1 Se acepta como distribución aproximada del peso de un tipo de arandelas una normal de media 0.350 gramos y desviación típica 0.050 gramos. Si se seleccionan 10 arandelas al azar, ¿cuál es la probabilidad de que haya más de una arandela que pese entre 0.300g y 0.325g.

Solución: 0.8029

7.2 La duración de la vida de ciertos elementos electrónicos sigue una distribución exponencial con media de 8 meses. Calcular:

a) La probabilidad de que la vida de un elemento este comprendida entre 3 y 12 meses.

b) Si un elemento supera los 10 meses de vida, la probabilidad de que viva más de 25.

Solución: a) 0.4642; b) 0.1534

7.3 Una sustancia radiactiva emite 3.87 partículas α cada 7.5 s por término medio. Calcular la probabilidad de que se emita al menos una partícula α en un segundo.

Solución: 0.4031

7.4 En la observación del número de glóbulos rojos (en millones por mm^3) de los habitantes de una gran ciudad se observó que seguían aproximadamente una distribución normal de media 4.5 y desviación típica 0.5.

a) Si el número de glóbulos de una persona supera la media, ¿cuál es la probabilidad de que tenga más de 5 millones?

b) Si se eligen 10 habitantes al azar, ¿cuál es la probabilidad de que sólo uno o dos de ellos tengan más de 5 millones?

c) Si se eligen 200 habitantes al azar, ¿cuál es la probabilidad de que tengan más de 5 millones de glóbulos el 20 % de ellos o menos?

Solución: a) 0.3173; b) 0.6195; c) 0.9552 (usando la aproximación normal de la binomial) 0.9517 (con la binomial y usando un ordenador)

7.5 Para efectuar una suma de n sumandos, una computadora sustituye cada sumando no entero por el entero más próximo, introduciéndose así por cada sumando un error aleatorio distribuido uniformemente en [-0.5, 0.5]. Si la computadora suma 2500 números, ¿cuál es la probabilidad de que el valor absoluto del error acumulado en la suma sea mayor que 18? Utilizar la distribución normal como aproximación.

Solución: 0.2124

7.6 El tiempo de vida de las bombillas de tipo A sigue una distribución exponencial de media 1000 horas, y el tiempo de vida de las bombillas de tipo B una distribución normal de media 1000 horas.

a) Compramos una bombilla de cada tipo. ¿Cuál es la probabilidad de que alguna de ellas dure más de 1000 horas?

b) Compramos 5 bombillas del tipo A y 5 del tipo B. Sea la variable aleatoria X, el número de bombillas que dura más de 1000 horas. Calcular la media de dicha variable.

Solución: a) 0.6839; b) 4.339

7.7 Sea X una variable normal que se usa como aproximación de una binomial con $P(X > 75) = 0.8944$ y $P(X < 78) = 0.3085$

a) Hallar la media y la desviación típica.

b) Calcular la probabilidad exacta de que haya 80 ensayos exitosos.

Solución: a) $\mu = 80, \sigma = 4$ b) 0.0993

7.8 Un aparato contiene dos piezas que pueden romperse, una rueda y un motor. Para que el aparato funcione ambas piezas son necesarias. La vida de las ruedas sigue una distribución normal N(1000, 200) y la de los motores una normal N(1100, 400). La duración de la rueda y el motor son independientes.

a) Tenemos una muestra de 6 aparatos. ¿Cuál es la probabilidad de que al menos 2 aparatos funcionen más de 1000 horas?

b) Tenemos una muestra de 60 aparatos. ¿Cuál es la probabilidad de que al menos 20 aparatos funcionen más de 1000 horas?

Solución: a) 0.5784; b) 0.3322 usando la aproximación normal y 0.3268 con la binomial y con la ayuda de un ordenador

7.9 En una comunidad autónoma hay un embalse de agua. El volumen de lluvia en miles de metros cúbicos que almacena el embalse al año sigue una distribución $N(2000, 500)$. El consumo de agua en la misma comunidad, medida en miles de metros cúbicos, es otra variable aleatoria, independiente a la anterior, que sigue una distribución $N(1500, 200)$.

a) Calcula la probabilidad de que en un año la cantidad de agua consumida sea mayor a la almacenada en el embalse.

b) ¿Cuál será la probabilidad de que la media de agua consumida en 5 años sea mayor que 1.600.000 m^3?

Solución: a) 0.1766; b) 0.1318

7.10 La producción diaria de hierro en una herrería (en toneladas) sigue una distribución con media 650 t y varianza 820 t^2.

a) Calcula el valor esperado de la producción de 200 días consecutivos.

b) ¿Cuál es la probabilidad de que la producción de hierro de 200 días consecutivos sea mayor que 130.500 t?

c) ¿Cuántas toneladas de hierro se producirán durante 200 días con una probabilidad de 0,95?

Solución: a) $1.3 \cdot 10^5$ toneladas; b) 0.1318; c) $1.3068 \cdot 10^5$ t

8

VARIABLES ALEATORIAS N-DIMENSIONALES

En este capítulo se dan algunas nociones básicas sobre cómo proceder con variables aleatorias n-dimensionales, sin embargo, este tema no es fundamental para poder comprender el siguiente capítulo del libro, por lo que los estudiantes que no estén especialmente interesados en el manejo de este tipo de variables aleatorias, pueden saltarse este capítulo.

Definición 8.1 *Una función definida en* Ω

$$\begin{aligned} \vec{X}: \ \Omega &\rightarrow \mathbb{R}^n \\ \omega &\rightarrow \vec{X}(\omega) = (X_1(\omega), X_2(\omega), \ldots, X_n(\omega)) = \vec{x} \end{aligned}$$

es una variable aleatoria n-dimensional o vector aleatorio *si*

$$\forall \vec{x} \in \mathbb{R}^n \quad \exists \quad A = \{\omega \in \Omega | \vec{X}(\omega) \leq \vec{x}\} \in \mathcal{P}(\Omega)$$

8.1 FUNCIÓN DE DISTRIBUCIÓN N-DIMENSIONAL

Definición 8.2 *Llamaremos* función de distribución n-dimensional *de una variable aleatoria* $X(\omega)$ *a*

$$F_{X_1 X_2 \ldots X_n}(\vec{x}) = P\{X_1(\omega) \leq x_1 \cap X_2(\omega) \leq x_2 \cap \ldots \cap X_n(\omega) \leq x_n \leq x\} \tag{56}$$

Propiedades:

- $F_{\vec{X}}(\vec{x})$ es monótona creciente para cada argumento:

$$F_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_i + n_i, \ldots, x_n) \geq F_{\vec{X}}(x_1, x_2, \ldots, x_i, \ldots, x_n) \quad \forall i = 1..n$$

- $F(x)$ es continua por la derecha:

$$\lim_{n_i \rightarrow \ o^+} F_{\vec{X}}(x_1, x_2, \ldots, x_i + n_i, \ldots, x_n) = F_{\vec{X}}(x_1, x_2, \ldots, x_i, \ldots, x_n) \quad \forall i = 1..n$$

- $$\lim_{\vec{x} \rightarrow \ \vec{\infty}} F_{\vec{X}}(\vec{x}) = 1$$

$$\lim_{x_i \rightarrow \ -\infty} F_{\vec{X}}((x_1, x_2, \ldots, x_i + n_i, \ldots, x_n) = 0 \quad \forall i = 1..n$$

En particular nos van a interesar la variables aleatoria bidimensionales $F_{XY}(x, y)$

Definición 8.3 *Se llama* distribución marginal *de* X *a la distribución de la variable aleatoria univaluada* X *deducida de la distribución conjunta* F_{XY}

$$F_X(x) = \lim_{y \rightarrow \ \infty} F_{XY}(x, y) \tag{57}$$

8.2 FUNCIÓN DENSIDAD DE PROBABILIDAD

Definición 8.4 *Una variable aleatoria* bidimensional *es* discreta *si cada una de la variables X e Y toma valores discretos.*

A cada valor de (x, y) se le puede asignar un suceso ω y por lo tanto su correspondiente probabilidad.

Se llama *función densidad de probabilidad*

$$f(x_i, y_j) = P(X = x_i, Y = y_j)$$

De modo que en este caso

$$F(x, y) = \sum_{x_i \leq x} \sum_{y_j \leq y} f(x_i, y_j)$$

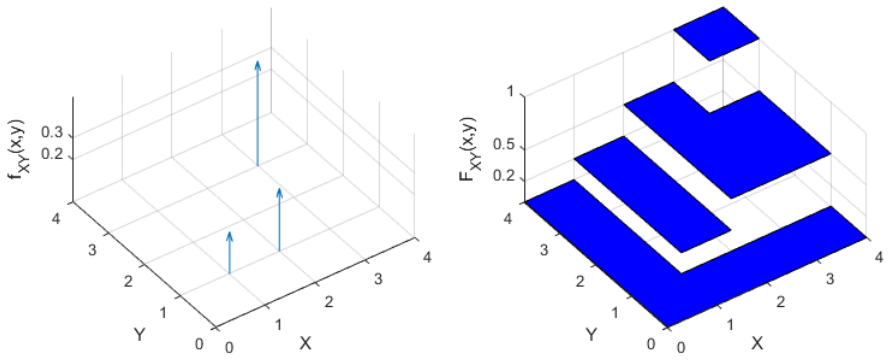
Propiedades

- $f(x_i, y_j) \geq 0$
- $\sum_{i,j} f(x_i, y_j) = 1$

Ejercicio 8.2.1

Supongamos que los únicos valores posibles que puede tomar una variable aleatoria bidimensional son (1,1), (2,1) y (3,3). Cada uno de los eventos correspondientes tiene probabilidad 0.2, 0.3 y 0.5 respectivamente. Encontrar $f_{XY}(x, y)$ y dibujar $F_{XY}(x, y)$

Solución:



Definición 8.5 *Se dice que la variable aleatoria* bidimensional *es* continua *si lo es también su función de distribución y existe una función $f(x, y)$ densidad de probabilidad tal que para todo número real*

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u, v) du dv$$

En particular si $F(x, y)$ es diferenciable

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

Propiedades

- $f(x,y)) \geq 0$
- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$
- $P\{\omega|(x,y) \in G \subset \mathbb{R}^2\} = \int\int_G f(x,y)dxdy$

Ejercicio 8.2.2

Encontrar b para que la función

$$g(x,y) = \begin{cases} be^{-x}\cos y & 0 \leq x \leq 2 \text{ y } 0 \leq y \leq \pi/2 \\ 0 & \text{else} \end{cases}$$

sea una función densidad de probabilidad.

Solución: $\int_0^{\pi/2} dy \int_0^2 dx g(x,y) = 1 \Rightarrow b = 1/(1 - exp(-2))$

Ejercicio 8.2.3

La función densidad de probabilidad conjunta de las v.a. X e Y es

$$g(x,y) = \begin{cases} bxy & 0 \leq x \leq 1,\ 0 \leq y \leq 2 \\ 0 & \text{else} \end{cases}$$

a) ¿Cuánto vale b?

b) ¿Cuál es la probabilidad del suceso A tal que $X^2 + Y^2 \leq 1$?

Solución: a) c = 1 b) $P(A) = \frac{1}{8}$

8.3 FUNCIONES DE DENSIDAD MARGINALES

Definición 8.6 *Se llaman* funciones de densidad marginales *a las funciones*

$$f_X(x_i) = \sum_{y_j} f_{XY}(x_i, y_j) \tag{58}$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy \tag{59}$$

Definición 8.7 *Sean X e Y dos variables aleatorias,* la densidad de probabilidad condicional *de la variable aleatoria Y dado que $X = x$ es*

$$f_Y(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

Teniendo en cuenta la definición 4.6,

Definición 8.8 *Se dice que dos variables aleatorias son estocásticamente* independientes *si*

$$f_{XY}(x,y) = f_X(x)f_Y(y) \tag{60}$$

Ejercicio 8.3.1

La función densidad conjunta de dos variables aleatorias con distribución absolutamente continua es:

$$f(x,y) = \begin{cases} k(x+xy) & 0 < x < 1,\ 0 < y < 1 \\ 0 & \text{en caso contrario} \end{cases}$$

Calcular

a) El valor de k

b) Las funciones de densidad marginales.

c) ¿Son independientes?

solución: a) $1 = \int_0^1 dx \int_0^1 dy k(x+xy) = k\frac{3}{4}$. b) $f_X(x) = \int_0^1 f(x,y)dy = 2x$ si $0 < x < 1$ $f_Y(y) = \frac{2}{3}(1+y)$ si $0 < y < 1$ c) Son independientes puesto que $f(x,y) = f_X f_Y$

Ejercicio 8.3.2

Dada la función densidad conjunta

$$f(x,y) = \theta(x)\theta(y)xe^{-x(y+1)}$$

Donde $\theta(x)$ es la función escalón de Heaviside. Calcular:

a) Las funciones densidad marginales $f_X(x)$ y $f_Y(y)$.

b) Las probabilidades, $P(0 < x < 1)$, $P(0 < y < 1)$, $P(0 < x < 1 \cap 0 < y < 1)$, y comentar el último resultado.

Solución: a) $f_X(x) = \theta(x)e^{-x}$ y $f_Y(y) = \frac{\theta(y)}{(y+1)^2}$; b) $1-e$; $\frac{1}{2}$; $\frac{e^{-2}\,(e-1)^2}{2}$.

Ejercicio 8.3.3

Dada la función densidad de la figura 8.1, donde $P(x_1,y_1) = \frac{2}{15}$, $P(x_2,y_1) = \frac{3}{15}, \ldots$.

a) Calcular $P(y_3)$

b) Dibujar $f_X(x|Y = y_3)$

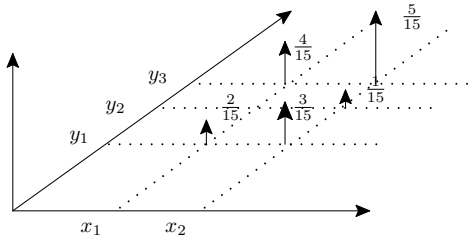


Figura 8.1: Función densidad de probabilidad conjunta de dos variables discretas

Solución: a) $P(y_3) = \frac{4}{15} + \frac{5}{15}$, b) $P(X = x_1|y_3) = \frac{4}{9}$, $P(X = x_2|y_3) = \frac{5}{9}$

Ejercicio 8.3.4

Calcular b y las funciones de densidad marginales si la función densidad conjunta es

$$g(x,y) = \begin{cases} bx & 0 \le x \le 1,\ |y| \le x^2 \\ 0 & \text{else} \end{cases}$$

Solución: $b = 2$, $f_X(x) = \int_{-x^2}^{x^2} 2x\,dy = 4x^3$ para $0 \leq x \leq 1$ y $f_Y(y) = \int_{\sqrt{|y|}}^{1} 2x\,dx = 1 - |y|$ para $-1 \leq y \leq 1$.

8.4 VARIANZA Y COVARIANZA

Cuando dos variables no son independientes, es interesante medir el grado de dependencia entre ellas, de ahí la importancia de los conceptos de covarianza y correlación.

Definición 8.9 *Sean X e Y dos variables aleatorias, la densidad de probabilidad conjunta $f_{XY}(x, y)$. El* valor esperado de la función $g(x, y)$ *es*[1]

$$E(g(x,y)) = \sum_x \sum_y g(x,y) f_{XY}(x,y) \tag{61}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y) dydx \tag{62}$$

En particular

Definición 8.10 *Se llama* correlación *de dos v.a. X e Y al momento de segundo orden m_{11}*

$$R_{X,Y} \equiv E[XY]$$

Definición 8.11 *Se llama* covarianza *de dos v.a. X e Y a*

$$Cov[X,Y] \equiv \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

Son fáciles de demostrar las siguientes
Propiedades

- $Cov[X,Y] = E[(XY)] - E[X]E[Y]$
- $Cov[X,X] = \sigma_X^2$
- $Cov[aX,Y] = aCov[X,Y]$
- $\sigma_{X\pm Y}^2 = \sigma_X^2 + \sigma_Y^2 \pm 2Cov[X,Y]$

Comentario 4 *En el caso particular que X e Y sean* ***independientes*** *$Cov(X,Y) = 0$. De modo que además de que*

$$E(aX + bY) = aE(x) + bE(Y) \tag{63}$$

resulta que

$$\sigma_{aX+bY}^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 \tag{64}$$

Definición 8.12 *El* coeficiente de correlación *de dos v.a. X e Y*

$$\rho_{XY} = \frac{Cov[X,Y]}{\sigma_X \sigma_Y}$$

[1] Del mismo modo que se explicó en la definición 5.8 éste cálculo podría hacerse de otra forma.

Propiedades

- $-1 \leq \rho_{XY} \leq 1$.
- Si $Y = aX + b$ entonces[2] $\rho_{XY} = \operatorname{sgn} a$

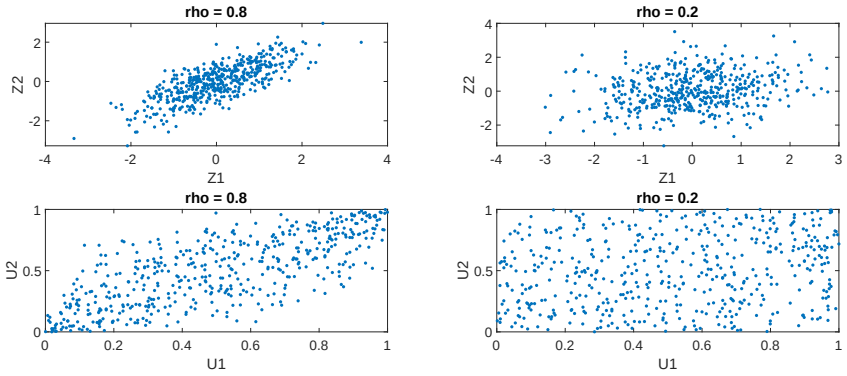


Figura 8.2: Representación de nube de puntos de variables aleatorias bidimensionales con distintos grados de correlación. Las variables Z son tipo normal, y las variables U tipo uniforme.

Ejercicio 8.4.1

La función densidad de probabilidad conjunta de las v.a. X e Y es

$$f_{XY}(x,y) = \begin{cases} xy & 0 \leq x \leq 1, 0 \leq y \leq 2 \\ 0 & \text{en otro caso} \end{cases}$$

Calcular σ_{XY}.

solución: $E[x] = \frac{2}{3}$, $E[y] = \frac{4}{3}$, $E[xy] = \frac{8}{9}$

8.5 SUMA DE VARIABLES ALEATORIAS

Sean X, Y dos variables aleatorias independientes, y sea W la variable aleatoria resultado de sumarlas

$$W = X + Y$$

Esta es una variable aleatoria que aparece muy a menudo, por ejemplo, X podría ser una señal e Y el ruido, para un instante dado, de modo que W sería la señal recibida en un instante dado.

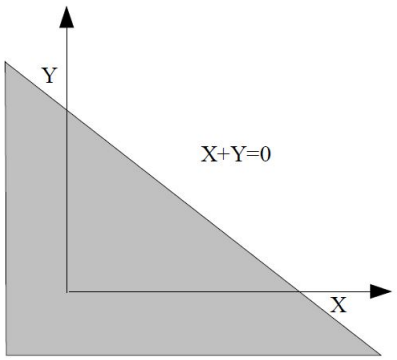
Considérese la función distribución de probabilidad de W:

$$F_W(w) = P\{W \leq w\} = \int_{-\infty}^{\infty} dy \int_{-\infty}^{w-y} f(x,y)dx$$

teniendo en cuenta que X e Y son independientes (60) se obtiene

$$F_W(w) == \int_{-\infty}^{\infty} f_Y(y)dy \int_{-\infty}^{w-y} f_X(x)dx$$

[2] ver el ejercicio 7



Para calcular la función densidad de probabilidad se deriva la expresión anterior

$$f(w) = \int_{-\infty}^{\infty} f_Y(y) f_X(w-y) dy \tag{65}$$

Comentario 5 *La función densidad de la suma de dos variables aleatorias independientes es el* **producto de convolución** *de las funciones densidad individuales.*

EJERCICIOS Y PROBLEMAS

8.1 Las variables aleatorias X e Y tienen una función densidad conjunta

$$f_{XY}(x,y) = \begin{cases} c & 0 \le x \le 3, 0 \le y \le 5 \\ 0 & \text{en otro caso} \end{cases}$$

Calcular el valor de

a) c

b) $P(A) = P(1 \le x \le 3, 2 \le y \le 3)$

c) $P(B) = P(Y > X)$

d) el valor esperado de $G(X,Y) = (XY)^2$

Solución: $c = \frac{1}{15}$, $P(A) = \frac{2}{15}$ $P(B) = \frac{7}{10}$

8.2 Lasvariables aleatorias X e Y tienen una función densidad conjunta

$$f_{XY}(x,y) = \begin{cases} 6(x+y^2)/5 & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{en otro caso} \end{cases}$$

Calcular las funciones de densidad marginales.

Solución: $f_X(x) = \frac{1}{5}(6x+2)$ para $0 \le x \le 1$ y $f_Y(y) = \frac{1}{5}(3+6y^2)$ para $0 \le y \le 1$.

8.3 La función densidad conjunta de dos variables aleatorias es:

$$f(x,y) = \begin{cases} 24xy & x > 0, y > 0, x+y \le 1 \\ 0 & \text{en caso contrario} \end{cases}$$

Calcular

a) Las funciones de densidad marginales.

b) ¿Son independientes?

solución: a) $f_X(x) = 12x(1-x)^2$ si $0 \le x \le 1$ $f_Y(y) = 12y(1-y)^2$ si $0 \le y \le 1$ c) No son independientes puesto que $f(x,y) \neq f_X f_Y$

8.4 La función densidad relacionada con el experimento aleatoria lanzar un dardo a una diana es:

$$f(x,y) = \frac{e^{-(x^2+y^2)/2\sigma^2}}{2\pi\sigma^2}$$

Calcular σ si el 80 % de los lanzamientos caen dentro de un círculo de radio 6cm.

Solución: $\sigma = r/\sqrt{2\ln 5} = 3.34cm$

8.5 El tiempo de conducción que le lleva a uno llegar al trabajo es una v.a. Y que depende del tráfico que depende de la hora de salida X. La salida se produce en un intervalo

aleatorio de longitud T_0 que empieza a las 7:30, el tiempo de conducción es de mínimo T_1. Se sabe que la función densidad conjunta es

$$f(x,y) = c(y-T_1)^3\theta(y-T_1)(\theta(x)-\theta(x-T_0))e^{-(y-T_1)(x+1)}$$

Calcular

a) c

b) La hora media a la que sale ($T_0 = 1h$)

c) Tiempo medio de conducción si la salida se produce a las 7:30

d) Tiempo medio de conducción si la salida se produce a las 8:30

Solución: a) $c = \frac{4}{7}$, b) $E(x) = \frac{2}{7}$, salida a la 7:47, c) $T_1 + 4h$, d)$T_1 + 2$

8.6 Se X una v.a de media 3 y varianza 2 y sea Y = -6X+22. Calcular E[Y], E[XY]

Solución: E[y] = 4, E[XY] = 0

8.7 Dos v.a. están relacionas por $Y = aX + b$. Calcular en función de μ_X y σ_X:

a) μ_Y y σ_Y^2

b) $Cov(X,Y)$ y ρ_{XY}

Solución: $\mu_Y = a\mu_x + b, \sigma_Y^2 = a^2\sigma_X^2, Cov(X,Y) = a\sigma_X^2$

8.8 Encontrar la función distribución de probabilidad conjunta F_{XY} si las v.a. X e Y tienen una función densidad conjunta

$$f_{XY}(x,y) = \begin{cases} 2 & 0 \leq y \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

Solución: Pag 170 Yates y Goodman, 1999

8.9 Usar (65) para comprobar que la suma y la resta de dos variables gaussianas independientes son variables gaussianas de $\mu = \mu_X + \mu_y$ y $\sigma = \sqrt{\sigma_x^2 + \sigma_y^2}$ en el caso de la suma y $\mu = \mu_X - \mu_y$ y $\sigma = \sqrt{\sigma_x^2 + \sigma_y^2}$ para la resta.

9 INFERENCIA ESTADÍSTICA

9.1 INTRODUCCIÓN

El objetivo principal de la estadística inferencial es analizar e interpretar los datos de una muestra para extraer conclusiones acerca de la ley de probabilidad del fenómeno en estudio. O sea, que estamos interesados en el estudio de una variable aleatoria X cuya función distribución $F_X(x)$ es mayor o menor grado conocida.

Tipos de inferencias:

- Inferencia paramétrica:

  En éste caso conocemos la distribución que sigue la población, pero no sabemos los parámetros que la caracterizan. Por ejemplo, se sabe que la población sigue una distribución normal pero no sabemos μ y/o σ. El objetivo, en este caso, será estimar estos parámetros.

- Inferencia no paramétrica:

  En este caso ni siquiera se sabe la distribución teórica de la población. En estos casos se suele tomar muestras "grandes", para poder recurrir al teorema del límite central y estimar los valores de μ y σ.

Para abordar el problema de la estimación de parámetros hay dos enfoques posibles:

- Enfoque clásico:

  Se basada en la idea de que la población está caracterizada por unos parámetros dados, sean o no conocidos, y por tanto el trabajo se limita a estimar estos parámetros ya sea asignándoles un valor, o un intervalo donde éstos se encuentran.

- Enfoque bayesiano:

  Se basan en la idea que los parámetros de la población son a su vez variables aleatorias que siguen distribuciones de probabilidad. En este caso se desea estimar el valor de los parámetros estadísticos de una población teniendo en cuenta la distribución "a priori" de estos datos, obteniéndose al final, una distribución "a posteriori" para los parámetros estadísticos.[1]

En lo que sigue sólo se tendrá en cuenta el enfoque clásico.

9.1.1 *Teoría de muestreo*

En el capítulo 2 ya se han dado algunas definiciones básicas desde la perspectiva de la estadística descriptiva. En particular en la sección 2.5 se define muestra (definición 2.3) pero es muy importante el método usado para determinar esta muestra. En esta sección se analizan algunos aspectos importantes relacionados estos métodos.

[1] Para ver algún ejemplo dirigirse a Walpole et al., 2012

- Muestreo aleatorio simple:

  El muestro aleatorio simple es aquel en que cada muestra tiene la misma probabilidad de ser elegida, en este caso para cada observación muestral, todos los individuos de la población tienen la misma probabilidad de ser elegidos, en la práctica se pude distinguir entre dos posibilidades, cuando la población es infinita o, finita, en el segundo caso los individuos son escogidos con reposición, en el primer caso es innecesario. Sin embargo, en el caso que la población sea mucho mayor que la muestra a menudo se considerará que la población es infinita.

  Para escoger los individuos con un muestro simple se suelen usar la tablas de números aleatorios (Ver Murray y Abellana, 1992 para encontrar tablas de números aleatorios, y Cao et al., 2001 para ver ejemplos de su uso).

- Muestreo sistemático

  Es el usado cuando la población N está ordenada, entonces para escoger los n individuos de la muestra se calcula la parte entera de $E(N/n) = k$ y se escoge aleatoriamente un valor $l \in [1, \ldots, k]$ de modo que se escogerán los individuos que ocupen las posiciones $\{l, l+k, l+2k, \ldots, l+(n-1)k\}$.

  Este muestreo es técnicamente más sencillo de llevar a cabo, y en el caso que los individuos cercanos tengan caracteres similares, sirve para cubrir más homogéneamente la población

- Muestreo estratificado

  Éste es el que se practica cunado la población está estratificada (por ejemplo, la población distingue por sexos, o por edades, o por nivel de estudios ...) en este caso se practica alguno de los otros dos métodos en cada uno de los estratos, pudiendo escoger el mismo número de individuos por cada estrato ("**afijación**"), o una afijación proporcional u otra[2].

- Muestreo conglomerado

  Éste es el que se da cuando la población está dividida en grupos que son homogéneos entre sí, (por ejemplo saber el estado de los ordenadores de una empresa con distintas sedes), en este caso se escogen algunos grupos (en este caso sedes) y se hace una estadística analizando todos los individuos del grupo escogido.

En la práctica la actividad más común es realizar **muestreos polietápicos**, esto es que combinan varios de los muestreos analizados. (por ejemplo para analizar el estado de las carreteras de un país, se hace un muestreo conglomerado para escoger alguna comunidades, dentro de ellas un muestreo estratificado, para escoger distintas carreteras según su categoría, dentro de cada estrato un muestreo conglomerado según la zona y finalmente uno de simple o sistemático para escoger el kilómetro de cada carretera a analizar)

En adelante consideraremos que el muestreo que se ha realizado es simple.

[2] Cao et al., 2001.

9.2 ESTIMACIÓN

Definición 9.1 *Una* muestra *de tamaño n, de una variable aleatoria X, son n variables aleatorias $X_1, X_2, \ldots, X_n$ independientes e igualmente distribuidas, con la misma distribución que X.*

Los valores particulares $x_1, x_2, \ldots x_n$ de la muestra son realizaciones de la muestra.

Como consecuencia de la definición anterior y de la definición 8.8 es evidente que la función densidad conjunta

$$f(x_1, x_2, \ldots, x_n) = f_X(x_1) f_X(x_2) \cdots f_X(x_n)$$

Definición 9.2 *Diremos* estadístico *a cualquier función T de la muestra*

$$T(X_1, X_2, \ldots, X_n)$$

Un estadístico es por lo tanto una función de variables aleatorias y que por lo tanto tiene su propia distribución de probabilidades

Definición 9.3 *Se llama* distribución muestral *a la distribución de probabilidad de un estadístico.*

Es evidente que la distribución muestral depende del tamaño de la muestra y de la distribución de la población.

Ejercicio 9.2.1

Se ha generado por ordenador un conjunto de 6 datos que siguen una distribución de Poisson y se han obtenido los números

$$P = \{3, 1, 2, 0, 2, 4\}$$

a) Calcular la media μ y la varianza de la población P

b) Considerar todas las muestras de tamaño 2 que se pueden obtener haciendo un muestreo aleatorio simple.

c) Calcular el valor del estadístico $\overline{X} = \frac{1}{2}(X_1 + X_2)$

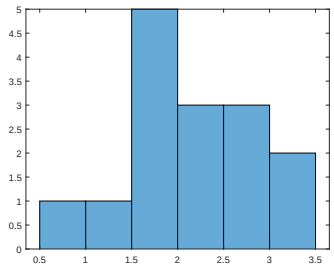
d) Calcular el valor esperado $E(\overline{X})$ y comparar-lo con μ

e) Calcular la varianza de $\overline{X}$ o sea $E((\overline{X} - E(\overline{X}))^2)$ y compararla con la varianza de P.

f) Representar gráficamente los resultados obtenidos en el apartado c) y compararlos con una distribución normal centrada en 2.

Solución: a) Usando las definiciones 2.10 y 2.17 $\mu = 2$ y $\sigma^2 = 1.\hat{6}$;
b) como el P hay 6 elementos el número total de muestras aleatorias[3] de 2 elementos es $C_{6,2} = 15$, algunas de ellas son $\{3, 1\}, \{2, 2\}, \{3, 2\} \ldots$;
c) los 15 resultados obtenidos son {2,2.5,1.5,2.5,3.5,1.5,0.5,1.5,2.5,1,2,1.5,1.5,2,3};
d) Como todas las muestras son igualmente probables, $E(\overline{X})$ es la media de los datos de c), $E(\overline{X}) = 1.933$; e) 0.5622

[3] ver la sección 3.4



9.2.1 *Estimación puntual*

La estimación puntual consiste en asignar el valor de un estadístico $\hat{\Theta} = T(X_1, \ldots, X_n)$ al parámetro θ desconocido de una población.

Definición 9.4 *Se llama* estimador *al estadístico usado para asignar un valor a un parámetro de* X

Evidentemente el problema es escoger un buen estimador para cada parámetro.

Definición 9.5 *Un estimador* $\hat{\Theta}$ *se denomina* insesgado *o* centrado *del parámetro* θ *cuando*

$$E(\hat{\Theta}) = \theta \tag{66}$$

Por consiguiente se llama **Sesgo** del estimador $\hat{\Theta}$ a $sesgo = E(\hat{\Theta}) - \theta$.

Así por ejemplo el estadístico $\overline{X} = \frac{1}{n}\sum_i^n X_i$, llamado media muestral es un estimador insesgado de la media de población, ya que

$$E(\overline{X}) = \mu \tag{67}$$

DEMOSTRACIÓN: Teniendo en cuenta que $\overline{X} = \frac{1}{n}\sum_i^n X_i$ y (63) es evidente que

$$E(\overline{X}) = E(\frac{1}{n}\sum_i^n X_i) = \frac{1}{n}\sum_i^n E(X_i) = \frac{1}{N}\sum_i^N \mu = \mu$$

□

Nótese que el estimador $\overline{X}$ toma el valor $\bar{x}$ para una realización concreta de la muestra, ambos reciben el nombre de media muestral. El mismo criterio de mayúsculas y minúsculas se usará para otros estadísticos.

Por otro lado, tal como se comentó en la sección 2.6, la varianza muestral (1) es un estimador sesgado de σ^2, ya que

$$E(S^2) = \frac{1}{n}\left(n\sigma^2 - \sigma^2\right) = \frac{n-1}{n}\,\sigma^2 \tag{68}$$

DEMOSTRACIÓN: Efectivamente, como

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

Tal como se ha visto, $E(\overline{X}) = \mu$, de modo que $\sigma^2_{\overline{X}} = E((\overline{X} - \mu)^2)$ y de la propiedad de no linealidad[4] de σ^2

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$$

de modo que

$$\begin{aligned} E(S^2) &= \frac{1}{n} E\left(\sum_{i=1}^{n}((X_i - \mu) - (\overline{X} - \mu))^2\right) = \frac{1}{n} E\left(\sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2\right) \\ &= \frac{1}{n}\left(\sum_{i=1}^{n} E\left((X_i - \mu)^2\right) - nE\left((\overline{X} - \mu)^2\right)\right) = \frac{1}{n}\left(n\sigma^2 - n\sigma^2_{\overline{X}}\right) \end{aligned}$$

y por tanto

$$E(S^2) = \frac{1}{n}\left(n\sigma^2 - \sigma^2\right) = \frac{n-1}{n}\sigma^2$$

como se quería demostrar □

Como ya se comentó (sección 2.6) a veces se define

Definición 9.6 *la* varianza muestral corregida o cuasi-varianza

$$S^2_{n-1} = \hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{69}$$

que es un estimador insesgado de la varianza de la población.

La diferencia entre ambos es despreciable para n grandes, en nuestro caso la consideraremos despreciable para $n > 30$.

Aunque se suelen usar preferentemente los estimadores centrados, a veces se suaviza la condición (66) y se definen **estimadores asintóticamente insesgados** cuando

$$\lim_{n \to \infty} E(\hat{\Theta}_n) = \theta \tag{70}$$

Por ejemplo, la varianza es un estimador asintóticamente insesgado de la varianza de la población.

Ejercicio 9.2.2

Demostrar que el estimador $\bar{\mu}_n = \frac{1}{n}(2X_1 + X_2 + \ldots + X_n)$ es un estimador asintóticamente insesgado.

Para valorar la conveniencia o no de un estimador no solamente se toma en consideración si está centrado o no, también se valora su error cuadrático medio.

Definición 9.7 *El* error cuadrático medio (ECM) *de un estimador* $\hat{\Theta}$ *es*

$$ECM(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2]$$

Es fácil de demostrar que

$$ECM(\hat{\Theta}) = \sigma^2_{\hat{\Theta}} - Sesgo(\hat{\Theta})^2$$

[4] Ver la sección 5.6 o mejor aún la ecuación (64)

de modo que, aunque para un mismo parámetro pueden haber varios estimadores insesgados, (por ejemplo para estimar la media se puede usar el estimador $\overline{X}$ o el estimador $\bar{\mu} = \sum_1^n a_i X_i$ con $\sum_i^n a_i = 1$.), si $\hat{\Theta}_1$ y $\hat{\Theta}_2$ son dos estimadores insesgados del parámetro θ se elige el que tiene menor varianza.

Definición 9.8 $\hat{\Theta}$ *es un* estimador eficiente *de* θ *si es insesgado y tiene varianza mínima*

Ejercicio 9.2.3

Demostrar que la varianza[5] del estimador insesgado $\overline{X}$ es

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \tag{71}$$

Solución: Dado que $E(\overline{X}) = \mu$, está claro que $E[(\overline{X} - E(\overline{X}))^2] = E(\overline{X}^2) - \mu^2)$.

$$E(\overline{X}^2) = \frac{1}{n^2} E[(\sum X_i)^2] = \frac{1}{n^2} E\left(\sum X_i^2 + 2X_i X_j \frac{n!}{2!(n-2)!}\right) = \frac{1}{n^2}\left(\sum E(X^2) + \mu^2 n(n-1)\right)$$

de donde se sigue (71).

Ejercicio 9.2.4

Dada X una variable aleatorias normal de media μ y desviación típica σ.

a) Calcular las probabilidades $P(X > \mu)$ y $P(X > \mu + \sigma)$.

b) Sea una muestra de 50 datos sacados de X. Calcular las probabilidades $P(\overline{X} > \mu)$ y $P(\overline{X} > \mu + \sigma)$.

Solución: a) 0.5, 0.15865525003546; b) Dado que X es normal, $\overline{X}$ también es normal, pero incluso en el caso que x no fuera normal, como la muestra es grande, podemos entender que la variable aleatoria $\overline{X}$ sigue una distribución normal de media $E(\overline{X}) = \mu$, y $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$, de modo que, las respuestas son $P(\overline{X} > \mu) = 0.5$ y $P(\overline{X} > \mu + \sigma) = P(Z > \sqrt{n}) \approx 10^{-13}$.

Hasta ahora se han visto algunos estadísticos para estimar los parámetros más habituales (μ, σ^2), alguno de los estimadores usados $(\overline{X})$ se han obtenido de la definición de media en estadística descriptiva, éste es pues un estimador empírico, sin embargo, para estimar la varianza se ha acordado en usar el estimador S_{n-1}^2 (69) porque se trata de un estimador insesgado, aunque no coincide con la definición clásica de varianza. Pero, ¿cuál es el estimador para la p de una distribución binomial?

9.2.2 *Cálculo de estimadores (Opcional)*

Para encontrar estimadores hay varios métodos: el método de máxima verosimilitud y el método de los momentos.

[5] Este resultado es válido para muestreos simples, sin embargo para el caso de muestreos sin reposición se puede comprobar Spiegel, (Mexico, 1998) que

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

donde N es el tamaño de la población

- **Método de máxima verosimilitud o de Fisher**

  Dado que X es una v.a. con función densidad $f(x, \theta_i)$ donde θ_i son los parámetros que caracterizan la distribución, y dada una realización de la muestra $\{x_1, \ldots, x_n\}$ que son n resultados independientes de la variable aleatoria. La probabilidad de que se dé esta muestra viene dada por

$$g(\theta_i) = \prod_{i=1}^{n} P_\theta(X = x_i) \tag{72}$$

  La función $g(\theta_i)$ recibe el nombre de **función de verosimilitud**.

  Comentario 6 *En el caso que la población sea una v.a. continua, para calcular la función de verosimilitud en lugar de usar $P_\theta(X = x_i)$ se usa $f_\theta(X = x_i)$.*

  El método consiste en estimar el valor de los parámetros θ_i que hacen dicha función máxima, o sea, que hacen que la realización de la muestra sea más probable.

  Para hacerlo basta con hacer nula la derivada de (72). A veces resulta más fácil derivar la función $\ln(g(\theta_i))$ cuyos máximos coinciden con los de $g(\theta_i)$ por ser la función logarítmica monótona creciente.

Ejercicio 9.2.5

Demostrar, usando el método de máxima verosimilitud que el estimador de p de variable Binomial es

$$\hat{p} = \frac{c}{n} \tag{73}$$

donde c es el número de éxitos de una muestra con n individuos.

Solución: Dado que X puede tomar dos valores, éxito con probabilidad p y fracaso con probabilidad $1 - p$, si de una muestra de n individuos, hay c éxitos, la función de máxima verosimilitud será

$$g(p) = p^c(1-p)^{(n-c)}$$

de modo que

$$\frac{d\ln(g(p))}{dp} = 0 \rightarrow \; c\frac{1}{p} - (n-c)\frac{1}{1-p} = 0$$

cuya solución es (73).

Ejercicio 9.2.6

Suponga que usan 10 ratas en un estudio biomédico donde se inyecta un medicamento con la intención de aumentar el periodo de supervivencia a un cáncer. Los tiempos de supervivencia son $14, 17, 27, 18, 12, 15, 20, 18, 16, 15$. Suponga que se aplica una distribución exponencial y estimar el parámetro β.

Solución:

$$g(\beta) = \frac{1}{\beta^{10}} e^{-\frac{\sum x_i}{\beta}}$$

de modo que

$$\frac{d\ln(g(p))}{dp} = 0 \rightarrow \; -\frac{10}{\beta} + \frac{\sum x_i}{\beta^2} = 0$$

cuya solución es $\beta = \frac{\sum x_i}{10}$

Ejercicio 9.2.7

Demostrar que los estimadores de máxima verosimilitud para μ y σ^2 de una variable aleatoria normal son respectivamente $\overline{X}$ y S_n^2

Para ver otros ejemplos consultar por ejemplo Narvaiza et al., 2001b.

- **Método de los momentos o de Pearson**

  Este método suele dar peores resultados que el anterior pero és más sencillo de aplicar. Consiste en calcular tantos momentos como parámetros queramos estimar. Dada su definición estadística

$$E(x^j) = \frac{\sum_{i=1}^{n} x_i^j}{N} \tag{74}$$

  como los momentos, en general, dependen de los parámetros, se trata de aislar de la ecuación (74) el parámetro deseado.

Ejercicio 9.2.8

Calcular el estimador de la media y la varianza de una distribución normal.
Solución:

$$E(x) = \mu \qquad E(x^2) = \sigma^2 + \mu^2$$

de modo que $\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{N}$ y $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} x_i^2}{N} - (\frac{\sum_{i=1}^{n} x_i}{N})^2$

Ejercicio 9.2.9

Calcular por el método de los momentos el estimador del parámetro λ de una distribución de Poisson.
Solución: $\lambda = \frac{\sum_{i=1}^{n} x_i}{N}$

Tabla 9.1: Principales estadísticos para hacer estimaciones puntuales.

Parámetro	Estimador
μ	$\overline{X} = \frac{1}{n}\sum_1^n x_i$
σ^2	$\hat{S}^2_{n-1} = \frac{1}{n-1}\sum_1^n (X_i - \overline{X})^2$
p (Binomial)	$\hat{P} = \frac{c}{n}$ donde c número de éxitos

9.3 ESTIMACIÓN POR INTERVALOS DE CONFIANZA

Hasta ahora sólo se ha inferido el valor de los parámetros de una variable aleatoria a partir del cálculo de unos estadísticos de una muestra, pero no se tiene ningún criterio que permita saber hasta que punto los datos obtenidos son fiables o no lo son. La estimación por intervalos de confianza permite superar esta dificultad. La idea es estimar un parámetro ubicándolo en un intervalo

$$\theta_i < \theta < \theta_s$$

Se procede definiendo dos estadísticos $\hat{\Theta}_i$ y $\hat{\Theta}_s$ tales que

$$P(\hat{\Theta}_i < \theta < \hat{\Theta}_s) = 1 - \alpha$$

donde $0 < \alpha < 1$. Lo que significa que hay una probabilidad $1 - \alpha$ de seleccionar una muestra que contenga θ entre $\hat{\theta}_i$ y $\hat{\theta}_s$. El parámetro $1 - \alpha$ recibe el nombre de **coeficiente de confianza**.

Los estadísticos dependen de la distribución muestral del estimador puntual ($\hat{\Theta}$), mientras que su valor concreto depende de la muestra.

De modo que $1 - \alpha$ no mide la probabilidad de que $\theta \in (\theta_i, \theta_s)$ si no que en realidad asegura que de 100 muestras posibles, en el $(1 - \alpha)\,\%$, θ estaría en intervalo calculado con este método.

En las siguientes secciones se va a explicar como hacer estimaciones por intervalos de algunos casos típicos. Empezaremos con el ejemplo de estimar la media de una población con sigma conocida, se trata de un caso poco realista, pero va a servir para ilustrar la idea que se acaba de exponer con un caso simple.

9.3.1 *Estimación de la media de una población con σ conocida*

Hemos visto que para estimar la media de una población, el estadístico más adecuado es la media muestral $\overline{X}$, puesto que se trata de un estimador insesgado, además, se ha visto (71) que la varianza de este estadístico es $\frac{\sigma^2}{n}$.

El teorema central del límite (TCL) asegura que para n suficientemente grande la variable aleatoria $\overline{X}$ está bien aproximada por una distribución $N(\mu, \sigma/\sqrt{n})$. Como se comenta en la sección 7.2.2 se considera que la aproximación es buena para $n \geq 30$, sin embargo cuando la población es a su vez gaussiana la aproximación es buena para cualquier valor de n.[6]

Veamos con un ejemplo como usar el TCL para realizar esta estimación. Supongamos que durante el control de calidad de una empresa que fabrica ruedas para trenes tenemos que estimar el diámetro medio de las ruedas fabricadas. Para ello se coge una muestra de 100 ruedas y se mide el diámetro medio muestral, obteniendo $\bar{x} = 45.2$ cm, con una desviación estándar de 0.1 cm. En este caso como la muestra es grande vamos a suponer que le desviación estándar muestral coincide con la poblacional.

El TCL nos dice que $\overline{X}$ sigue una distribución normal $N(\mu, 0.1/\sqrt{100})$. La figura 9.1 muestra esta distribución, la zona blanca representa un 95 % del área, de modo que para el 95 % de las muestras, la media muestral caerá dentro del intervalo marcado. Evidentemente no sabemos si nuestra muestra concreta se encuentra dentro de este 95 % o forma parte del 5 % restante de muestras.

Lo que se suele hacer es crear un intervalo del mismo tamaño alrededor de $\bar{x}$, no podemos estar seguros que μ estará dentro de este intervalo, no lo estaría en el caso de tener una muestra especialmente alejada de μ, pero tenemos una **confianza** del 95 % en que va ser así, pues con un 95 % de las muestras así sería.

[6] Para el caso, poco común, en que la muestra sea pequeña y la población no tenga una distribución parecida a la normal se puede usar el teorema de Txebyshev (teorema 5.1) para estimar el tamaño del intervalo de confianza.

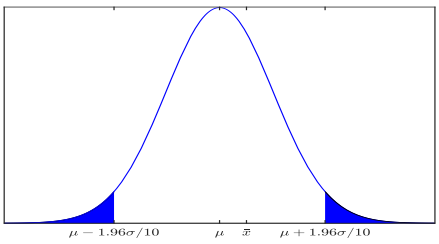


Figura 9.1: De cada 100 muestras en 95 la media estará dentro del intervalo $(\mu - 1.96\,\sigma/\sqrt{n}, \mu + 1.96\,\sigma/\sqrt{n})$.

De ahí que en éstos casos para determinar el intervalo de confianza para una coeficiente de confianza $(1-\alpha)$ se usan los valores $z_{\alpha/2}$ tales que

$$P(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha \tag{75}$$

donde $z_{\alpha/2}$ delimita la zona con área igual a $\alpha/2$ a su derecha. Es evidente que esta ecuación es equivalente a

$$P(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

de modo que el intervalo superior e inferior son respectivamente

$$x_i = \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad x_s = \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{76}$$

Definición 9.9 *El valor* $\frac{\sigma}{\sqrt{n}}$ *lleva el nombre de error estándar o* incertidumbre estándar *de* $\bar{x}$

Teorema 9.1 *Si se utiliza* $\overline{X}$ *como estimador de la media* μ *de una población con desviación típica* σ*, podemos tener una confianza del* $(1-\alpha)100\,\%$ *de que el error no excederá*

$$Error = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{77}$$

Ejercicio 9.3.1

Un ingeniero repite una medición de una variable física X 31 veces y reporta como resultado $\bar{x} \pm s/\sqrt{31}$

a) ¿Qué nivel de confianza podemos tener que el verdadero valor de X está dentro del intervalo dado?

b) ¿Cuál sería la confianza si el valor reportado fuera $\bar{x} \pm 2s/\sqrt{30}$?

c) Con ayuda de un ordenador genere 100 muestras de 31 individuos con una distribución normal N(0,1). Estime el valor de la media como lo hace el ingeniero. ¿En cuántas muestras 0 no está en el intervalo dado?

Solución: a)68.27 %; b) 95.45 %; c) Evidentemente el resultado variará para cada simulación. En nuestro caso las muestras obtenidas en son algo peores de lo esperado, pues solo el 61 % contienen el valor verdadero. La Figura 9.2 ilustra los resultados obtenidos con los intervalos del apartado a) y b).

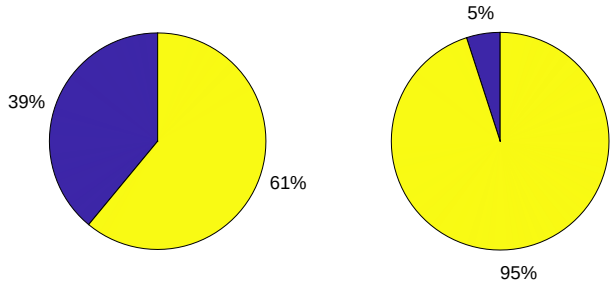


Figura 9.2: En el caso $\bar{x} \pm s/\sqrt{n}$ un 39 % de las muestras no incluyeron el valor verdadero. En el caso de indicar la incertidumbre como $2s/\sqrt{n}$ sólo el 5 % de las muestras no incluyeron el 0.

La ecuación (77) puede usarse para determinar el tamaño que debería tener la muestra para que el error estimado de la medida no exceda se uno determinado:

Teorema 9.2 *Si se utiliza $\overline{X}$ como estimador de la media μ de una población con desviación típica σ, podemos tener una confianza del $(1-\alpha)100\,\%$ de que el error no excederá de una cantidad L si el número de individuos de la muestra es el siguiente entero de*

$$n = \left(\frac{z_{\alpha/2}\sigma}{L}\right)^2 \tag{78}$$

Ejercicio 9.3.2

Los pesos en gramos de unos determinados tubos son: 506, 508, 499, 503, 504, 497, 512, 510, 514, 505, 493, 496, 506, 502, 509 y 496.

a) Se supone que estos están distribuidos normalmente con una desviación típica de 5 g. Obtener los intervalos de confianza al 90 %, 95 % y 99 % la media de estos tubos.

b) Determinar el tamaño muestral necesario para que con $\alpha = 0.05$ la longitud del intervalo esa menor o igual a la unidad.

Solución: a) Si se calcula la media se obtiene $\bar{x} = 503.75$, esta variable aleatoria está distribuida normalmente con $\sigma_{\bar{x}} = 5/\sqrt{16}$. Hay que encontrar x_i y x_s tal que $P(x_i < \bar{x} < x_s) = 0.90, 0.95, 0.99$, para hacerlo se tiene en cuenta que $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{16}}$ es una variable aleatoria normal tipificada.

Considerando un intervalo será simétrico, se busca $z_{\alpha/2}$ tal que $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 0.90, 0.95, 0.99$. Con ayuda de un ordenador o de la tabla del apéndice A, se obtienen, respectivamente $z_{\alpha/2} = 1.65, 1.96, 2.58$, y usando (76) se encuentran los intervalos $(501.7, 505.8)$, $(501.3, 506.2)$ y $(2005., 507.0)$

b) Para responder a la segunda pregunta, basta con usar (78) obteniéndose

$$n = \left(\frac{z_{\alpha/2}\sigma}{L/2}\right)^2 = \left(\frac{1.965}{0.5}\right)^2 = 384.1600 \rightarrow \; n \geq 385$$

Ejercicio 9.3.3

Una muestra al azar de 50 notas de matemáticas, revela que la media es 7.5 y que la desviación típica es 1.

a) ¿Cuáles son los límites de confianza 95 % para estimaciones de la media de las notas?

b) ¿Con qué grado de confianza podríamos decir que la media es 7.5 ± 0.1?

Solución: a) $\mu = 7.5 \pm 0.28$ b) $z_{\alpha/2} = 0.707 \rightarrow (1-\alpha) = 0.52)$

Ejercicio 9.3.4

De una población infinita de desviación típica 4, se ha cogido una muestra de tamaño 250 cuya media muestral es 7.4

a) Dar una estimación por intervalos de la media con un coeficiente de confianza del 99 %

b) ¿Cuál debería ser el valor de n para que $\mu = 7.4 \pm 0.4$?

Solución: a) $z_{\alpha/2} = 2.58$, de modo que $\mu = (6.75, 8.05)$ b)$n = \left(\frac{2z_{\alpha/2}\sigma}{L}\right)^2 = \left(\frac{2.58 \cdot 4}{0.4}\right)^2 \cong 666$

9.3.2 *Estimación de la media de una población, con σ desconocida*

Hasta ahora se ha supuesto que se conocía la varianza de la población, pero esta situación no siempre se da, en muchos casos será necesario usar la propia muestra para estimar la varianza de la población. El estimador usado para tal propósito suele ser el estimador insesgado S^2_{n-1} (69) sin embargo, en este caso, la convergencia de la $\overline{X}$ a una normal no es muy rápida, porque s no es necesariamente cercana a σ, por lo que se suelen usar otras aproximaciones:

En particular para el caso que X sea una variable aleatoria $N(\mu, \sigma)$, la variable

$$T = \frac{\overline{X} - \mu}{S_{n-1}/\sqrt{n}} = \frac{\overline{X} - \mu}{S_n/\sqrt{n-1}} \tag{79}$$

tiene una distribución t de Student con $\nu = n - 1$ grados de libertad[7].

Se puede demostrar que esta variable aleatoria es simétrica y tiene

$$\mu = 0 \qquad \sigma = \frac{n-1}{n-3} \quad \text{para } n > 3$$

En este caso los límites del intervalo vienen dados por

$$x_i = \bar{x} - t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \qquad x_s = \bar{x} + t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} \tag{80}$$

Los valores típicos de $t_{\nu,\alpha/2}$ se pueden encontrar en la tabla del apéndice B, aunque también es posible usar algún software para calcularlos, como se explica en el capítulo 11.

Definición 9.10 *El valor $\frac{s}{\sqrt{n}}$ lleva el nombre de **incertidumbre estándar estimada** de $\bar{x}$.*

Comentario 7 *Aunque el resultado (80) es válido sólo cuando X está normalmente distribuida se puede usar el resultado como aproximación para distribuciones en forma de campana.*

Para este tipo de distribuciones, los datos atípicos son muy extraños, por lo que no conviene usar la distribución t cuando la muestra contenga datos atípicos.

[7] La distribución t de Student debe su nombre al seudónimo que usaba su descubridor el británico William Sealey Gosset en 1908. Un estadístico que trabajaba en las destilerías de Guinness en Dublín que había prohibido a sus empleados publicar cualquier resultado por temor a que se violaran secretos industriales.

La Figura 9.3 muestra la función densidad de probabilidad de varias variables t de Student con grados de libertad $\nu = 5$, 15 y 30. Como se puede observar, muestras con $n > 30$, esta variable aleatoria es casi indistinguible de la normal.

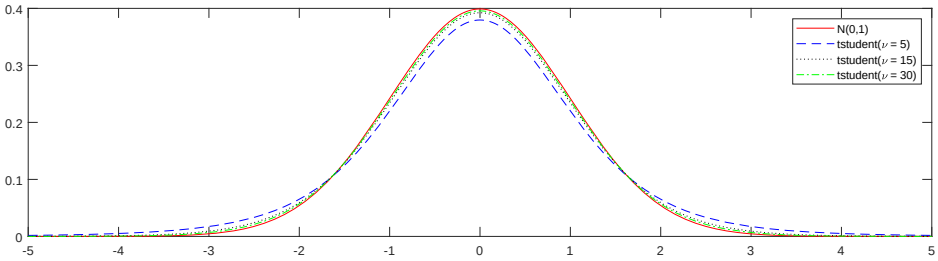


Figura 9.3: Gráficas de la función de densidad de la variable aleatoria t de Student para distintos grados de libertad en comparación con $N(0,1)$

Comentario 8 *Para el caso que la muestra sea superior a 30 la distribución t de Student y la normal son prácticamente iguales, pero para los otros casos la diferencia es significativa.*

Ejercicio 9.3.5

Un metalúrgico estudia un nuevo proceso de soldadura. Fabrica cinco uniones soldadas y mide la resistencia producida por cada uno. Los cinco valores (en ksi) son 56.3, 65.4, 58.7, 70.1 y 63.9. Suponga que estos valores son una muestra aleatoria de una población aproximadamente normal y determine un intervalo de confianza para la media de la resistencia de las soldaduras hechas por este proceso con un nivel de confianza del 95 %.

Solución: $\overline{X} = 62.88$ y $s = 5.48$ como $\frac{\overline{X}-\mu}{s/\sqrt{n}} = t_{n-1}$ y $n = 5$, en la tabla vemos que $t_{4,0.025} = 2.7765$ de modo que $\mu = 62.88 \pm 2.7765 \cdot 54.8/\sqrt{5} = 62.88 \pm 6.81$

9.3.3 *Estimación de la varianza de una población*

Como se ha visto el estimador puntual para la varianza de la población es la quasivarianza o varianza muestral (69), pero ahora nos interesa hacer une estimación por intervalos, para ello como en el caso de la media se usará una variable aleatoria auxiliar.

En el caso que X sea una variable aleatoria $N(\mu, \sigma)$ la variable $Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$ es una variable χ^2 con 1 grado de libertad. En el caso que $X_1, \ldots, X_n$ sean variables aleatorias independientes $N(\mu, \sigma)$ entonces

$$\sum_1^n \left(\frac{X_i - \mu}{\sigma}\right)^2 = \chi^2_\nu$$

es una variable aleatoria χ^2 con $v = n - 1$ grados de libertad. Se puede demostrar que para esta variable

$$\mu = n \qquad \sigma = 2n$$

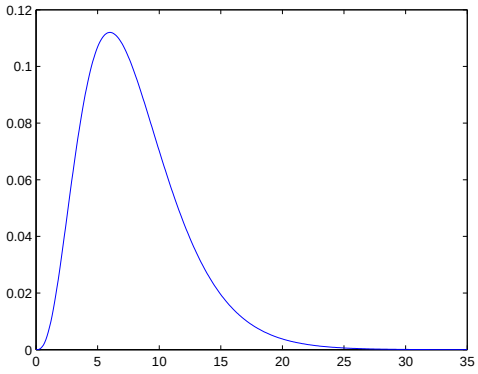


Figura 9.4: Gráfica de la función densidad de probabilidad de la distribución χ^2_8.

Comentario 9 *A diferencia de lo que pasa con la distribución normal o la t de Student, la χ^2_ν no es una distribución simétrica. Ver figura 9.4.*

Como

$$\sum_1^n \left(\frac{X_i - \bar{x}}{\sigma}\right)^2 = \frac{(n-1)S^2_{n-1}}{\sigma^2}$$

entonces,

$$\chi^2_{n-1} = \frac{(n-1)S^2_{n-1}}{\sigma^2} \tag{81}$$

de modo que la distribución χ cuadrado es la usada para determinar los intervalos de correspondiente a la varianza muestral. Efectivamente, con ayuda de las tablas del anexo C, o con algún software[8], se pueden encontrar los valores críticos tales que

$$P\left(\chi^2_{1-\alpha/2} < \frac{(n-1)S^2_{n-1}}{\sigma^2} < \chi^2_{\alpha/2}\right) = 1-\alpha$$

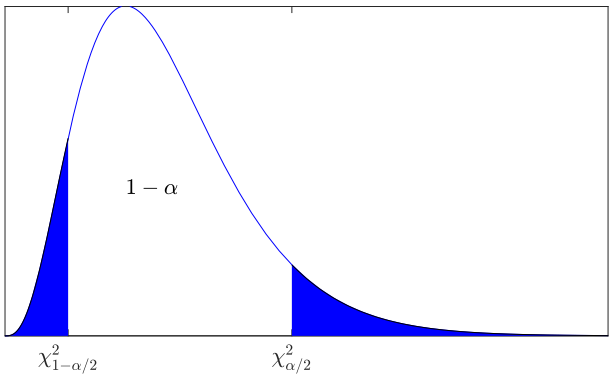


Figura 9.5: Las tablas permiten determinar los valores de χ^2 que tienen áreas $\alpha/2$ y $1-\alpha/2$ a su derecha.

y por tanto

[8] Ver explicación en el capítulo 11.

$$P\left(\frac{(n-1)S_{n-1}^2}{\chi_{1-\alpha/2}^2} > \sigma^2 > \frac{(n-1)S_{n-1}^2}{\chi_{\alpha/2}^2}\right) = 1-\alpha$$

Esta última expresión[9] prueba que, si s^2 es el valor de estimador de la varianza de una muestra de tamaño n de una población normal, el intervalo de confianza de $(1-\alpha)100\,\%$ para σ es

$$\sigma^2 \in \left[\frac{(n-1)s^2}{\chi_{\alpha/2}^2}, \frac{(n-1)s^2}{\chi_{1-\alpha/2}^2}\right] \tag{82}$$

El hecho que la χ^2 sea una distribución asimétrica impide poner es este intervalos como $S^2\pm$ incertidumbre de la medida.

Los límites del intervalo de confianza para la desviación estándar σ son las raíces de los valores dados en (82).

Ejercicio 9.3.6

Una agencia de alquiler de coches necesita estimar el número medio de kilómetros que realiza su flota de automóviles. Para ello, en varios días toma recorridos de 61 vehículos de su flota y obtiene que la media muestral es 165 km/día y que $S_{n-1} = 6$ km/día.

a) Construir un intervalo de confianza para la media a un nivel de confianza del 95 %.

b) Construir un intervalo de confianza para la varianza a un nivel de confianza del 90 %.

c) Construir un intervalo de confianza para la desviación típica a un nivel de confianza del 90 %.

Solución: Como $n = 61 > 30$ a pesar de ser desconocida la σ de la población se puede usar la distribución normal para determinar el intervalo: $\mu \in [\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}] = [163.4943, 166.5057]$, ya que $z_{\alpha/2} = 1.96$. Si se usara la t-Student, se obtendría $\mu \in [\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}] = [163.4633, 166.5367]$, ya que $t_{\alpha/2,n-1} = 2.0003$.

Para la varianza, el intervalo será $\sigma^2 \in [\frac{(n-1)S^1}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^1}{\chi^2_{1-\alpha/2,n-1}}] = [45.52, 83.36]$, ya que con $\alpha = 0.1$ y n = 61, $\chi^2_{1-\alpha/2,n-1} = 43.188$ y $\chi^2_{(\alpha)/2,n-1} = 79.082$.

Para estimar la desviación típica se calcula la raíz del intervalo de la varianza; $[6.75, 9.13]$.

9.3.4 *Estimación de la diferencia entre dos medias*

En ocasiones en lugar de estimar un parámetro de la población se trata de evaluar las diferencias entre dos muestras que presumiblemente corresponden a dos poblaciones $\{X, Y\}$. En esta situación, se estudiarán varios casos. Uno de ellos es la diferencia de medias $(\mu_X - \mu_Y)$. El estimador puntual es $\overline{X} - \bar{Y}$, es decir, en este caso, se toman 2 muestras distintas, una de cada población, con n_X y n_Y elementos respectivamente, y se calcula $\bar{x} - \bar{y}$.

Como se ha visto anteriormente, en el caso que las muestras sean grandes o siendo pequeñas en el caso de que las poblaciones sean normales, la variable aleatoria $\overline{X}$ sigue

[9] En algunos textos, los símbolos $\chi^2_{\alpha/2}$ y $\chi^2_{1-\alpha/2}$ se usan para determinar los puntos con áreas $\alpha/2$ y $1-\alpha/2$ a su izquierda, con lo cual habría que intercambiar las expresiones de los extremos del intervalo. Por ejemplo la función `chi2inv(p,nu)` de Matlab, da el valor de χ^2_{n-1} que delimita el área p a su izquierda.

una distribución normal y la combinación lineal de variables normales es normal, por lo que

$$\overline{X} - \bar{Y} \approx N(\mu_X - \mu_Y, \sqrt{\sigma^2_{\overline{X}} + \sigma^2_{\bar{Y}}}) = N(\mu_X - \mu_Y, \sqrt{\sigma^2_X/n_X + \sigma^2_Y/n_Y}) \tag{83}$$

Como en el caso de la estimación de la media, se puede usar esta aproximación para dar un intervalo de confianza al $(1-\alpha)100\,\%$, para la diferencia de medias

$$\mu_X - \mu_Y \in \left[\bar{x} - \bar{y} - z_{\alpha/2}\sqrt{\sigma^2_X/n_X + \sigma^2_Y/n_Y}, \bar{x} - \bar{y} + z_{\alpha/2}\sqrt{\sigma^2_X/n_X + \sigma^2_Y/n_Y}\right] \tag{84}$$

donde $z_{\alpha/2}$ es el valor de la normal tipificada que deja un área $\alpha/2$ a su derecha.

En el caso, muy probable, que las desviaciones estándar de las dos poblaciones sean desconocidas, si las muestras son suficiente mente grandes se suele substituid σ_i por s_i, es decir se asume que la estimación puntual de la desviación estándar, es suficiente mente buena para calcular el intervalo e confianza de la diferencia de medias, i.e.,

$$\mu_X - \mu_Y \in \left[\bar{x} - \bar{y} - z_{\alpha/2}\sqrt{s^2_X/n_X + s^2_Y/n_Y}, \bar{x} - \bar{y} + z_{\alpha/2}\sqrt{s^2_X/n_X + s^2_Y/n_Y}\right] \tag{85}$$

Ejercicio 9.3.7

Se ha desarrollado un diseño nuevo de foco que se piensa durará más que el diseño viejo. Una muestra aleatoria simple de 64 focos nuevos tiene un tiempo de vida promedio de 578 horas y una desviación estándar de 22 horas. Una muestra aleatoria simple de 144 focos viejos tiene tiempo de vida promedio de 551 horas y desviación estándar de 33 horas. Se quiere encontrar un intervalo de confianza de 95 % para la diferencia entre la media de los tiempos de vida de los focos de los dos diseños.

Solución: $\bar{F}_n - \bar{F}_v \approx N(578-551, \sigma_{\bar{F}_n - \bar{F}_v})$ con $\sigma_{\bar{F}_n - \bar{F}_v} = \sqrt{22^2/64 + 33^2/144} = 3.89$ de modo que tenemos una confianza del 95 % en que $\mu_{F_n - F_v} = 27 \pm 1.96 \cdot 3.89 = 27 \pm 7.62$

En el caso de que las muestras no sean grandes, y en el caso muy probable de que las varianzas de las poblaciones X e Y sean desconocidas estas, se pueden substituir por su estimación puntual s_X y s_Y, pero en este caso en lugar de usar la distribución normal se usara la distribución t:

$$\frac{\overline{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{s^2_X/n_X + s^2_Y/n_Y}} \approx t_\nu \quad \nu = \frac{(s^2_X/n_X + s^2_Y/n_Y)^2}{\frac{(s^2_X/n_X)^2}{n_X - 1} + \frac{(s^2_Y/n_Y)^2}{n_Y - 1}} \tag{86}$$

donde $t_{\nu,\alpha/2}$ es el valor de la distribución t con ν grados de libertad, que deja un área $\alpha/2$ a su derecha. El valor de ν así calculado raramente será entero, de modo que se suele redondear al entero más cercano.

En resumen, se puede estimar la diferencia de medias de las dos poblaciones con un nivel de confianza $100(1-\alpha)$ con el siguiente intervalo:

$$\bar{x} - \bar{y} \pm t_{\nu,\alpha/2}\sqrt{s^2_X/n_X + s^2_Y/n_Y} \tag{87}$$

donde ν se obtiene de (86).

Existe un caso particular, en el que esta última expresión se puede simplificar un poco. Esto se da cuando las dos poblaciones tienen la misma desviación estándar. En este caso, el intervalo para la diferencia de las medias se calcula de la siguiente manera:

$$\bar{x} - \bar{y} \pm t_{\nu,\alpha/2} s_p \sqrt{1/n_X + 1/n_Y} \qquad \nu = n_X + n_Y - 2 \tag{88}$$

pero, la desviación estándar común s_p, se calcula usando las desviaciones muestrales con la siguiente ecuación:

$$s_p^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2} \tag{89}$$

Para poder usar las expresiones (88) y (89) deberíamos saber que las dos poblaciones tienen la misma varianza, pero lo normal, es que de las poblaciones lo único conocido sea la información extraída de las muestras, de modo que, lo que se suele hacer es estimar el cociente de las varianzas.

9.3.5 *Estimación del cociente de varianzas*

Cuando se tienen dos muestras de dos poblaciones distintas X e Y, a veces se quieres valorar la posibilidad de que estas tengan la misma varianza, en este casos, en lugar de analizar la diferencia de las varianzas se estima el cociente de las mismas. El estimador puntual para σ_X / σ_Y es el cociente de ñlas varianzas mustrales s_X / s_Y.

Para hacer una estimación por intervalos de este cociente se usa una nueva distribución auxiliar, se trata de la distribución F de Snedecor, en honor al matemático estadounidense que trabajó en el ámbito del análisis de varianza y diseño de experimentos a principios del siglo XX.

Para poder hacer una estimación por intervalos se usa la siguiente igualdad

$$F = \frac{\sigma_Y^2 S_X^2}{\sigma_X^2 S_Y^2} \tag{90}$$

donde F es la distribución F de Snedecor con grados de libertad $\nu_1 = n_X - 1$ y $\nu_2 = n_Y - 1$ siendo n_i el tamaño de las muestras tomadas de cada población.

Usando esta ecuación, es evidente que

$$P\left(F_{1-\alpha/2}\left(\nu_1, \nu_1\right) < \frac{\sigma_Y^2 S_X^2}{\sigma_X^2 S_Y^2} < F_{\alpha/2}\left(\nu_1, \nu_2\right)\right) = 1 - \alpha \tag{91}$$

donde como es habitual $F_{\alpha/2}(\nu_1, \nu_2)$ es el valor de la distribución F con grados de libertad ν_1 y ν_2 que deja un área $\alpha/2$ a su derecha.

Teniendo en cuenta que $F_{1-\alpha/2}\left(\nu_1, \nu_2\right) = \frac{1}{F_{\alpha/2}(\nu_2, \nu_1)}$ es evidente que

$$P\left(\frac{S_X^2}{S_Y^2} \frac{1}{F_{\alpha/2}\left(\nu_1, \nu_2\right)} < \frac{\sigma_X^2}{\sigma_Y^2} < \frac{S_X^2}{S_Y^2} F_{\alpha/2}\left(\nu_2, \nu_1\right)\right) = 1 - \alpha \tag{92}$$

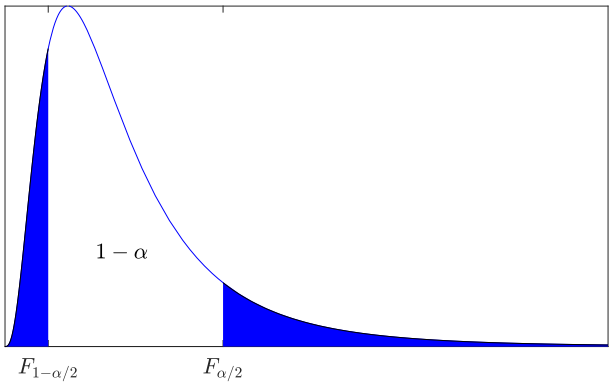


Figura 9.6: Ejemplo de la función de densidad de probabilidad de una distribución F de Snedecor.

De modo que el intervalo de confianza para el cociente de las varianzas viene dado por

$$\frac{\sigma_X^2}{\sigma_Y^2} \in \left[\frac{S_X^2}{S_Y^2} \frac{1}{F_{\alpha/2}(\nu_1,\nu_2)}, \frac{S_X^2}{S_Y^2} F_{\alpha/2}(\nu_2,\nu_1) \right] \tag{93}$$

9.3.6 *Datos Pareados*

Hay algún tipo de estudios en los que se comparan dos poblaciones con datos muy dispersos pero pareados, veamos un ejemplo: consideremos un estudio realizado por una compañía de taxis que tratara de decidir si el uso de llantas radiales en lugar de llantas regulares mejora la economía de combustible. Pare ello se seleccionan una serie de vehículos para hacer mediciones, pero es evidente que el consumo de dependerá del vehículo, del conductor y del recorrido seleccionado para hacer las medidas, una manera de filtrar el efecto de estos factores y poder decidir sobre la conveniencia o no de cambiar as llantas es hacer un estudio pareado: para cada vehículo se selecciona un recorrido y un conductor. Éste realiza las mediciones de consumo con los dos tipos de llantas, de modo que para cada muestra tendrá sentido analizar la diferencia de consumo, ya que los factores ajenos al estudio afectan a las dos medidas.

De algún modo podemos decir que tenemos una única muestra a la que se le han hecho dos "tratamientos", o podemos decir que tenemos 2 muestras, pero en la que cada individuo de una le corresponde un único individuo de la otra, y por tanto estas muestras no son independientes la una de la otra. Otro ejemplo típico de este tipo de datos es el análisis de los efectos de un tratamiento sobre una población, podemos considerar una muestra y medir el parámetro X al que se supone que afecta la medicación antes y después del tratamiento, en este caso para cada individuo tendríamos 2 medidas, una "antes" y otra "después" del tratamiento.

En el caso de tener 2 muestras pareadas, justamente por el hecho de que la diversidad de los individuos afecta a la dispersión del resultado, en el ejemplo de los taxis, el efecto del coche, conductor y recorrido, no tiene sentido analizar la media del "antes" y el "después", llanta radial versus llanta regular, pero sí lo tiene ver la media de las diferencias, es decir para cada trio, (coche, conductor, recorrido) calcular la diferencia de consumo con y sin llanta radial, de modo que, para cada individuo de la muestra se calcula la

diferencia de los datos, obteniendo una única muestra de diferencias, $D = \{d_1, d_2, \ldots, d_n\}$. En este caso, para dar un intervalo de confianza a la media de las diferencias μ_D, se usarán las mismas fórmulas que para el caso de la media de una única población:

$$\mu_D = \left[\bar{D} - t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}}, \bar{D} + t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}}\right] \tag{94}$$

Ejercicio 9.3.8

Consideremos que se hace el estudio sobre llantas mencionado en esta sección. Se equipan 12 autos con llantas radiales, cada vehículo es conducido por su conductor sobre un recorrido de prueba. Sin cambiar de conductores, los mismos autos se equipan con llantas comunes y realiza de nuevo el recorrido de prueba. El consumo de gasolina se registra en la tabla siguiente. Calcule un intervalo de confianza al 95 % para μ_D

litros por kilómetro		
Coche	**Llanta radial**	**Llanta regular**
1	4.2	4.1
2	4.7	4.9
3	6.6	6.2
4	7.0	6.9
5	6.7	6.8
6	4.5	4.4
7	5.7	5.7
8	6.0	5.8
9	7.4	6.9
10	4.9	4.7
11	6.1	6.0
12	5.2	4.9

Solución: Como se trata de datos pareados, calculamos la diferencia de la columna 2 y 3, luego calculamos la media de las diferencias $\bar{D} = 0.1417$ l/km y la desviación estándar muestral $s_D = 0.1975$ l/km. Como la muestra es menor de 30, suponiendo que la diferencia sigue una distribución normal, podemos usar la siguiente expresión para definir el intervalo de confianza: $\left[\bar{D} - t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}}, \bar{D} + t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}}\right]$, como $t_{11,0.025} = 2.201$ podemos decir que tenemos una confianza del 95 % que $\mu_D = [0.016, 0.267]$, es decir, $\mu_D > 0$ lo que significa que, en general, el consumo es mayor con las llantas radiales.

9.3.7 *Intervalos de confianza unilaterales*

En la tabla 9.2, se pueden observar las fórmulas dadas para el cálculo de intervalos de confianza mencionados, como veremos en el capítulo siguiente estas fórmulas y pequeñas modificaciones de las mismas se usan para ayudar a los ingenieros y científicos para tomar decisiones.

Tabla 9.2: Resumen de los principales intervalos de confianza al $(1-\alpha)100\,\%$.

Parámetros	Condiciones	Intervalo
μ	σ conocida o $n>30$	$\left[\bar{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$
μ	σ desconocida, $n\leq 30$	$\left[\bar{x}-t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}, \bar{x}+t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}\right]$
σ^2	Población normal	$\left[\frac{(n-1)s^2}{\chi^2_{\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}\right]$
$\mu_1-\mu_2$	σ_1,σ_2 conocidas o $n_1,n_2\geq 30$	$\bar{x}_1-\bar{x}_2 \pm z_{\alpha/2}\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}$
$\mu_1-\mu_2$	σ_1,σ_2 desconocidas	$\bar{x}_1-\bar{x}_2 \pm t_{\nu,\alpha/2}\sqrt{s_1^2/n_1+s_2^2/n_2}$
$\mu_1-\mu_2$	$\sigma_1=\sigma_2$ desconocidas	$\bar{x}_1-\bar{x}_2 \pm t_{n_1+n_2-2,\alpha/2}s_p\sqrt{1/n_1+1/n_2}$
μ_D	Datos pareados	$\left[\bar{D}-t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}}, \bar{D}+t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}}\right]$
$\frac{\sigma_X^2}{\sigma_Y^2}$	Poblaciones normales	$\left[\frac{S_X^2}{S_Y^2}\frac{1}{F_{\alpha/2}(n_1-1,n_2-1)}, \frac{S_X^2}{S_Y^2}F_{\alpha/2}\left(n_2-1,n_1-1\right)\right]$

$$s_p^2=\frac{(n_1-1)\,s_1^2+(n_2-1)\,s_2^2}{n_1+n_2-2}, \qquad \nu=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}}$$

Las modificaciones más relevantes van a consistir en realizar estimaciones no centradas, es decir, en estimar con intervalos abiertos. En lugar de estimar el parámetro θ dentro de un intervalo cerrado con una confianza del x %, $\theta\in[\theta_{inf},\theta_{sup}]$, se pueden usar intervalos abiertos

$$P(\hat{\Theta}_i<\theta)=1-\alpha$$

de modo que $\theta\in[\theta_{inf},\infty)$, o para el otro intervalo, $\theta\in(-\infty,\theta_{sup}]$

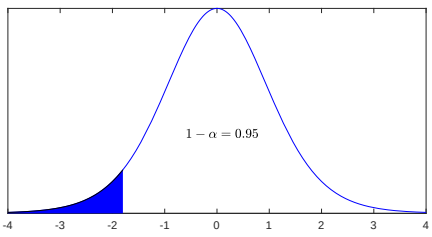
$$P(\theta<\hat{\Theta}_s)=1-\alpha$$

En general los cambios que hay que hacer para determinar estos intervalos son evidentes.

Ejercicio 9.3.9

Sigamos con el ejemplo anterior (Ejercicio 9.3.8). Estimar μ_D con un intervalo abierto por la derecha, con una confianza del 95 %.

Solución: En este caso, para determinar el valor de t que marca el límite del intervalo, se buscará aquel punto con un área de 0.95 a su derecha. Como la distribución t es simétrica, $t_{11,0.95}=-1.796$, de modo que tenemos una confianza del 95 % que $\mu_D=[0.0393,\infty)$



El otro intervalo lateral sería $\mu_D = [-\infty, 0.244)$, lo cual es un resultado poco interesante, en tanto que el intervalo incluye el 0, y por tanto no aporta información clara sobre si es bueno para la reducción de consumo el uso de llantas radiales.

EJERCICIOS Y PROBLEMAS

9.1 En una cadena de autoservicios, el número de clientes que tiene cada una de las tiendas sigue una distribución normal, de media $\mu = 5000$ y desviación típica $\sigma = 500$. Se escoge una muestra de 25 tiendas.

a) ¿Cuál es la probabilidad que la media muestral salga menor que 5075?

b) Diga con una probabilidad del 95 % entre que valores estará la media muestral.

Solución: a) 0.7734; b) [4804,5196].

9.2 Calcular los intervalos de confianza de la media de la población en los siguientes casos

$\bar{x}$	σ	S_n	S_{n-1}	S_n^2	S_{n-1}^2	n	Normal?	Confianza(%)
15.3	3.1					50	sí	95
15.3	3.1					10	sí	95
15.3		3.1				41	?	95
15.3					9.61	10	sí	95
254.9			51.8			26	sí	99
210				20		13	sí	99
15				4		150	?	99

Solución: $I_1 = [14.4407, 16.1593]$; $I_2 = [13.3786, 17.2214]$;$I_3 = [14.3738, 16.2262]$; $I_4 = [13.0824, 17.5176]$; $I_5 = [226.5830, 283.2170]$; $I_6 = [206.3072, 213.6928]$; $I_7 = [14.6310, 15.3690]$

9.3 ¿Cuántos elementos debería tener la muestra en el segundo y tercer caso para que el intervalo se redujera a la mitad sin cambiar el coeficiente de de confianza? **Solución:** a) 40; b) ≈ 157

9.4 Para estimar al diámetro medio de los ejes que fabrica una máquina se escoge una muestra de 10 ejes, obteniéndose lo siguientes resultados en *cm*:

2.02,1.98, 2.04, 1.99, 2.05, 2.00, 2.02, 1.98, 2.03, 2.00

Suponer que el diámetro sigue una distribución normal.

a) Hallar con un coeficiente de confianza del 95 % el intervalo de confianza para la media.

b) Suponiendo que $\sigma = 0.025$ cm, hallar con un coeficiente de confianza del 95 % el intervalo de confianza para μ.

Solución: $I_a = [1.99, 2.029]$; $I_b = [1.995, 2.026]$

9.5 Se tiene una población normal, de media y varianza desconocidas. Se escoge una muestra de 26 elementos y se obtiene

$$\bar{x} = 1905 \quad s^2 = 38025$$

a) Hallar el intervalo de confianza para la media con una confianza del 95 %.

b) Hallar el intervalo de confianza para la varianza con una confianza del 95 %.

Solución: $I_{\mu}^{0.95} = [1826.2, 1983.8]$; $I_{\sigma^2}^{0.95} = [23388, 72458]$

9.6 La desviación estándar en la medida de una pieza concreta es $\sigma = 1$ mm. Se desea determinar la media en un intervalo de confianza de 3 mm de longitud con un coeficiente de confianza del 95 %. ¿Cuantos elementos deber tener la muestra?

Solución: $n = (z_{\alpha/2}\sigma/1.5)^2 \approx 2$

9.7 Se toman 25 elementos de una población con $\sigma_1 = 5$ y se obtiene que $\bar{x}_1$=80, de otra población con $\sigma_2 = 3$ se toman 36 elementos, obteniéndose $\bar{x}_2 = 3$ Calcular el intervalo de confianza al 95$ para la diferencia de medias.

Solución: (2.81,7.19)

9.8 Se han analizado la concentración en metales de dos subsuelos distintos, del primero se han tomado 15 muestras obteniéndose una media de 3.84 g/m^3 y una desviación típica de 3.07 g/m^3. De la segunda zona se han extraído 12 muestras y los resultados han sido una media de 1.49 g/m^3 y una desviación típica de 0.80 g/m^3. Calcular el intervalo de confianza al 95 % de la diferencia en la concentración de metales. Suponer que ambas poblaciones son normales.

Solución: (0.600, 4.10)

9.9 Un investigador de la UCLA afirma que la vida promedio de un ratón se puede prolongar ocho meses más cuando las calorías de su comida se reducen aproximadamente 40 % desde el momento en que se destetan. Las dietas restringidas se enriquecen a niveles normales con vitaminas y proteína. Suponga que se alimenta una muestra aleatoria de 10 ratones con una dieta normal y tiene una vida promedio de 32.1 meses con una desviación estándar de 3.2 meses, mientras que una muestra aleatoria de 15 ratones se alimenta con la dieta restringida y viven un promedio de 37.6 meses con una desviación estándar de 2.8 meses. Suponga que las distribuciones de las vidas con las dietas regular y restringida son aproximadamente normales.

a) Calcula el intervalo de confianza de las medias de las vidas de los dos grupos de ratones con un nivel de confianza del 95 %.

b) Calcula el intervalo de confianza de las varianzas de las vidas de los dos grupos de ratones con un nivel de confianza del 95 %.

c) ¿Qué podemos decir sobre estas varianzas?

d) Estima la diferencia de las medias con un nivel de confianza del 98 %.

Solución: a) $I_{\mu_1}^{0.95} = [29.811, 34.389]$, $I_{\mu_2}^{0.95} = [36.049, 39.151]$;b) $I_{\sigma_1}^{0.95} = [4.84, 34.1]$, $I_{\sigma_2}^{0.95} = [4.20, 19.5]$ c) del apartado anterior como los intervalos se solapan podemos asumir que son iguales, con una confianza del 95 %. Pero para hacerlo al 98 % de confianza, podemos estimar el cociente: $\frac{\sigma_1^2}{\sigma_2^2} \in [0.40698, 4.9606]$, como el 1 está dentro del intervalo podemos decir que es aceptable que las sigmas sean iguales. c)$I_{\mu 1-\mu_2}^{0.98} = [-14.46, 3.4597]$

9.10 Se investiga la resistencia a la tensión de ruptura del hilo proporcionado por dos fabricantes. De la experiencia con los procesos de los fabricantes se sabe que

$$\sigma_1 = 5$$

$$\sigma_2 = 4$$

Una muestra aleatoria de 20 especímenes de prueba proveniente de cada fabricante arroja como resultados

$$\bar{X}_1 = 88$$

$$\bar{X}_2 = 91$$

a) Encuentra un intervalo de confianza del 90 % para la diferencia de las medias de la tensión de ruptura.

b) ¿Existe alguna evidencia que apoye la afirmación de que el hilo del fabricante 2 tiene una mayor resistencia media?

c) Supongamos que queremos construir un intervalo de confianza del 90 % para $\mu_1 - \mu_2$ de modo que el error al estimar esta cantidad sea menos que 1.5 ¿Cuál es el tamaño de la muestra que se debe tomar de cada población?

Solución: a)(-5.355,-0.645)b) como el 0 no está en el intervalo, la respuesta es no con una confianza del 90 %. c) el tamaño debería ser de 50

d) Un experto en eficiencia desea determinar el tiempo promedio que toma perforar tres hoyos en cierta placa metálica. ¿De qué tamaño debe ser una muestra para tener un 95 % de confianza en que esta media muestral estará dentro de 15 segundos de la media verdadera? Suponga que por estudios previos se sabe que $\sigma = 40$ segundos.

Solución: 28

e) Un artículo publicado en *Nuclear Engineering International* describe varias características de las varillas de combustible utilizadas en un reactor propiedad de una empresa noruega de electricidad. Las mediciones notificadas sobre el porcentaje de enriquecimiento de 12 varillas son las siguientes:

2.94	2.75	2.75	2.81
2.90	2.90	2.82	2.95
3.00	2.95	3.00	3.05

Encuentra un intervalo de confianza del 99 % para el porcentaje promedio de enriquecimiento. ¿Estás de acuerdo con la afirmación de que el porcentaje promedio de enriquecimiento es del 2.95 %? ¿Por qué?

Solución: (2.81,2.99)

10

TEORÍA ESTADÍSTICA DE LA TOMA DE DECISIONES

Una de las principales tareas por las que se contrata un ingeniero o a un científico es que este tome decisiones comprometidas, y que lo haga con cierto rigor, evaluando los riesgos que conciernen a dicha decisión. En este capítulo se explica una de las técnicas que usan estos ingenieros y científicos para realizar esta tarea, pero para ello vamos a introducir un léxico técnico apropiado.

En la práctica es frecuente tomar decisiones relativas a una población sobre la base de la información obtenida de una o varias muestras. Tomar una decisión es en realidad, aceptar o rechazar una proposición relativa a la población como puede ser el valor de su media, o el tipo de distribución que sigue. Estas proposiciones reciben el nombre de **hipótesis estadísticas**.

En este capítulo se supone que la distribución es conocida salvo el valor de uno o varios parámetros, de modo que, para nosotros, una hipótesis estadística hace referencia a los parámetros de la distribución de probabilidad que sigue la población.

Definición 10.1 *Una* hipótesis estadística *es una proposición sobre uno o más parámetros de una o varias poblaciones.*

El valor de dichos parámetros se conjetura en función de la experiencia que se tienen sobre la población o experimentos previos que la analizan, en función de un modelo teórico que la describe o sencillamente porqué son los deseados para una población. En cualquiera de estos casos se conjetura un valor o rango de valores y luego se realiza un **contraste de hipótesis**, o test de hipótesis, con el objetivo de validar o rechazar dicha conjetura.

Para realizar este contraste, lo que se hace es recoger información de una o varias muestras aleatorias de la población o poblaciones estudiadas, si esta información es consistente con la hipótesis, esta será aceptada y, si no lo es, será rechazada. En cualquier caso, es importante resaltar que la hipótesis versa sobre propiedades de la población y no sobre la muestra concreta usada para su contraste. Como se verá, hay una relación muy estrecha entre esta técnica de para tomar decisiones y la de definir intervalos de confianza, por ejemplo, en el ejemplo 9.3.8 podríamos tener que tomar una decisión sobre si merece o no la pena cambiar las llantas de los vehículos de alquiler con el objetivo de reducir el consumo (es decir, contrastar si ¿es $\mu_D < 0$?), el intervalo obtenido nos inclina a pensar que efectivamente se trata de una mala idea cambiar de llantas para conseguir dicho objetivo (Tenemos cierta confianza en que $\mu_D > 0$).

Definición 10.2 *Se llama* hipótesis nula H_0 *a la que hipótesis que se va a contrastar.*

Para aceptar o rechazar una hipótesis se usa un estadístico de contraste, que permita tomar una decisión a partir del valor obtenido con una muestra, de modo que

dicho estadístico definirá una región de aceptación de H_0 y una o varias regiones de rechazo. Cuando se rechace H_0, se tomará por buena una **hipótesis alternativa** H_1. Normalmente la hipótesis nula será una hipótesis simple, tipo $\mu = \mu_0$, mientras que H_1 será compuesta, tipo $\mu \neq \mu_0$, $\mu < \mu_0$ o $\mu > \mu_0$.

En cualquier caso, el rechazo o validación de una hipótesis siempre está sujeto a un error, a menos, claro está, que se analice la población en su totalidad. Un test de hipótesis debe evaluar estos errores. En estas condiciones se presentan las siguientes posibilidades:

Decisión	H_0 verdadera	H_0 falsa
Aceptar H_0	No hay error	Error tipo II (β)
Rechazar H_0	Error tipo I (α)	No hay error

Normalmente el error de tipo I recibe el nombre de **nivel de significación** y se representa con α, (su complementario corresponde al nivel de confianza en la decisión de rechazar H_0). El error de tipo II, que se designa con β, es más difícil de controlar, para evaluar-lo deberíamos saber el verdadero valor del parámetro que se contrasta[1].

Cuando se rechaza la hipótesis nula se hace en virtud de ciertas evidencias muestrales, se ha evaluado α, y por tanto, se tiene confianza en la decisión tomada, sin embargo, cuando las evidencias no permiten rechazar H_0, no se pueden descartar otras alternativas, de modo que en la mayoría de los casos el objetivo del contraste es rechazar la hipótesis nula en favor de una H_1. Es decir, a menudo se presenta como H_1 la hipótesis que quiere ser probada por el analista, a la que se llega después de rechazar H_0.

En general, cuando uno de los errores disminuye el otro aumenta, salvo en el caso que el método usado para disminuir los errores sea aumentar el tamaño de la muestra.

Para hacer un contraste de hipótesis los pasos a seguir siempre son los mismos:

(a) Identificar el parámetro poblacional de interés.

(b) Establecer la hipótesis nula H_0.

(c) Establecer la hipótesis alternativa apropiada H_1.

(d) Seleccionar un nivel de significación α.

(e) Establecer el estadístico de prueba adecuado.

(f) Establecer la región crítica para ese estadístico y el nivel de significación elegido.

(g) Calcular el valor del estadístico para una muestra de la población a estudio.

(h) Decidir sobre la aceptación o rechazo de la hipótesis nula.

El hecho de que se acepte o se rechace la hipótesis nula depende exclusivamente de α, puede que, con el mismo valor experimental del estadístico, otra persona decidiese tomar otra decisión, para ello es habitual presentar además el p-valor asociado al valor experimental obtenido

[1] Para ver como estimar el error tipo II ver el ejercicio 10.1.2.

Definición 10.3 *Llamamos* p-valor *al error de tipo I asociado al contraste de hipótesis si el valor obtenido del estadístico fuera el que delimita la zona crítica.*

El conocimiento del p-valor permite, en muchos casos, reforzar la confianza en la decisión tomada. Para ilustrar esta afirmación, ver el ejemplo al final de la siguiente sección.

A continuación veremos algunos casos concretos sobre como plasmar lo mencionado hasta ahora.

10.1 CONTRASTES DE HIPÓTESIS SOBRE LA MEDIA DE LA POBLACIÓN

Para contrastar el valor de la media de una población con un nivel de significación α, H_0 es $\mu = \mu_0$ y el estadístico de contraste será $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$. En el caso que la varianza no sea conocida se usa el estadístico $t_{n-1} = \frac{\bar{x}-\mu}{s/\sqrt{n}}$

- En el caso que H_1 sea $\mu \neq \mu_0$, el **contraste es bilateral** entonces se determinará $z_{\alpha/2}$ de modo que si $z \in (-z_{\alpha/2}, z_{\alpha/2})$ se acepta H_0. El valor de $z_{\alpha/2}$ se obtiene de la tabla del apéndice A, de modo que el área de la curva normal a la derecha de este valor sea $\alpha/2$.

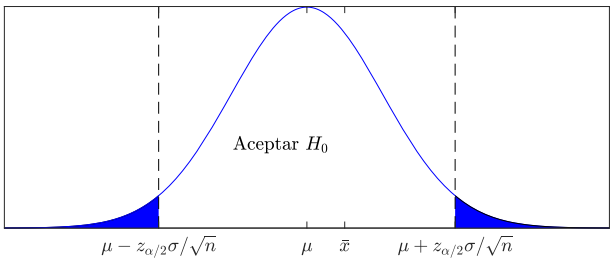


Figura 10.1: Área de aceptación de H_0 en un contraste bilateral.

  En la figura 10.1, se ha dibujado la $\mathrm{N}(\mu_0, \sigma)$ correspondiente a la aceptación de la hipótesis nula, se ha encontrado $z_{\alpha/2}$ que determina los límites de aceptación de H_0 y se ha marcado una hipotética $\bar{x}$ que implicaría la aceptación de H_0.

  Si este fuera el resultado de un contraste deberíamos decir que no tenemos pruebas suficientes que permitan rechazar H_0 por lo que no nos queda más remedio de aceptarla como válida.

  Si, por el contrario, $\bar{x}$ hubiese caído en alguna de las zonas azules el resultado del test sería que tenemos una confianza del (1-α)100 % de que H_0 es falsa y por tanto, aceptamos H_1. La probabilidad de que, siendo H_0 cierta, $\bar{x}$ caiga en la zona azul es α.

- En el caso que H_1 sea $\mu < \mu_0$, el **contraste es unilateral por la izquierda** y se determinará $-z_\alpha$ de modo que si $z > -z_\alpha$ se acepta H_0, y en caso contrario se rechaza.

  De nuevo en esta caso, en la figura 10.2, de ha marcado una $\bar{x}$ que implicaría el no rechazo de H_0.

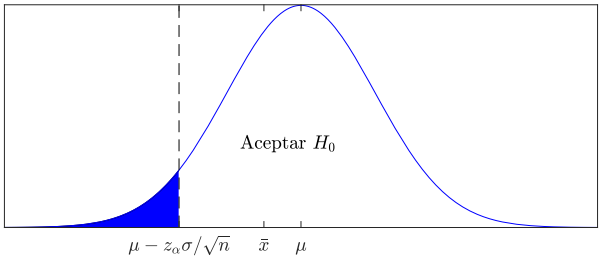


Figura 10.2: Área de aceptación de H_0 en un contraste unilateral.

- En el caso que H_1 sea $\mu > \mu_0$, el **contraste es unilateral por la derecha** y se determinará z_α de modo que si $z < z_\alpha$ se acepta H_0, y en caso contrario se rechaza.

Veamos con un ejemplo lo que se acaba de explicar: Suponga que una empresa que vende tubos fluorescentes dice que la vida media de sus productos es de 1600 h. Se toma una muestra de 100 tubos fluorescentes y se obtiene que $\bar{x}$ es 1570 h con una desviación típica de 120 h. Se va a contrastar la hipótesis $\mu = 1600$ h contra la hipótesis alternativa $\mu \neq 1600$ h usando un nivel de significación de a) 0.05 y b) 0.01.

Como $n > 30$ la media se aproxima bien por una normal. como H_1 es $\mu \neq 1600h$, se trata de un contraste bilateral de modo que en el caso a) $z_{\alpha/2} = 1.96$ y en el b) $z_{\alpha/2} = 2.58$.

Si usamos el estadístico de contraste $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ obtenemos -2.5,lo que implicaría que en el caso a) $\bar{x}$ cae fuera del área de aceptación [-1.96,1.96] y por tanto podemos decir que tenemos evidencias que nos permiten rechazar H_0, y por tanto tenemos una confianza del 95 % de que $\mu \neq 1600$ h.

Sin embargo, en el caso b) $\bar{x}$ sí cae en área de aceptación [-2.58,2.58] y por tanto no podríamos rechazar H_0 con un nivel de significación de 0.01, a pesar de que con este test, no tenemos mucha confianza en que $\mu = 1600$ h.

En un caso así, lo más razonable sería calcular el p-valor, es decir el área sombreada en la siguiente gráfica, que da 0.0124, o sea que podemos rechazar H_0 con un nivel de significación de 0.0124, o lo que es lo mismo, tenemos una confianza del 98.8 % de que $\mu \neq 1600$ h.

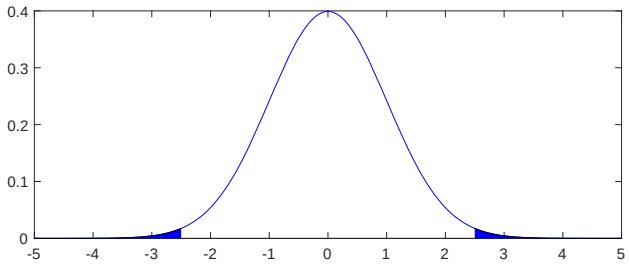


Figura 10.3: El p-valor indica la probabilidad de obtener un valor igual o más extremo que el observado, suponiendo que H_0 sea cierta.

Finalmente, es también posible analizar el resultado de este contraste si la alternativa a H_0 fuera $H_1 : \mu < 1600$. En este caso el p-valor solo sería la mitad del área de la gráfica anterior, de modo que en este caso diríamos que tenemos evidencias de que H_0 es falsa, con un nivel de significación de 0.0062, aceptamos que $\mu < 1600$ h. y tenemos una confianza del 99.37 % en esta afirmación.

Con este último ejemplo es evidente la importancia de seleccionar de manera adecuada H_1 en el momento de diseñar un contraste de hipótesis.

Ejercicio 10.1.1

Una empresa quiere cambiar el sistema iluminación de una máquina, que ahora se ilumina con bombillas clásicas y que tienen un tiempo medio de encendido de 250 ms. Se propone substituir las bombillas por leds, con el objetivo que el tiempo de encendido disminuya. Si se compran 25 leds y se comprueba que, $\bar{x} = 220$ ms con $S = 15$ ms, ¿merece la pena cambiar la configuración de la máquina?

Solución:

Como $n < 30$, la media no se aproxima bien por una normal, además como la varianza no es conocida usaremos, para el contraste la distribución t_{n-1}.

En éste caso $H_0 : \mu = 250$ y como $H_1 : \mu < 250$, se trata, por tanto de un contraste unilateral. Si imponemos un nivel de significación $\alpha = 0.01$ se obtiene $-t_{n-1}(\alpha) = -2.492$ de modo que el intervalo de aceptación de H_0 es: Tiempo de espera ≥ 242.52 por lo que con una confianza del 99 % podemos asegurar que el tiempo de dencendido con leds es menor que con las bombillas tradicionales.

Ejercicio 10.1.2

En una población infinita cuya desviación típica vale $\sigma = 5$, y que se distribuye según una ley normal, se realiza una muestra de tamaño 25, y nos planteamos el siguiente contraste de hipótesis:

$$H_0 : \mu = 15$$

$$H_1 : \mu \neq 15$$

Aceptamos H_0 si la media muestral pertenece al intervalo (13.5, 16.5) Calcula:

a) Calcula la probabilidad de error de Tipo I

b) Suponiendo que el verdadero valor de la media es 17 calcula la probabilidad de error del Tipo II

Solución: Sea X la variable aleatoria $N(\mu, 5)$, de modo que $\bar{X} \in N(\mu, 1)$ a) α =P(Error tipo I)=P(de rechazar H_0 siendo cierta)= $1 - P(13.5 < \bar{x} < 16.5|\mu = 15) = 1 - P(-1.5 < z < 1.5) = 0.1336$

b) P(Error tipo II) = $P(13.5 < \bar{x} < 16.5|\mu = 17) = P(-3.5 < z < -0.5) = 0.3083$

10.2 CONTRASTES DE HIPÓTESIS SOBRE LA VARIANZA

En esta y las siguientes secciones se van a usar los estadísticos comentados en el capítulo anterior, y resumidos en la tabla 9.2 para realizar contraste de hipótesis sobre otros parámetros de la población. En todos los casos se hará con un ejemplo.

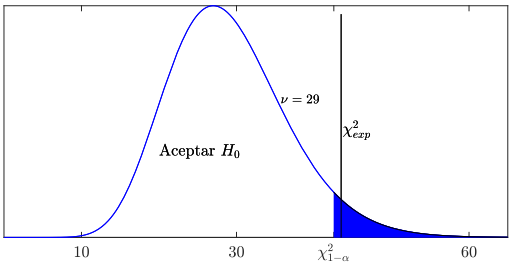
Ejercicio 10.2.1

Se suele decir que una máquina para hinchar ruedas de coche esta estropeada si la varianza del peso de las ruedas una vez hinchadas es mayor que 0.02 kg^2. Se toma una muestra aleatoria de 30 neumáticos y se obtiene que la varianza es de 0.03 kg^2. Suponiendo que la población sigue una distribución normal de media desconocida, indique con un nivel de significación del 5 % si la máquina está estropeada.

Solución: Sospechamos que la máquina está estropeada, por lo que vamos a plantear las siguientes hipótesis: $H_0 : \sigma^2 \leq 0.02$ kg^2 y $H_1 : \sigma^2 > 0.02$ kg^2. En este caso el estadístico de contraste es

$$\frac{(n-1)s^2}{\sigma^2} = \chi^2_\alpha$$

el nivel de significación 0.05 implica que limite del intervalo de acepción es 42.557.



Cono los datos del enunciado $\chi^2_{exp} = \frac{(n-1)s^2}{\sigma^2} = 43.5$ como queda fuera de la zona de aceptación podemos decir que la máquina está estropeada con un nivel de significación de 0.05, es más, si calculamos el valor del área a la derecha del χ^2_{exp}, el p-valor, podemos disminuir el error tipo I a 0.0409.

10.3 CONTRASTES DE HIPÓTESIS SOBRE LA DIFERENCIAS DE DOS MEDIAS

Como se explicó en la sección 9.3.4 hay varios estadísticos para valorar la diferencia entre medias de dos poblaciones dependiendo: i) del número de elementos de las muestras y ii) en el caso de muestras pequeñas, de si la varianza es igual o distinta. Veamos cómo se trasladan estos estadísticos al caso de un contraste de hipótesis.

Ejercicio 10.3.1

Los sindicatos de la construcción de una gran empresa sospechan que los trabajadores cobran distinto en función de su nacionalidad a pesar de realizar tareas parecidas y aseguran que la diferencia de sueldos entre los colectivos nacionales y los extranjeros es de 10000 €. Se sabe que los sueldos de los trabajadores siguen una distribución normal.

Se hace un estudio estadístico con 500 trabajadores nacionales y otros 700 extranjeros, obteniéndose sueldos medios de 55100 y 48700 € respectivamente, siendo las desviaciones típicas 6600 € para los nacionales y 6000 € para los extranjeros.

Realice un contraste de hipótesis con un nivel de significación del 1 % para verificar si los sindicatos tienen razón.

Solución: En este caso las dos poblaciones corresponden a los trabajadores nacionales y los extranjeros. Como las muestras son muy grandes vamos a tomar la desviación estándar muestral como valor de las desviación estándar poblacional ($s_i = \sigma_i$).

Las hipótesis a contrastar serán $H_0 = \mu_N - \mu_E \leq 10000$ y $H_1 : \mu_N - \mu_E > 10000$. El estadístico de contraste será

$$z = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}.$$

En este caso, el valor crítico (z_α) que marca el límite de aceptación de H_0 es 2.33, y el valor experimental obtenido es z_{exp}= -9.67, de modo que no podemos rechazar H_0.

Para arreglar el desaguisado, podemos decir que los sindicatos deberían ser menos ambiciosos en su proclama y decir que la diferencia de sueldos es mayor que 5000 €, en cuyo caso, el valor z_{exp} sería 3.76 y podríamos rechazar H0 y aceptar que $\mu_N - \mu_E > 5000$, con un nivel de significación menor que 0.01, ya que el p-valor en este caso sería de $8.5 \cdot 10^{-5}$.

De forma análoga se haría en los otros casos. La tabla 10.1 recoge los estadísticos de prueba para los casos más comunes en los contraste de hipótesis que involucran medias. Se insta a lector a realizar los ejercicios del final del capítulo para probar distintas de estas opciones.

Tabla 10.1: Contrastes de hipótesis relacionadas con las medias.

H_0	**Estadístico de prueba**	H_1	**Región crítica** Rechazar H_0
$\mu = \mu_0$	$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ σ conocida o n>30	$\mu < \mu_0$	$z < -z_\alpha$
		$\mu > \mu_0$	$z > z_\alpha$
		$\mu \neq \mu_0$	$z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$
$\mu = \mu_0$	$t_{n-1} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$ σ desconocida	$\mu < \mu_0$	$t < -t_\alpha$
		$\mu > \mu_0$	$t > t_\alpha$
		$\mu \neq \mu_0$	$t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$
$\mu_D = d_0$ Datos pareados	$t_{n-1} = \frac{\overline{D}-d_0}{s_D/\sqrt{n}}$	$\mu_D < d_0$	$t < -t_\alpha$
		$\mu_D > d_0$	$t > t_\alpha$
		$\mu_D \neq d_0$	$t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$z = \frac{(x_1-x_2)-a_0}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$ σ_1 y σ_2 conocidas	$\mu_1 - \mu_2 < d_0$	$z < -z_\alpha$
		$\mu_1 - \mu_2 > d_0$	$z > z_\alpha$
		$\mu_1 - \mu_2 \neq d_0$	$z < -z_{\alpha/2}$ o $z > z_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \frac{(\bar{x}_1-\bar{x}_2)-d_0}{s_p\sqrt{1/n_1+1/n_2}}$ $\nu = n_1 + n_2 - 2$ $\sigma_1 = \sigma_2$ pero desconocidas $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	$\mu_1 - \mu_2 < d_0$	$t < -t_\alpha$
		$\mu_1 - \mu_2 > d_0$	$t > t_\alpha$
		$\mu_1 - \mu_2 \neq d_0$	$t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$
$\mu_1 - \mu_2 = d_0$	$t = \frac{(\bar{x}_1-\bar{x}_2)-d_0}{\sqrt{s_1^2/n_1+s_2^2/n_2}}$ $\nu = \frac{\left(s_1^2/n_1+s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1-1}+\frac{\left(s_2^2/n_2\right)^2}{n_2-1}}$ $\sigma_1 \neq \sigma_2$ y desconocidas	$\mu_1 - \mu_2 < d_0$	$t < -t_\alpha$
		$\mu_1 - \mu_2 > d_0$	$t > t_\alpha$
		$\mu_1 - \mu_2 \neq d_0$	$t < -t_{\alpha/2}$ o $t > t_{\alpha/2}$

10.4 CONTRASTES SOBRE PROPORCIONES

Como se ha visto, (ver ejercicio 9.2.5), un buen estimador para el parámetro p de una distribución binomial es $\hat{p} = \frac{x}{n}$ dónde x es el número de éxitos. Si X es el estadístico muestral que denota el número de éxitos entonces, es evidente que éste sigue una distribución binomial. Esto se usará para realizar contraste de hipótesis sobre proporciones, veámoslo con un ejemplo.

Consideremos que se lanza la moneda $n = 12$ veces y se cuenta el número de caras ($x = 8$). Se va a diseñar una regla de decisión para contrastar la hipótesis de si la moneda está trucada o no con un nivel de significación de $\alpha = 0.05$.

Si la moneda no está trucada, lo esperado es que la probabilidad de que sea cara sea $1/2$, de modo que $H_0 : p = 1/2$, como en nuestro caso $\hat{p} = \frac{2}{3}$, la hipótesis alternativa será $H_1 : p > 1/2$. La probabilidad de que siendo cierta H_0, se tenga x éxitos o más, es

$$P(X \geq x) = \sum_{i=0}^{i=x} B(i; n, p)$$

si este valor es menor o igual que α, se acepta H_0, en caso contrario se rechaza H_0 para quedarnos con la hipótesis alternativa H_1.

En el ejemplo, $P(X \geq 8) = 0.073$, por lo que no tenemos evidencias de que la moneda esté trucada.

Muy a menudo, cuando n es grande, se usan el teorema central del límite para aproximar la distribución binomial con una normal, usando que $\mu = np$ y $\sigma = \sqrt{npq}$,

Ejercicio 10.4.1

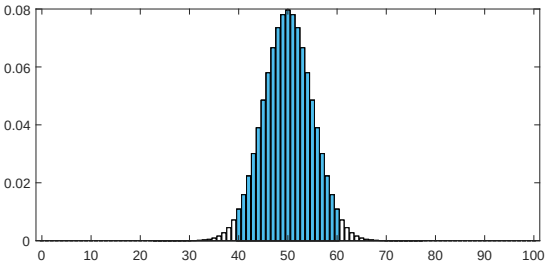
Para saber si una moneda esta trucada o no se toma la siguiente solución: Si el número de caras de una muestra de 100 tiradas está entre 40 y 60 se asume que no está trucada.

a) Hallar la probabilidad de cometer un error de tipo I

b) Representar gráficamente la regla de decisión y el resultado de la parte a).

c) Comentar las conclusiones en caso que el número de caras obtenido fuera 56 y 68.

Solución: a) Con ayuda de un ordenar de puede calcular esta probabilidad $P(40 < X \leq 60) = 0.95396$, otra alternativa es usar la aproximación normal de la binomial, entonces $P(40 < X \leq 60) \approx P(40.5 < N(x; 50, 5) < 60.5) = 0.95342$ lo que significa que el nivel de significación del test es 0.046 b) $z_{\alpha/2} = 1.9901$



c) En el caso x = 56, $\hat{p} = 0.5$ y $\hat{\mu} = 56$, por tanto, $z_{exp} = 1.2 \in (-1.99, 1.99)$ se acepta H_0. En el caso $x = 68$ $z_{exp} = 3.6 \notin (-1.99, 1.99)$ se rechaza H_0 y se pude afirmar que tenemos pruebas de que la moneda está trucada.

Ejercicio 10.4.2

Un laboratorio farmacéutico sostiene que uno de sus productos es 90 % efectivo para reducir una alergia en 8 h. En una muestra de 200 personas con esa alergia el medicamento dio buen resultado en 160 determinar si la afirmación del laboratorio es legítima.

Solución: $H_0 : \mu = n\,0.9 \quad H_1 : \mu < n\,0.9$ el valor obtenido para $\hat{p} = 0.8$ si se se uas un nivel de significación del 1 % resulta que $z_\alpha = -2.33$ de modo que si $z < z_\alpha$ se rechaza H_0 en favor de H_1 si no se acepta H_0. En este caso

$$z = \frac{160 - 180}{\sqrt{200\,0.9\,0.1}} = -4.7$$

de modo que se rechaza H_0

EJERCICIOS Y PROBLEMAS

10.1 Un estudio rebela que los resultados de las pruebas de resistencia a la adhesión de 22 especímenes de aleación U-700 son tales que en promedio, la carga de falla, es $\bar{x} = 13.714$ MPa, mientras que la desviación estándar de la muestra es 3.55. ¿Sugieren estos datos que el promedio de falla es superior a 10 MPa?. Supóngase que la carga donde se produce la falla tiene una distribución normal y utilícese un nivel de significación del 0.05.

Solución: H_0 $\mu = 10$; H_1 $\mu > 10$ como $t_{0.05,\,21} = 1.721$ y en este caso $t_{exp} = 4.90$ se rechaza H_0 y se concluye que con un grado de confianza del 95 % la caraga de fallo promedio es mayor que 10 MPa.

10.2 Se lleva a cabo un experimento para determinar si el acabado superficial tiene un efecto sobre el límite de aguante del acero. Una teoría existente dice que el pulido aumenta el límite de aguante promedio (flexión inversa). El experimento se realiza sobre acero al carbón al 0.4 % con el uso de especímenes sin pulir y con pulido suave. Se obtienen los datos de la tabla.

Resistencia (psi)	
Acero pulido	**Acero sin pulir**
85 500	82 600
91 900	82 400
89 400	81 700
84 000	79 500
89 900	79 400
78 700	69 800
87 500	79 900
83 100	83 400

Desde un punto de vista práctico el pulido no debería tener ningún efecto sobre la desviación estándar del límite de aguante que a partir de los resultados de numerosos experimentos del límite de aguante se sabe que es de 4000 psi.

Encuentre un intervalo de confianza de 95 % para la diferencia entre las medias poblacionales. Suponga que las poblaciones se distribuyen normalmente. ¿Qué podemos concluir de los resultados obtenidos?

Solución: $I^{95\,\%}_{\mu:1-\mu_2} = (17287, 25745)$; si hacemos un contraste con $H_0 = \mu_1 - \mu 2 = 0$ vs. $H_1 : \mu_1 - \mu 2 > 0$ $z_c = 1.64$, $z_{exp} = 2.972$ por tanto rechazamos H_0 con un p-valor de 0.0015.

10.3 El Amstat News (diciembre de 2004) lista los sueldos medios de profesores asociados de estadística en instituciones de investigación, en escuelas de humanidades y en otras instituciones en Estados Unidos. Suponga que una muestra de 30 profesores asociados de instituciones de investigación tiene un sueldo promedio de 70 750$ anuales con una desviación estándar de 6000$. Suponga también que una muestra de 30 profesores asociados de otros tipos de instituciones tiene un sueldo promedio de 65 200$ con una desviación estándar de 5000$.

a) Analiza la veracidad de la siguiente hipótesis con un nivel de significación de 0.01. Los profesores asociados de instituciones de humanidades ganan más de 2000$ más que los de otras instituciones.

b) Consideramos ahora únicamente la muestra de los profesores asociados de humanidades. Calcula entre qué valores se encuentra la desviación estándar de la población con un nivel de confianza del 98 %.

c) El responsable de la escuela de humanidades afirma que la desviación estándar del sueldo de los profesores asociados es menor que 4000$. ¿Es cierta esta afirmación? Razona tu respuesta.

Solución: a) $H_0 : \mu_1 - \mu_2 = 2000; H_1 : \mu_1 - \mu_2 > 20000$ como $z_{exp} = 2.489 > z_c = 2.33$ se rechaza H_0; b) $\sigma \in (4.59, 8.55)$k$; c) No.

10.4 Se analiza en 10 personas el efecto de un medicamento en la concentración de proteína A en sangre (mg/l). A cada paciente se le realiza un análisis antes y después de inyectar el tratamiento, los resultados son lo siguientes:

[A](mg/l)	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
Antes	0.56	0.83	0.34	1.23	0.50	1.13	0.62	0.45	0.39	1.02
Después	0.59	1.03	0.28	1.25	0.85	1.30	0.58	0.59	0.54	0.93

a) Contrastar con un nivel de significación de 0.01 si el medicamento cambia o no la concentración en sangre de la vitamina A.

b) Realizar otro contraste para ver si el medicamento aumenta la concentración de A en sangre.

Solución: Datos pareados. a) Aceptamos $H_0 : \mu_D = 0$ ya que $t_{exp} = 1.9877 \in (-3.2498, 3.2498)$; b) $\mu_D > 0$ el p-valor es 0.039.

10.5 Un vendedor de electrodomésticos dice que unos de sus aparatos consume 46 kWh. Se toman 12 de estos aparatos y se comprueba que consumen de media 42 kWh, con una desviación típica de 11.9 kWh.

a) Realice un contraste de hipótesis con un nivel de significación del 5 % para comprobar si el consumo es 46 kWh o no lo es.

b) Realice un contraste de hipótesis al 5 % para comprobar si el consumo es 46 kWh o menor.

Solución: a)$t_{exp} = -1.122 \in (-2.201; 2.201)$, por tanto $H_0 : \mu = 46$ se acepta, no podemos decir que $H_1 : \mu \neq 46$; b) $t_{exp} = -1.164 > -1.7959$. No podemos asegurar que sea menor.

10.6 Se registran las siguientes mediciones del tiempo de secado, en horas, de cierta marca de pintura vinílica:

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Suponga que las mediciones representan una muestra aleatoria de una población normal.

a) Realice una estimación puntual de la media y la varianza de la población.
b) ¿Podemos afirmar que la media de la población es 4.5 con un nivel de significación del 5 %?
c) Calcula el p-valor del anterior apartado.
d) Estime con un intervalo de confianza del 95 % el tiempo medio de secado de esta pintura vinílica.
e) Suponga ahora que la media de la población es 4.5 h y su desviación estándar 1 h.¿Cuál es la probabilidad de que la media del tiempo de secado de una muestra de 15 ensayos esté dentro del intervalo definido en el apartado anterior?

Solución: a) $\bar{x} = 3.787$ h $s = 0.971$ h; b) No; c) 0.013; d) (3.25,4.32); e) 0.248.

10.7 Si al efectuar mediciones simultáneas de voltaje eléctrico usando dos tipos diferentes de voltímetro se obtienen las siguientes medidas

Voltímetro 1 (V):	10.8	8	15	10.3	10.3	17	12.6	25.2
Voltímetro 2 (V):	10	7.8	15.3	10.3	10.2	16.3	12.1	25

¿Puede afirmarse al nivel del 5 % que no existe diferencia significativa en la calibración de los dos instrumentos? (Se supone normalidad)

Solución: Se trata de un problema de datos pareados. $t_{exp} = 2.11 \in (-2.36, 2.36)$ no podemos rechazar que midan distinto.

10.8 Se afirma que la resistencia del alambre A es mayor que la del alambre B. Un experimento sobre los alambres muestra los resultados de la tabla (en Ω).

A	B
0.140	0.135
0.138	0.140
0.143	0.136
0.142	0.142
0.144	0.138
0.137	0.140

10.8 Calcular la resistencia media de cada alambre al 98 %.
a) Suponga que las varianzas son iguales y haga un test que permita decir cuál de los 2 tiene mayor resistencia. ¿Cuál es el nivel de significación?

Solución: a) $I_{\mu_1}^{0.98} = [0.136, 0.145]$, $I_{\mu_2}^{0.98} = [0.135, 0.143]$; b) $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 > 0$, $d_0 = 0.022$ $t_{exp} = 1.37$ p-valor= 0.1.

10.9 En la caja de un medicamento se indica que contiene 10000 píldoras, pero la asociación de consumidores sospecha que no es cierto, para ello se toma una muestra de 100 unidades y se cuentan el número de píldoras obteniéndose los siguientes resultados:

$$\bar{x} = 973, \qquad s = 114$$

Con un nivel de significación $\alpha = \%1$ indique si el laboratorio miente.

Solución: Se pude rechazar $H_0 : \mu = 1000$ frente $H_1 : \mu < 1000$ $z_{exp} = -2.37 < z_c = -2.33$.

10.10 Se encontró que el contenido de nicotina de dos marcas de cigarrillos, medido en miligramos, es el que se muestra en la tabla.

A	2.1	4.0	6.3	5.4	4.8	3.7	6.1	3.3	3.2	5.4
B	4.1	0.6	3.1	2.5	4.0	6.2	1.6	2.2	1.9	5.4

Responde las siguientes cuestiones:

a) Calcula el intervalo de confianza de las varianzas del contenido en nicotina de las dos marcas, con una confianza del 95 %.

b) A un nivel de significancia de 0.05 pruebe la hipótesis de que las medias del contenido de nicotina de las dos marcas son iguales.

Solución: a) $I^{0.95}_{\sigma_1^2} = [0.91, 6.38]$, $I^{0.95}_{\sigma_2^2} = [1.46, 10.3]$; b) $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$, $t_{exp} = 1.79 \in (-2.11, 2.11)$ no se puede rechazar H_0.

10.11 Josetxo es Youtuber desde hace unos años. Hasta ahora recibía una media de 20 visitas al día en su canal. Hace un mes consiguió un nuevo patrocinador y ha tenido 650 visitas (todo el mes, 31 días).

a) ¿Cuál era el promedio de visitas al canal de Josetxo antes de conseguir el patrocinador? ¿Qué distribución sigue el mencionado número de visitas?

b) ¿Podemos decir con un nivel de significancia del 5 % que el patrocinador le ha ayudado a subir el número de visitas al canal?

Solución: a) 620; b) $z_{exp} = 1.20 < z_c = 1.64$ no podemos rechazar H_0.

10.12 Se analiza la concentración de colesterol (mg/dl) en dos grupos. Los de primer grupo son personas cuyos padres han muerto por problema cardiovasculares, mientras que en los historiales médicos de los del segundo grupo no aparecen problemas relacionados con el corazón. Se han obtenido los siguiente resultados

$$n1 = 100; \bar{x}_1 = 20703; s_1 = 1506$$

$$n2 = 74; \bar{x}_2 = 19304; s_2 = 1703$$

a) ¿Se puede asegurar, con un nivel de significación de 0.05, que los del primer grupo tienen mayor colesterol? (suponer que la varianza de los dos grupos es la misma)

b) Calcular el p-valor e interpretar el resultado.

Solución: a) $H_0 : \mu_1 - \mu 2 = 0$ vs. $H_1 : \mu_1 - \mu 2 > 0$ $z_{exp} = 5.5 >> z_c = 1.645$; por tanto se rechaza H_0; b) con un p-valor de 2.3639e-08. La probabilidad de que las dos poblaciones tuviesen la misma cantidad de colesterol y las medias fuesen las encontrada muy pequeña, no es razonable pensar que así sea.

11

PROBABILIDAD Y ESTADÍSTICA CON MATLAB

Para hacer los cálculos que se han venido explicando a lo largo de todo este curso son de gran utilidad algunos programas informáticos. Es este capítulo se van a explicar algunos comandos de Matlab MathWorks, 2008 que permiten realizarlos. Éste es un software de carácter generalista, que además de poderse usar en el ámbito de la estadística y la probabilidad tiene muchas otras aplicaciones, desde el cálculo matricial, al cálculo diferencial, Inteligencia artificial y Deep Learning, simulación de sistemas mecánicos, eléctricos o electrónicos, entre otras muchas posibilidades.

Siguiendo el esquema general del libro, este capítulo se divide en varias secciones donde se describen las principales funciones de Matlab que permiten realizar los cálculos de los capítulos anteriores. Sin embargo, el objetivo de este capítulo no es desarrollar un curso de Matlab, si no dejar constancia de los comandos más típicos, y se recomienda hacer uso extensivo de la fantástica ayuda de Matlab y los "live-scripts" que la ayuda ofrece. Además, con este libro se pondrán a disposición del lector unos live-scripts con los ejemplos aquí dados.

11.1 ESTADÍSTICA DESCRIPTIVA CON MATLAB

La estadística descriptiva es la parte de la estadística que se dedica organizar, representar gráficamente y analizar los datos experimentales, que incluye, además, las técnicas que permiten inferir características de la *población* en estudio. De modo que si se desea hacer estadística descriptiva en Matlab el primer paso es introducir los datos de la muestra.

Matlab es un software que está pensado para manejar los datos en forma matricial, por este motivo en este capítulo se manejarán los datos cuando estos están almacenados en este formato, a pesar de que hay estructuras alternativas de almacenar los datos (tablas o celdas)

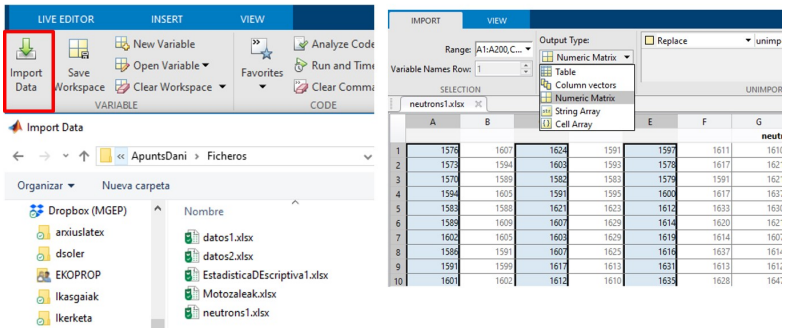
11.1.1 *Importación de datos en Matlab*

Una manera de introducir datos en el software es escribiéndolas directamente en un script. El siguiente código introduce una matriz M_{8x2} y un vector M_{3x2} en la memoria del programa

```
a=[1 1
2 4
3 9
4 16
5 25
6 36
7 49
8 64
```

```
]
b=[ i, 2; 3, 4;5 pi]
```

Pero en general, los datos de las muestras estarán almacenados en fichero externos que deben importarse, para ello se recomienda usar la "Import tool". que se puede activar clickando sobre el icono Import Data que se observa en la figura. Después, se selecciona el fichero y se pude interaccionar con él para seleccionar el conjunto de datos a importar y la forma de almacenamiento dentro de Matlab.



11.1.2 *Variable discreta*

Dada una muestra cuyos datos están el almacenados en el vector Matlab `data`, el comando `tabulate(data)` genera una tabla de frecuencias que incluye las frecuencias absolutas y relativas (ver sección 2). Si se quisieran añadir las frecuencias acumuladas, se puede hacer uso del comando `cumsum`.

Ejemplo: Tabla de frecuencias

```
clear
close all
clc
data=[ -7     5     5     5     7
8     6     8     6     5
5     -7     8     8     8
8     8     6     7     8
7     8     7     8     7];
tabla=tabulate(data(:));
tabla = [tabla, cumsum(tabla(:,2)), cumsum(tabla(:,3))]
```

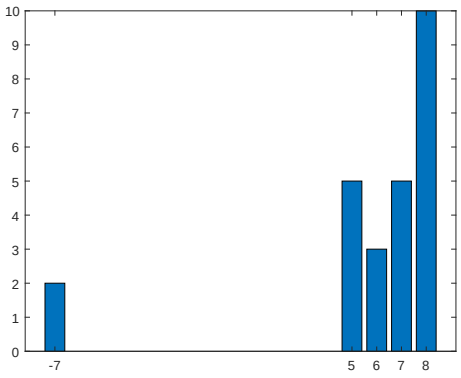
```
tabla =

-7     2     8     2     8
 5     5    20     7    28
 6     3    12    10    40
 7     5    20    15    60
 8    10    40    25   100
```

11.1.3 *Variables discretas, gráficos*

Para representar gráficamente estos datos podemos hacer un diagrama de barras o un diagrama de sectores, pare ello los comandos son respectivamente `bar` y `pie`.

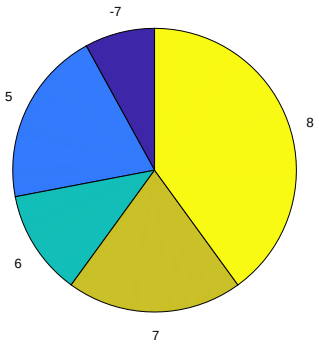
Por ejemplo, para representar las frecuencias absolutas se debe introducirla siguiente instrucción

```
bar(tabla(:,1),tabla(:,2))
```



Para generar el diagrama de sectores, de las frecuencias relativas la siguiente:

```
labels=int2str(tabla(:,1))
labels=cellstr(labels);
pie(tabla(:,2),labels);
```



11.1.4 *Variable continua*

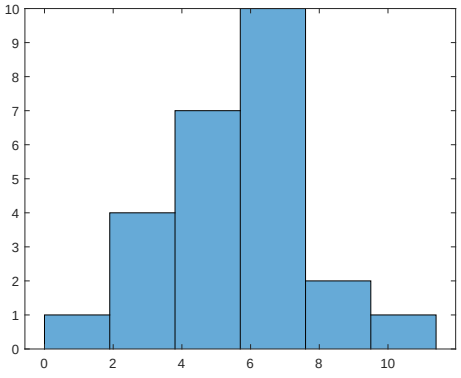
Si los datos corresponden a una variable continua, el comando `histogram` sirve para los dos propósitos: hace una representación gráfica de los mismo (histograma) y proporciona los datos que permiten generar la tabla de frecuencias. Aunque el comando `histcounts` también sirve para obtener estos datos.

Por defecto, la marca de clase para Matlab es el inicio del intervalo, por eso, en el código propuesto se ha añadido el término `h.BinWidth/2` en la definición de x_i.

Histograma y tabla de frecuencias

```
data = [0.80152       6.7553       2.8096        4.021       3.8844
         6.277       7.0672       4.4648       10.949       4.7869
        5.7429       5.8396       5.3731       3.7548       4.5697
        4.2516       6.2021       6.9019       8.8406        5.947
        6.3907        3.652        3.419       6.9223       7.7313];
histogram(data,6);
xi=h.BinEdges(1:end-1)+h.BinWidth/2;
ni=h.Values;
[Ni,Xi]=histcounts(data,6);
tabla=[xi',ni',ni'/sum(ni),cumsum(ni')]
```

```
h =
Histogram with properties:
Data: [5x5 double]
Values: [1 4 7 10 2 1]
NumBins: 6
BinEdges: [0 1.9000 3.8000 5.7000 7.6000 9.5000 11.4000]
BinWidth: 1.9000
BinLimits: [0 11.4000]
Normalization: 'count'
FaceColor: 'auto'
EdgeColor: [0 0 0]
```



La tabla de frecuencias se genera con el código:

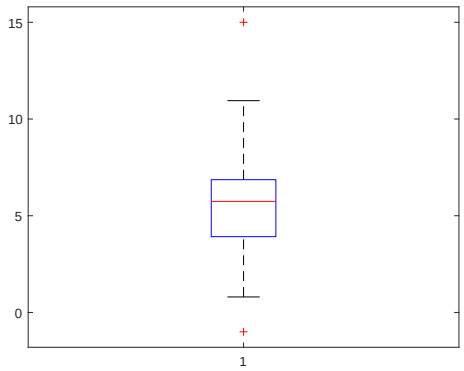
```
tabla=[xi',ni',ni'/sum(ni),cumsum(ni')]
```

```
tabla =
       0.95           1        0.04           1
       2.85           4        0.16           5
       4.75           7        0.28          12
       6.65          10         0.4          22
       8.55           2        0.08          24
      10.45           1        0.04          25
```

Diagrama de caja y bigotes. Box plot

A los datos se le añaden dos puntos atípicos para que se vea, en el diagrama, como son resaltados.

```
boxplot([data(:);15;-1])
```



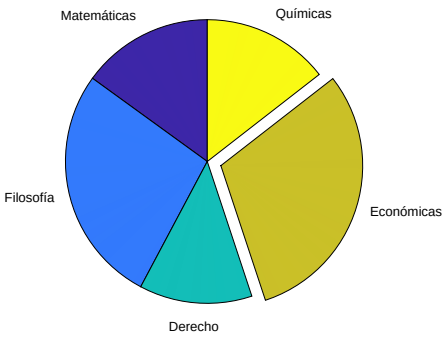
11.1.5 *Variable Cualitativa*

En este caso los gráficos típicos son otra vez el diagrama de sectores y el diagrama de barras. Suponga que los datos de matriculación en una universidad son los siguientes:

Facultad	**Num. Alumnos**
Matemáticas	2136
Filosofía	3870
Derecho	1830
Económicas	4328
Químicas	2060

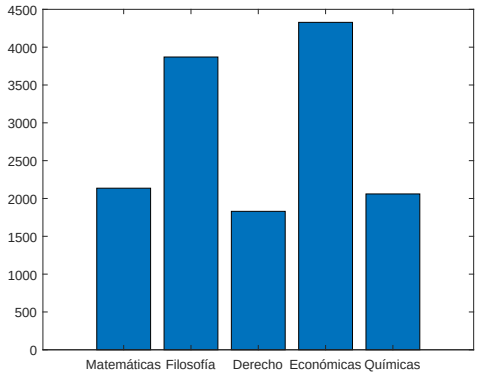
Para representar gráficamente esto datos podemos usar las siguientes instrucciones, en la que se resalta el sector con mayor número de alumnos,

```
data=[2136,3870,1830,4328,2060];
labels={'Matemáticas', 'Filosofía','Derecho','Económicas','Químicas'};
pie (data,[0,0,0,1,0],labels);
```



o bien,

```
bar(data);
set(gca,'XTickLabel',labels)
```



11.1.6 *Medidas de centralización*

Usando los datos una muestra A, los comando para calcular los principales parámetros de centralización y dispersión están en la tabla 11.1

```
data=[ 7     5     5     5     7
8     6     8     6     5
5     7     8     8     8
8     8     6     7     8
7     8     7     8     7];
```

Tabla 11.1: Principales instrucciones de Matlab para calcular los parámetros de centralización y dispersión de una muestra

Nombre	**def nº**	**símbolo**	**comando**	**ejemplo**	**respuesta**
Media	2.10	$\overline{X}$	mean	mean(A(:))	6.8800
Mediana	2.13	M	median	median(A(:))	7
Moda	2.14	M_d	mode	mode(A(:))	8
Cuartiles	2.15	$P_{1/4}\, M\, P_{3/4}$	prctile	prctile(A(:),[25 50 75])	6 7 8
Rango	2.16	R	range	range(A(:))	3
Varinaza	2.17	S^2	var	var(A(:),1)	1.3056
Varianza muestral	9.6	S^2_{n-1}	var	var(A(:))	1.3600
Desviación típica	2.18	S_{n-1}	std	std(A(:))	1.1662
Rango intercuartílico	2.21	IQR	iqr	iqr(A(:))	2

11.2 PROBABILIDAD Y DISTRIBUCIONES DE PROBABILIDAD CON MATLAB

11.2.1 *Combinatoria*

Sobre métodos de conteo, Matlab tiene definidas funciones para calcular el factorial de un número

$$n! = \texttt{factorial(n)}$$

y los números combinatorios

$$\binom{m}{n} = \texttt{nchoosek(m,n)},$$

pero no una función que calcule variaciones de m elementos cogidos de n en n, para ello se recomienda definir una función de Matlab que las calcule, por ejemplo, la función `permn`

```
function s = permn(m,n)
% This function count number of arrangements posibles with m distincts
% elements grouped in groups of size n elements, without repetition, and
% being order important
% $  nPr = P(n,r) = n! / (n-r)!$
switch nargin
case 1
s =factorial(m);
case 2
if m<n    error('n must be an integer between 0 and m.')
else
```

```
s = prod(m:-1:m-n);
end
otherwise
error('Too many input arguments.')
end
```

Si lo que interesa es ver las distintas permutaciones que se pueden hacer con unos elementos se usa el comando `perms`:

```
p=perms(['a' 'b' 'c'])
length(p)
```

```
p =
cba
cab
bca
bac
abc
acb

ans = 6
```

11.2.2 *Generación de números aleatorios*

En Matlab hay varios comandos que permiten generar números aleatorios que sigan una determinada distribución, el más común es el comando `random(name,A,B,[m,n])` que genera una matriz de m filas y n columnas de números aleatorios que signe la distribución *name*, por ejemplo

```
N = 5;
p=0.4;
random('Binomial',N,p,[3,1])
```

genera una vector columna de 3 elementos que siguen una distribución $B(5, 0.4)$.

Las principales distribuciones comentadas en este libro están incluidas dentro de la función `random`, además de la ya comentada `'Binomial'`, están las más comunas `'Normal', 'Exponential', 'Uniform', 'Poisson', 'Discrete Uniform'` y las `'Chisquare', 'F','Gamma', 'Geometric','T'` entre otras.

Además, de la instrucción `random`, para generar un conjunto de n datos aleatorios que sigan una distribución concreta también existen funciones específicas, por ejemplo, para el caso de una distribución Gaussiana $N(\mu, \sigma)$ se puede usar el comando `normrnd(`μ`,`σ`, [1 n])`, y si la normal es tipificada existe el comando `randn([m n])`.

Para las otras distribuciones existen funciones similares: `poissrnd(lambda,[m,n]))`, `binornd(N,p,[m,n])`,...

Los números aleatorios generados por ordenador se usan en el juego, para generar muestras estadísticas, para hacer simulaciones de sistemas físicos, para la criptografía, entre otras aplicaciones.

Ejercicio 11.2.1

En la sección 5.7.2 el capítulo 5 vimos que en el caso de variables aleatorias $X + X \neq 2X$. Se por pone comprobar experimentalmente esta desigualdad. Generar dos muestras, (a, b) de 100 elementos cada una que sigan un distribución cualquiera, por ejemplo una binomial con $N = 20$ y $p = 0.2$. Calcular la media y la varianza de $a + b$ y de $2a$. Comparar los resultados obtenidos.

Solución: El código para hacer este ejercicio podría ser:

```
N = 20;
p=0.2;
a=binornd(20,0.2,[100,1]);
b=binornd(20,0.2,[100,1]);
clc
[M,V]=binostat(20,0.2)
x=[mean(a),var(a,1);mean(b),var(b,1)]
[mean(a+b),var(a+b,1);mean(2*a),var(2*a)
sum(x)
```

se puede observar claramente, que $S^2_{2X} > S^2_{X+X}$

11.3 DISTRIBUCIONES DE PROBABILIDAD

Las funciones de densidad de probabilidad de las distribuciones más comunes también estás definidas en Matlab, como en el caso de la función `random` existe la función `pdf(name,x,A,B,...)` que da el valor de la función densidad de probabilidad de la distribución "name" cuyos parámetros son A, B, ... en el punto x. Pero para las distribuciones típicas existen sus funciones específicas.

Lo mismo se puede decir de la funciones de distribución de probabilidad, en esta caso, la función de Matlab a llamarlas es `cdf(name,x,A,B,...)`

11.3.1 *Distribuciones Discretas*

Las funciones densidad de probabilidad de las distribuciones discretas más comunes también están definidas en Matlab:

- **Binomial**: Si se desea calcular $B(3; 5, 0.4)$ (38) hay que ejecutar:

  `binopdf(3,5,0.4)`

  da como resultado 0.2304, que es lo mismo que se obtendría de ejecutar `nchoosek(5,3)*0.4^3*(1-0.4)^2`.

- **Poisson**: Si se desea calcular $P(3, 5)$ (45) hay que ejecutar:

  `poisspdf(3,5)`

  da como resultado 0.14037, que es lo mismo que se obtendría de ejecutar `exp(-5)*5^3/factorial(3)`.

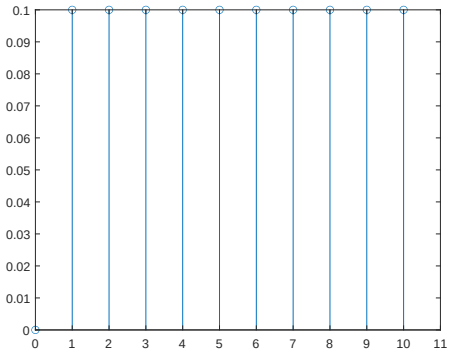
- **Uniforme discreta**: La función `unidpdf(x,n)`, da la probabilidad del evento x, de N.

  Por ejemplo, las instrucciones siguientes:

```
n = 10;
[m,v] =unidstat(n) % media y varianza
x = 0:10;
y = unidpdf(x,n);
figure;
stem(x,y)
h = gca;
h.XLim = [0 11];
```

generan las siguientes salidas

```
m = 5.5000
n= 8.2500
```



La instrucción `[mu,sigma2]=*stat(A,B)` da la media μ y la varianza σ^2 de la distribución "*". Por ejemplo, usar `[mu,sigma2]=poisstat(3)`.

Ejercicio 11.3.1

Usar Matlab para:

a) Calcular y representar gráficamente la función densidad de probabilidad $B(10, 0.4)$.

b) Calcular y representar gráficamente los 50 primeros términos la función densidad de probabilidad de una distribución de Poisson de media 10.

c) Generar un conjunto de 100 números que sigan las distribuciones de los apartados a y b, hacer un histograma, calcular la media, la varianza y comentar los resultados.

d) Calcular $P = B(3; 5000, 0.0004)$ y usar la distribución de Poisson para estimar el valor de P

Solución c) 0.18048 y 0.18045 respectivamente

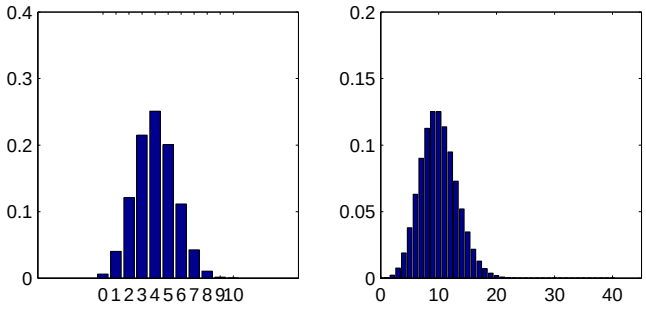


Figura 11.1: Representación gráfica de B(10,0.4) y Poisson(10).

11.3.2 *Distribuciones continuas*

- **Normal**: La función `normpdf(x,`μ,σ`)` da la función densidad de probabilidad de una distribución normal de media μ y desviación típica σ es decir es

$$f = \texttt{@(x) normpdf}(x,\mu,\sigma) \Longleftrightarrow f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

- **Exponencial**: La función `exppdf(x,`μ`)` da la función densidad de probabilidad de una distribución exponencial de media μ es decir es

$$f = \texttt{@(x) exppdf}(x,\mu) \Longleftrightarrow f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \quad x > 0$$

- **Uniforme continua**: La función `unifpdf(x,a,b)` da la probabilidad de la distribución uniforme continua definida en el intervalo $[a,b]$ en el punto x, por ejemplo

```
x = 1:5;
unifpdf(x,2,4)
```

da como salida, el vector

```
0    0.5000    0.5000    0.5000         0
```

Las funciones densidad de probabilidad normalmente se usan, vía integración o suma, para calcular probabilidades, este cálculo está definido dentro de las funciones distribuciones de probabilidad. Las variables típicamente tienen su funciones distribución de probabilidad definidas, por ejemplo, la función `poisscdf(7,lambda)` da la probabilidad $P(X \leq 7)$ siendo X una variable aleatoria de Poisson con media λ. Equivalentemente, `binocdf(B,N,p)-binocdf(A,N,p)` es la probabilidad $P(A < X \leq B)$ donde X es una binomial.

Para el caso de las variables continuas la idea es la misma: para la exponencial, `expcdf(B,beta)-expcdf(A,beta)` $= P(A < X \leq B)$.

Un caso que merece especial atención es el caso de la variable normal, ya que como se ha visto en las secciones 7.2 y 9, la distribución normal es muy usada, por esto se consideran interesantes las siguientes funciones de Matlab:

Tabla 11.2: Funciones de Matlab relacionadas con la normal

Función	¿Qué hace?
`normrnd(`μ`,`σ`, [m n])`	Genera una matriz de m filas n columnas de números aleatorios que siguen una $N(\mu, \sigma)$.
`normpdf(x,`μ, σ`)`	Calcular el el valor de la función densidad de probabilidad $N(\mu, \sigma)$ para $X = x$.
`normcdf(x,`μ, σ`)`	Calcular el el valor de la función distribución de probabilidad $N(\mu, \sigma)$ para $X = x$, i.e., $P\{X \leq x\}$ siendo X, $N(\mu, \sigma)$.
`normspec(`$[x_{inf}, x_{sup}], \mu, \sigma$`)`	Dibuja la curva $N(\mu, \sigma)$ resaltando el área entre el intervalo $[x_{inf}, x_{sup}]$, y da el valor de la misma.
`norminv(P,`μ, σ`)`	Da el valor x tal que $P\{X \leq x\} = P$ siendo X, $N(\mu, \sigma)$.
$[\bar{x}, s, intM, intS]$`=normfit(data)`	Estima la media y la varianza de la muestra `data` de una población normal, y da los intervalos de confianza de las mismas al 95 %.

Como se explica en la tabla 11.2, la instrucción $p =$ `normspec(`$[X_i\, X_s], \mu, \sigma$`)` dibuja el área bajo la normal $N(\mu, \sigma)$ correspondiente a la probabilidad de que $X \in (X_i, X_s)$

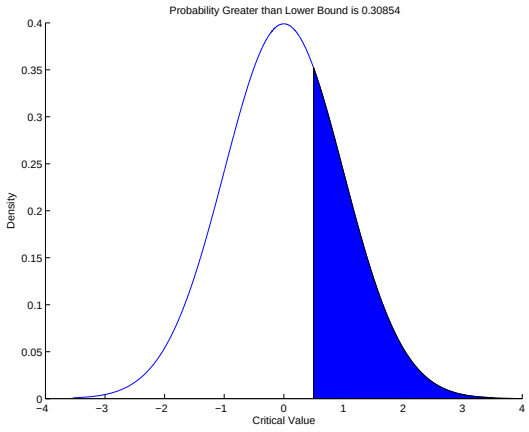


Figura 11.2: Resultado de introducir en Matlab $p =$ `normspec(`$[0.5\, Inf], 0, 1$`)`.

X = `norminv(P,`μ, σ`)` es la inversa de la función `normcdf`, es decir da el punto x de la curva $N(\mu, \sigma)$ tal que el área a su derecha es P.

Por ejemplo, para encontrar lo puntos que delimita una área del 95 % centrada en la curva $N(0, 1)$ la instrucción sería `x = norminv([0.025 0.975],0,1)` y la respuesta es

```
x =  -1.9600    1.9600
```

11.4 ESTIMACIÓN POR INTERVALOS CON MATLAB

Dado un conjunto de datos `data`, para hacer una estimación por intervalos con una confianza de $(1 - \alpha)$. existe la instrucción

$$[\bar{x}, s, \mathrm{I}\mu, \mathrm{I}\sigma] = \texttt{normfit(data,}\alpha\texttt{)}$$

que da, respectivamente, la media muestral, la desviación típica muestral, el intervalo de estimación para la media ($\mathrm{I}\mu$), que son lo números X_i, X_s (calculados usando la distribución t-Student), y el intervalo de confianza de la σ ($\mathrm{I}\sigma$), que se calcula usando la χ^2, tal como se muestra en la tabla 9.2.

Veamos un ejemplo, generamos con ayuda de Matlab una muestra aleatoria de 1000 individuos de una población $N(3, 5)$ y estimamos la media y la desviación estándar con un nivel de confianza del 99 %.

```
x = normrnd(3,5,[1000,1]);
[xb,s,muCI,sigmaCI] = normfit(x,0.01)
```

```
xb = 2.8184
s = 4.9164
muCI =
2.4172
3.2197
sigmaCI =
4.6476
5.2158
```

Este mismo cálculo se pude hacer usando directamente las expresiones de la tabla 9.2. Para ello primero hay que realizar una estimación puntual de la media y la desviación típica. Esto, en Matlab, se hace usando las instrucciones ya vistas `mean` y `std`, o el comando `mle` que da la estimación usando el estimador de máxima verosimilitud, esta última instrucción admite indicar cual es la distribución teórica de la población, en caso de no indicarlo, se supone que es normal.

```
xb=mean(x);
s = std(x);
ph=mle(x)
```

de modo que, usando la distribución t, el intervalo para la media sería:

```
t=tinv(1-al/2,n-1)
e = t*s/sqrt(n)
mus = xb-e
mui = xb+e
```

y usando la distribución normal:

```
z=norminv(1-al/2)
xb+[-z,z]*s/sqrt(n)
```

Evidentemente, como n es muy grande, ambos resultados coinciden y coinciden con el resultado dado por `normfit`.

Finalmente, para estimar la varianza, se podría usar el siguiente código

```
xi2i=chi2inv(al/2,n-1)
xi2s=chi2inv(1-al/2,n-1)
sigmai = sqrt((n-1)*s^2/xi2s)
sigmas = sqrt((n-1)*s^2/xi2i)
```

Comentario 10 *Es importante notar que en el capítulo 9 se ha indicado que z_α, t_α y χ_α corresponde al punto de la curva cuya área a la derecha él es α, sin embargo, las funciones de Matlab* `norminv(α)`, `tinv(α,ν)` *y* `chi2inv(α, n-1)` *dan el punto cuya área es α a su izquierda.*

Del mismo modo que se ha hecho para la media y la desviación típica de una sola muestra, código similar se pude aplicar para estimar la diferencia de medias o el cociente de varianzas, para ello será necesarias la función de Matlab

`X = finv(P,`ν_1`,`ν_2`)`

que calcula la inversa de la función distribución de probabilidad de la F de Snedecor de grados de libertad ν_1 y ν_2. Sin embargo, tal como se comentó en el capítulo 10, el contraste de hipótesis y la estimación por intervalos están muy relacionados, de modo que para hacer estimación de diferencias de medias o cocientes de varianzas quizás sea más útil usar las funciones comentadas en la siguiente sección.

11.5 CONTRASTES DE HIPÓTESIS CON MATLAB

En Matlab hay varias funciones destinadas a hacer contraste de hipótesis, en esta sección se comentan algunas de ellas, ver tabla 11.3, pero se recomienda usar la ayuda del propio programa para ver otras opciones y más detalles.

Tabla 11.3: Funciones de Matlab para realizar contraste de hipotesis.

Función	¿Qué hace?
`ztest(x,`μ_0,σ`)`	Indica rechace o no de $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
`ttest(x,`μ_0`)`	$H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. Para test de hipótesis con sigma desconocida, y para datos pareados.
`ttest2(x,y)`	$H_0 : \mu_1 - \mu_2 = 0$ vs. $H_1 : \mu_1 - \mu_2 \neq 0$. Por defecto supone varianzas iguales.
`vartest(x,`σ_0^2`)`	$H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 \neq \sigma_0^2$.
`vartest2(x,y)`	$H_0 : \sigma_X^2/\sigma_Y^2 = 1$ vs. $H_1 : \sigma_X^2/\sigma_Y \neq 1$.

Por defecto, el nivel significación es del 5 %, pero todas las funciones admiten la opción de cambiar H_1 e indicar otro nivel de significación. Además, si se pide, también dan el p-valor, el intervalo de confianza correspondiente y el valor experimental obtenido al hacer el test.

Contrastes sobre la media de una población

A modo de ejemplo consideremos un conjunto de alumnos cuya nota en estadística sigue una distribución normal de media 6 y desviación típica 1.2. El siguiente código genera una muestra.

```
clear
n=100;
x= normrnd(6,1.2,n,1);
```

Para contrastar $H_0 : \mu = 5$ vs. $H_1 : \mu \neq 5$ con un nivel de significación del 0.01 debería ejecutarse:

```
s=std(x);
[h,p,ci,zval]=ztest(x,5,s,alpha)
```

El resultado es

```
h = 1
p = 4.3538e-10
ci =
5.4895
6.1777
zval =6.2408
```

Que indica que: (`h=1`) se rechaza H_0, el p-valor es $4.3{\cdot}10^{-10}$, el intervalo de confianza para a media es [5.49,6.18] y que $z_{exp} = 6.24$.

Para hacer un contraste unilateral, por ejemplo $H_0 : \mu \geq 5$ vs. $H_1 : \mu < 5$, con un nivel de significación del 5 %, la instrucción a ejecutar debe ser

`[h,p,ci,zval]=ztest(x,5,s,'Tail','left')` que nos arroja los siguientes resultados:

```
h = 0
p = 1
ci =
-Inf
6.0533
zval =6.2408
```

Donde se ve, que claramente se acepta H_0.

En este ejemplo se ha tenido en cuenta que como $n >> 30$ $\sigma \approx s$, pero se podría hacer el mismo test usando la distribución t, con el comando

`[h,pvalue,ci,z]=ttest(x,5,'Alpha',alpha)`

que en este caso, arroja el mismo resultado que se ha obtenido con la normal.

Finalmente, el comando `ztest` también funciona si en lugar de tener toda la muestra, se tiene sólo el valor de $\bar{x}$ y s o σ, pero en este caso habría que llamar a la función así:

`[h,pvalue,ci,z]=ztest(xb,5,s/sqrt(n),'Tail','right').`

Diferencia de medias

Un ejemplo de uso de la función `ttest2` sería el siguiente:

```
x=normrnd(5,1.2,[10,1]);
y=normrnd(5.5,1.1,[10,1]);
[h,pvalue,ci,z]=ttest2(x,y)
[h,pvalue,ci,z]=ttest2(x,y,'Vartype','unequal')
```

Donde se realiza un contraste de semejanza de medias suponiendo varianzas igual y distintas sobre dos muestras que siguen poblaciones N(5,1.2) y N(5.5,1.1) respectivamente.

Cociente de varianzas

El siguiente ejemplo simula que se tienen las notas de dos clases de 20 y 25 alumnos, los de la primera clase siguen una distribución normal de media 5 y desviación típica 0.8, mientras que los de la segunda clase siguen una N(6,1.3). Se contrasta la hipótesis $H_0 : \sigma_1^2/\sigma_2^2 = 1$ vs. $H_1 : \sigma_1^2/\sigma_2 \neq 1$.

```
x=normrnd(5,0.8,[20,1]);
y=normrnd(6,1.3,[25,1]);
[h,pvalue]= vartest2(x,y)
```

Como el resultado es

```
h = 1
pvalue = 0.0126
```

se puede concluir que las varianzas no son iguales y además la confianza el 98.7 % en que la decisión no es equivocada.

Ejercicio 11.5.1

Antes de sacar a la venta unas píldoras para adelgazar, la compañía farmacéutica norteamericana, realiza un test sobre un grupo de 64 personas. Antes del test se sabe que el peso de la población sigue una distribución normal de media 95 kg con una desviación típica de 12 kg. Diseñar un experimento de contraste de hipótesis basado en la media muestral que permita determinar si las píldoras tienen un efecto relevante. Nivel de significación $\alpha = 0.01$.

Solución:

Evidentemente se tratará de un test unilateral, con $H_0 : \mu \geq 95$ y hipótesis alternativa $H_1 : \mu < 95$.

Como $n > 30$ y además se supone que σ es conocida, se puede usar que $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$.

Usando Matlab este ejercicio se resuelve con las instrucciones

```
mu=95
sigma=12/sqrt(64)
x = norminv([0.01,1],mu,sigma)
z=norminv([0.01,1],0,1)
z = normspec([x],mu,sigma)
```

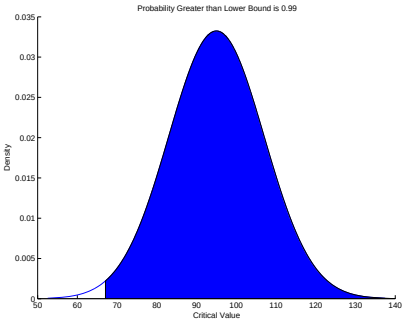
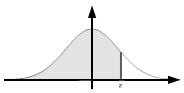


Figura 11.3: El área sombreada representa una probabilidad del 99 %.

A

DISTRIBUCIÓN NORMAL TIPIFICADA N(0,1)

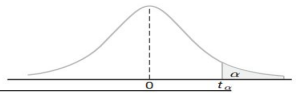
$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\, dx \qquad \text{para} \quad z \geq 0$$



z	0'00	0'01	0'02	0'03	0'04	0'05	0'06	0'07	0'08	0'09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92786	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96637	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99897	0.99900
3.1	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929
3.2	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950
3.3	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965
3.4	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976
3.5	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983
3.6	0.99984	0.99985	0.99985	0.99986	0.99986	0.99987	0.99987	0.99988	0.99988	0.99989
3.7	0.99989	0.99990	0.99990	0.99990	0.99991	0.99991	0.99991	0.99992	0.99992	0.99992
3.8	0.99993	0.99993	0.99993	0.99994	0.99994	0.99994	0.99994	0.99995	0.99995	0.99995
3.9	0.99995	0.99995	0.99996	0.99996	0.99996	0.99996	0.99996	0.99996	0.99997	0.99997
4.0	0.99997	0.99997	0.99997	0.99997	0.99997	0.99997	0.99998	0.99998	0.99998	0.99998
4.1	0.99998	0.99998	0.99998	0.99998	0.99998	0.99998	0.99998	0.99998	0.99999	0.99999

B

VALORES CRÍTICOS DE LA DISTRIBUCIÓN t



ν	α 0,25	0,2	0,15	0,1	0,05	0,025	0,01	0,005	0,0005
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656	636,619
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,600
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,768
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,689
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,660
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	3,460
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,290

C

VALORES CRÍTICOS DE LA DISTRIBUCIÓN χ^2_ν

	α										
ν	0.001	0.005	0.010	0.025	0.050	0.100	0.125	0.200	0.250	0.333	0.500
1	0.000	0.000	0.000	0.001	0.004	0.016	0.025	0.064	0.102	0.186	0.455
2	0.002	0.010	0.020	0.051	0.103	0.211	0.267	0.446	0.575	0.811	1.386
3	0.024	0.072	0.115	0.216	0.352	0.584	0.692	1.005	1.213	1.568	2.366
4	0.091	0.207	0.297	0.484	0.711	1.064	1.219	1.649	1.923	2.378	3.357
5	0.210	0.412	0.554	0.831	1.145	1.610	1.808	2.343	2.675	3.216	4.351
6	0.381	0.676	0.872	1.237	1.635	2.204	2.441	3.070	3.455	4.074	5.348
7	0.598	0.989	1.239	1.690	2.167	2.833	3.106	3.822	4.255	4.945	6.346
8	0.857	1.344	1.646	2.180	2.733	3.490	3.797	4.594	5.071	5.826	7.344
9	1.152	1.735	2.088	2.700	3.325	4.168	4.507	5.380	5.899	6.716	8.343
10	1.479	2.156	2.558	3.247	3.940	4.865	5.234	6.179	6.737	7.612	9.342
11	1.834	2.603	3.053	3.816	4.575	5.578	5.975	6.989	7.584	8.514	10.341
12	2.214	3.074	3.571	4.404	5.226	6.304	6.729	7.807	8.438	9.420	11.340
13	2.617	3.565	4.107	5.009	5.892	7.042	7.493	8.634	9.299	10.331	12.340
14	3.041	4.075	4.660	5.629	6.571	7.790	8.266	9.467	10.165	11.245	13.339
15	3.483	4.601	5.229	6.262	7.261	8.547	9.048	10.307	11.037	12.163	14.339
16	3.942	5.142	5.812	6.908	7.962	9.312	9.837	11.152	11.912	13.083	15.338
17	4.416	5.697	6.408	7.564	8.672	10.085	10.633	12.002	12.792	14.006	16.338
18	4.905	6.265	7.015	8.231	9.390	10.865	11.435	12.857	13.675	14.931	17.338
19	5.407	6.844	7.633	8.907	10.117	11.651	12.242	13.716	14.562	15.859	18.338
20	5.921	7.434	8.260	9.591	10.851	12.443	13.055	14.578	15.452	16.788	19.337
21	6.447	8.034	8.897	10.283	11.591	13.240	13.873	15.445	16.344	17.720	20.337
22	6.983	8.643	9.542	10.982	12.338	14.041	14.695	16.314	17.240	18.653	21.337
23	7.529	9.260	10.196	11.689	13.091	14.848	15.521	17.187	18.137	19.587	22.337
24	8.085	9.886	10.856	12.401	13.848	15.659	16.351	18.062	19.037	20.523	23.337
25	8.649	10.520	11.524	13.120	14.611	16.473	17.184	18.940	19.939	21.461	24.337
26	9.222	11.160	12.198	13.844	15.379	17.292	18.021	19.820	20.843	22.399	25.336
27	9.803	11.808	12.879	14.573	16.151	18.114	18.861	20.703	21.749	23.339	26.336
28	10.391	12.461	13.565	15.308	16.928	18.939	19.704	21.588	22.657	24.280	27.336
29	10.986	13.121	14.256	16.047	17.708	19.768	20.550	22.475	23.567	25.222	28.336
30	11.588	13.787	14.953	16.791	18.493	20.599	21.399	23.364	24.478	26.165	29.336
35	14.688	17.192	18.509	20.569	22.465	24.797	25.678	27.836	29.054	30.894	34.336
40	17.916	20.707	22.164	24.433	26.509	29.051	30.008	32.345	33.660	35.643	39.335
45	21.251	24.311	25.901	28.366	30.612	33.350	34.379	36.884	38.291	40.407	44.335
50	24.674	27.991	29.707	32.357	34.764	37.689	38.785	41.449	42.942	45.184	49.335
60	31.738	35.534	37.485	40.482	43.188	46.459	47.680	50.641	52.294	54.770	59.335

	α										
ν	0.600	0.667	0.750	0.800	0.875	0.900	0.950	0.975	0.990	0.995	0.999
1	0.708	0.936	1.323	1.642	2.354	2.706	3.841	5.024	6.635	7.879	10.828
2	1.833	2.197	2.773	3.219	4.159	4.605	5.991	7.378	9.210	10.597	13.816
3	2.946	3.405	4.108	4.642	5.739	6.251	7.815	9.348	11.345	12.838	16.266
4	4.045	4.579	5.385	5.989	7.214	7.779	9.488	11.143	13.277	14.860	18.467
5	5.132	5.730	6.626	7.289	8.625	9.236	11.070	12.833	15.086	16.750	20.515
6	6.211	6.867	7.841	8.558	9.992	10.645	12.592	14.449	16.812	18.548	22.458
7	7.283	7.992	9.037	9.803	11.326	12.017	14.067	16.013	18.475	20.278	24.322
8	8.351	9.107	10.219	11.030	12.636	13.362	15.507	17.535	20.090	21.955	26.125
9	9.414	10.215	11.389	12.242	13.926	14.684	16.919	19.023	21.666	23.589	27.877
10	10.473	11.317	12.549	13.442	15.198	15.987	18.307	20.483	23.209	25.188	29.588
11	11.530	12.414	13.701	14.631	16.457	17.275	19.675	21.920	24.725	26.757	31.264
12	12.584	13.506	14.845	15.812	17.703	18.549	21.026	23.337	26.217	28.300	32.910
13	13.636	14.595	15.984	16.985	18.939	19.812	22.362	24.736	27.688	29.819	34.528
14	14.685	15.680	17.117	18.151	20.166	21.064	23.685	26.119	29.141	31.319	36.123
15	15.733	16.761	18.245	19.311	21.384	22.307	24.996	27.488	30.578	32.801	37.697
16	16.780	17.840	19.369	20.465	22.595	23.542	26.296	28.845	32.000	34.267	39.252
17	17.824	18.917	20.489	21.615	23.799	24.769	27.587	30.191	33.409	35.718	40.790
18	18.868	19.991	21.605	22.760	24.997	25.989	28.869	31.526	34.805	37.156	42.312
19	19.910	21.063	22.718	23.900	26.189	27.204	30.144	32.852	36.191	38.582	43.820
20	20.951	22.133	23.828	25.038	27.376	28.412	31.410	34.170	37.566	39.997	45.315
21	21.991	23.201	24.935	26.171	28.559	29.615	32.671	35.479	38.932	41.401	46.797
22	23.031	24.268	26.039	27.301	29.737	30.813	33.924	36.781	40.289	42.796	48.268
23	24.069	25.333	27.141	28.429	30.911	32.007	35.172	38.076	41.638	44.181	49.728
24	25.106	26.397	28.241	29.553	32.081	33.196	36.415	39.364	42.980	45.559	51.179
25	26.143	27.459	29.339	30.675	33.247	34.382	37.652	40.646	44.314	46.928	52.620
26	27.179	28.520	30.435	31.795	34.410	35.563	38.885	41.923	45.642	48.290	54.052
27	28.214	29.580	31.528	32.912	35.570	36.741	40.113	43.195	46.963	49.645	55.476
28	29.249	30.639	32.620	34.027	36.727	37.916	41.337	44.461	48.278	50.993	56.892
29	30.283	31.697	33.711	35.139	37.881	39.087	42.557	45.722	49.588	52.336	58.301
30	31.316	32.754	34.800	36.250	39.033	40.256	43.773	46.979	50.892	53.672	59.703
35	36.475	38.024	40.223	41.778	44.753	46.059	49.802	53.203	57.342	60.275	66.619
40	41.622	43.275	45.616	47.269	50.424	51.805	55.758	59.342	63.691	66.766	73.402
45	46.761	48.510	50.985	52.729	56.052	57.505	61.656	65.410	69.957	73.166	80.077
50	51.892	53.733	56.334	58.164	61.647	63.167	67.505	71.420	76.154	79.490	86.661
60	62.135	64.147	66.981	68.972	72.751	74.397	79.082	83.298	88.379	91.952	99.607

BIBLIOGRAFÍA

BIPM, IEC, ISO LFCC y OIML IUPAC (1995). "Guide to Expression of Uncertainty in Measurement (Corrected and Reprinted GUM)". En: *Geneve Switerland: ISO*.

Canavos, G.C., E.G.U. Medal y G.J.V. Ramírez (1987). *Probabilidad y estadística: aplicaciones y métodos*. McGraw-Hill México.

Cao, R. et al. (2001). *Introducción a la estadística y sus aplicaciones*. Ediciones Piramide.

Cuadras, C.M. ((Barcelona, 1990)). *Problemas de probabilidades y estadística* VOL. 1. PPU.

Gutiérrez Pulido, Humberto y Román De la Vara Salazar (2008). "Análisis y diseño de experimentos". En.

Masoliver, J. y J. Wagensberg (1996). *Introducció a la teoria de la probabilitat i de la informació*. Edicions Proa.

MathWorks (2008). *Symbolic Math Toolbox 5 MuPAD® Tutorial*. The MathWorks, Inc.

Mitchell, Tom M et al. (1997). *Machine learning. 1997*. Ed. por Burr Ridge. Vol. 45. 37. McGraw Hill.

Montgomery, Duglas C. y George C. Runger ((Mexico, 2000)). *Probabilidad y estadística aplicadas a la inegeniería*. McGraw-Hill.

Murray, Spiegel R. y L. Abellana (1992). *Fórmulas y tablas de matemática aplicada*. McGraw-Hill.

Narvaiza, J.l. et al. (2001a). *Estadística aplicada a la gestión y a las ciencias sociales. Estadística Descriptiva y Probabilidad*. dsclée de brouwer S.A.

— (2001b). *Estadística aplicada a la gestión y a las ciencias sociales. Inferencia Estadística*. dsclée de brouwer S.A.

Navidi, William Cyrus (2006). *Estadística para ingenieros y científicos*. McGraw-Hill.

Papoulis, Athanasios (1991). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill.

Papoulis, Athanasios y S. Unnikrishna Pillai (2002). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill.

Peebles, Z. Peyton (2001). *Probability, Random Variables, and Random signal Principles*. McGraw-Hill.

Proakis, J.G. y M. Salehi (2002). *Communication systems enginering*. Prentice Hall.

Quesada, V., A. Isidoro y L. A. Lopez (2000). *Curso y ejercicios de estadística*. Alhambra longman.

Ross, Sheldon M. (1996). *Stochastic Processes, 2nd Edition*. John Wiley & Sons, Inc.

Rovelli, Carlo (2021). *Helgoland*. Flammarion.

Spiegel, Murray R. ((Mexico, 1998)). *Estadística*. McGraw-Hill.

Taylor, John (1997). *Introduction to error analysis, the study of uncertainties in physical measurements*. Vol. 1.

Troconiz, A. Fz. de (1993). *Probabilidades estadística muestreo*. Tébar Flores, S.L.

Walpole, Ronald E et al. (2012). *Probabilidad y estadística para ingeniería y ciencias*. 9.ª ed. Pearson Educación. ISBN: 978-607-32-1417-9.

Yates, Roy D. y David J. Goodman (1999). *Probability and stochastic processes: a friendly introduction for electrical & computer engineers*. John Willey & Sons.

ÍNDICE ALFABÉTICO

Mondragon Unibertsitatea

Escuela Politécnica Superior

A cualquier ingeniero, a lo largo de su carrera profesional, se le exigirá que tome decisiones racionales, en el sentido que deben ser tomadas en base a una serie de datos experimentales. La Estadística es la ciencia que estudia la recopilación, presentación, análisis y uso de datos para tomar decisiones y resolver problemas. Pero para poder tomar decisiones a partir de un conjunto de datos es necesario todo un conjunto de herramientas matemáticas que son en sí un campo de las matemáticas: la teoría de la probabilidad, que es aquella que estudia los fenómenos aleatorios. En este libro se introducen nociones de estadística descriptiva que permiten analizar y presentar un conjunto de datos, así como los principales conceptos de probabilidad y variables aleatorias que permiten inferir de estos datos conclusiones que ayuden al ingeniero a tomar decisiones y a justificarlas con rigor.