Dra. Josefa Marín Fernández

Departamento de Estadística e Investigación Operativa

Universidad de Murcia

Estadística. Teoría, problemas y prácticas

Grado en Información y Documentación

Curso 2011-12

Contenidos

1

Tabulación y representación gráfica de los datos

1.1. Desarrollo de los contenidos fundamentales

1.1.1. Introducción a la Estadística

Estadística: ciencia que se ocupa de recoger, clasificar, representar y resumir los datos de muestras, y de hacer inferencias (extraer conclusiones) acerca de las poblaciones de las que éstas proceden.

1. *Estadística descriptiva*: parte de la estadística que se ocupa de recoger, clasificar, representar y resumir los datos de las muestras.
2. *Estadística inferencial*: parte de la estadística que se ocupa de llegar a conclusiones (inferencias) acerca de las poblaciones a partir de los datos de las muestras extraídas de ellas.

CONCEPTOS GENERALES:

- *Población*: conjunto de individuos con propiedades comunes sobre los que se realiza una investigación de tipo estadístico.
- *Muestra*: subconjunto de la población.
- *Tamaño muestral*: número de individuos que forman la muestra.
- *Muestreo*: proceso de obtención de muestras representativas de la población.
- *Variable*: propiedad o cualidad que puede manifestarse bajo dos o más formas distintas en un individuo de una población.
- *Modalidades, categorías o clases*: distintas formas en que se manifiesta una variable.
- Las variables se clasifican en:

1. *Cuantitativas*: se expresan numéricamente. Se clasifican en:
   a) *Discretas*: toman valores numéricos aislados, por lo que, fijados dos consecutivos, no pueden tomar ningún valor intermedio.
   b) *Continuas*: pueden tomar cualquier valor dentro de unos límites, por lo que entre dos valores cualesquiera, por próximos que sean, siempre pueden encontrarse valores intermedios.
2. *Cualitativas*: no se expresan numéricamente. Se clasifican en:
   a) *Ordinales*: admiten una ordenación de menor a mayor aunque sus resultados no son numéricos.
   b) *Nominales*: no admiten una ordenación de menor a mayor.

1.1.2. Tabulación de los datos

Los datos se agrupan en clases si son cualitativos o discretos, o en intervalos de clase (de igual longitud, generalmente) si son continuos (o discretos con muchos valores distintos).

- Número adecuado de intervalos: $k = 1 + 3'322 \log n$, siendo n el número total de datos. Si los datos no están agrupados en intervalos, también denotaremos por k al número de datos (o de categorías) diferentes.
- *Amplitud* del intervalo de clase $(\ell_i, \ell_{i+1}]$: $d_i = \ell_{i+1} - \ell_i$.
- *Marca de clase* del intervalo $(\ell_i, \ell_{i+1}]$: $x_i = \dfrac{\ell_i + \ell_{i+1}}{2}$.
- *Frecuencia absoluta* de la clase i-ésima: f_i=número de observaciones contenidas dentro de ella.
- *Frecuencia relativa* o *proporción* de la clase i-ésima: $h_i = \dfrac{f_i}{n}$.
- *Porcentaje* de la clase i-ésima: $\%_i = 100\, h_i$.
- *Frecuencia acumulada absoluta* o *frecuencia absoluta acumulada* de la clase i-ésima: $F_i = f_1 + f_2 + \cdots + f_i$.
- *Frecuencia acumulada relativa* o *frecuencia relativa acumulada* o *proporción acumulada* de la clase i-ésima: $H_i = h_1 + h_2 + \cdots + h_i = \dfrac{F_i}{n}$.
- *Distribución de frecuencias*: tabla conteniendo las distintas clases y las frecuencias correspondientes a cada una de ellas.

1.1.3. Representaciones gráficas

1. Variables cualitativas
   a) *Diagrama de barras*: se sitúan en el eje horizontal las clases y sobre cada una de ellas se levanta un segmento rectilíneo (o un rectángulo) de altura igual a la frecuencia (absoluta o relativa) o al porcentaje de cada clase.

b) *Gráfico de sectores*: se divide el área de un círculo en sectores circulares de ángulos proporcionales a las frecuencias absolutas de las clases.

2. Variables cuantitativas con datos no agrupados en intervalos

   a) *Diagrama de barras*: se sitúan en el eje horizontal los diferentes resultados de la variable y sobre cada uno de ellos se levanta un segmento rectilíneo de altura igual a la frecuencia (absoluta o relativa) o al porcentaje de cada resultado.

   b) *Polígono de frecuencias*: se sitúan los puntos que resultan de tomar en el eje horizontal los distintos valores de la variable y en el eje vertical sus correspondientes frecuencias (no acumuladas), uniendo después los puntos mediante segmentos rectilíneos.

   c) *Gráfico de frecuencias acumuladas*: es la representación gráfica de las frecuencias acumuladas, para todo valor numérico. Siempre es una gráfica en forma de *escalera*.

3. Variables cuantitativas con datos agrupados en intervalos

   a) *Histograma*: se sitúan en el eje horizontal los intervalos de clase y sobre cada uno se levanta un rectángulo de área igual o proporcional a la frecuencia absoluta.

   b) *Polígono de frecuencias*: se sitúan los puntos que resultan de tomar en el eje horizontal las marcas de clase de los intervalos y en el eje vertical sus correspondientes frecuencias (no acumuladas), uniendo después los puntos mediante segmentos rectilíneos.

   c) *Gráfico de frecuencias acumuladas*: es la representación gráfica de las frecuencias acumuladas para todo valor numérico, teniendo en cuenta que dentro de cada intervalo de clase se supone que el número de observaciones se distribuye uniformemente. Siempre es un polígono.

1.2. Ejemplos que se van a resolver en clase

En este tema vamos a utilizar los resultados de las tres variables siguientes: **sexo**, **edad** y **altura**, en metros, observadas en todos/as los/as alumnos/as que han asistido a clase el primer día.

Ejemplo 1.1. Con los datos de la variable sexo:

a) Determinar la distribución de frecuencias absolutas.

b) Determinar la distribución de frecuencias relativas (o proporciones).

c) Determinar la distribución de porcentajes.

Ejemplo 1.2. Con los datos de la variable edad:

a) Determinar la distribución de frecuencias absolutas, frecuencias relativas y porcentajes.

b) Determinar la distribución de frecuencias acumuladas absolutas.

c) Determinar la distribución de frecuencias acumuladas relativas (o proporciones acumuladas).

d) Determinar la distribución de porcentajes acumulados.

Ejemplo 1.3. Con los datos de la variable altura:

a) Agrupar los datos en intervalos de la misma amplitud.

b) A partir de la agrupación anterior determinar la distribución de frecuencias absolutas, relativas, acumuladas absolutas y acumuladas relativas.

Ejemplo 1.4. Dibujar el diagrama de barras de frecuencias absolutas de los datos de la variable sexo.

Ejemplo 1.5. La siguiente tabla muestra el país de procedencia de los documentos primarios de los resúmenes contenidos en un determinado volumen de las tres revistas siguientes: Computer Abstracts, Lead Abstracts y Sociological Abstracts. Dibujar el diagrama de barras conjunto de frecuencias absolutas.

Tabla 1.4

país de procedencia	Computer Abstracts	Lead Abstracts	Sociological Abstracts
Países Bajos	42	34	22
Francia	55	7	76
Alemania	162	37	14
Gran Bretaña	310	147	24
EEUU	966	265	552
Rusia	191	37	42
Otros	265	79	239
suma	1.991	606	969

Ejemplo 1.6. Dibujar el gráfico de sectores de los datos de la variable sexo.

Ejemplo 1.7. Dibujar el diagrama de barras de frecuencias absolutas de los datos de la variable edad.

Ejemplo 1.8. Dibujar el polígono de frecuencias relativas de los datos de la variable edad.

Ejemplo 1.9. Dibujar el gráfico de frecuencias acumuladas absolutas de los datos de la variable edad.

Ejemplo 1.10. Dibujar el histograma de los datos de la variable altura agrupados en intervalos de la misma amplitud.

Ejemplo 1.11. Dibujar el polígono de frecuencias absolutas de los datos de la variable altura agrupados en intervalos de la misma amplitud.

Ejemplo 1.12. Dibujar el gráfico de frecuencias acumuladas absolutas de los datos de la variable altura agrupados en intervalos de la misma amplitud.

1.3. Actividades de aplicación de los contenidos

1.3.1. Problemas propuestos

Problema 1.1. El gasto de una biblioteca, en euros, durante un año determinado, es:

Gasto en personal	6.570
Gasto en libros	3.450
Otros gastos	2.380

Hacer un diagrama de barras de frecuencias absolutas y un gráfico de sectores.

Problema 1.2. Una biblioteca contiene una cantidad de estantes de libros en varios idiomas tal como muestra la siguiente tabla:

Idioma	Nº de estantes
Francés	78
Alemán	47
Ruso	20
Español	30

Determinar la distribución de frecuencias relativas. Hacer un diagrama de barras de frecuencias relativas y un gráfico de sectores.

Problema 1.3. La estadística de fotocopias de una biblioteca, durante un año determinado, es la siguiente:

Reproducción de catálogos	16.110
Trabajo del personal de la biblioteca	63.350
Préstamo interbibliotecario	2.600
Copias para usuarios de la biblioteca	43.540

Determinar la distribución de porcentajes. Hacer un diagrama de barras de porcentajes y un gráfico de sectores.

Problema 1.4. La estadística de fotocopias de 4 bibliotecas (A, B, C y D), durante un año, está recogida en la siguiente tabla:

	A	B	C	D
Reproducción de catálogos	16.110	3.640	0	3.400
Trabajo del personal de la biblioteca	63.350	11.360	3.080	5.500
Préstamo interbibliotecario	2.600	1.090	560	250
Copias para usuarios de la biblioteca	43.540	58.040	1.980	0

Hacer un diagrama de barras conjunto de frecuencias absolutas.

Problema 1.5. El número de citas en diferentes campos de investigación y en distintos años viene dado en la tabla siguiente:

	1970	1980	1990
Sociología	330	414	547
Economía	299	393	295
Política	115	357	137
Psicología	329	452	258

Hacer un diagrama de barras conjunto de frecuencias relativas.

Problema 1.6. El número de palabras clave (*keywords*) de 72 artículos de investigación viene dado por:

Nº de palabras clave	3	4	5	6	7	8	9	10	11	12	13	14
Nº de artículos	5	8	12	7	9	9	10	5	3	2	1	1

Hacer un diagrama de barras de frecuencias absolutas.

Problema 1.7. La altura, en centímetros, de una colección de libros es la siguiente:

Altura	15	16	17	18	19	20	21	22	23	24	25	26	27
Nº de libros	1	0	3	4	4	2	4	5	2	2	2	1	1

Determinar la distribución de frecuencias relativas y hacer un polígono de frecuencias relativas.

Problema 1.8. El número de palabras por línea de una página de un libro viene dado por:

Nº de palabras por línea	4	5	8	9	10	11	12	13	14	15	16	17
Nº de líneas	1	1	2	3	2	7	11	14	3	2	1	1

Determinar la distribución de frecuencias acumuladas absolutas y hacer el gráfico de frecuencias acumuladas absolutas.

Problema 1.9. Los siguientes datos corresponden al número de palabras por resumen (*abstract*) de los artículos científicos de autores españoles que han publicado en una determinada revista de investigación durante un año concreto:

10	15	16	20	17	19	21	14	13	19
11	14	17	19	20	20	22	15	13	12
12	15	17	19	18	23	22	17	21	20
15	18	16	18	12	17	14	15	17	15

Determinar la distribución de frecuencias absolutas, relativas, acumuladas absolutas y acumuladas relativas. Hacer un diagrama de barras de frecuencias absolutas, un polígono de frecuencias relativas y un gráfico de frecuencias acumuladas relativas.

Problema 1.10. Los siguientes datos agrupados en intervalos se refieren al número de llamadas telefónicas recibidas en el servicio de información de una biblioteca pública durante 45 días elegidos al azar:

Nº de llamadas	(9,15]	(15,21]	(21,27]	(27,33]	(33,39]	(39,45]	(45,51]
Nº de días	2	4	8	14	10	6	1

Dibujar el histograma, el polígono de frecuencias y el gráfico de frecuencias acumuladas absolutas.

Problema 1.11. El número de socios de 84 bibliotecas públicas viene dado por:

1.995	1.050	2.500	3.000	3.000	1.500	2.500
995	995	3.000	3.000	1.200	1.450	2.500
2.750	3.000	1.600	3.000	2.250	2.750	1.800
1.250	3.250	1.800	1.750	3.250	2.100	4.500
2.100	995	3.500	2.500	1.700	2.100	1.250
3.500	3.250	1.200	950	3.250	1.700	3.000
1.500	3.500	1.500	995	2.750	3.500	2.150
1.750	2.000	2.200	1.750	2.800	750	2.000
1.500	3.500	4.500	1.950	3.000	2.200	1.600
1.200	2.400	750	1.850	2.400	1.250	3.000
800	2.750	4.000	2.050	5.500	3.750	950
995	3.750	1.500	1.800	1.200	2.500	1.250

Aunque la variable es cuantitativa discreta, se desea agrupar los datos en intervalos de la misma amplitud. A partir de esta agrupación, determinar la distribución de frecuencias y dibujar el histograma, el polígono de frecuencias y el gráfico de frecuencias acumuladas relativas.

1.3.2. Soluciones de los problemas propuestos

Solución del problema 1.1. La variable estadística es el *tipo o modalidad de gasto*. Es cualitativa nominal. Tiene 3 categorías, clases o modalidades. Cada vez que se realiza un gasto en la biblioteca se observa dicha variable (cada individuo es cada gasto que se hace).

Categorías (Tipos de gasto)	f_i	ángulos
Gasto en personal	6570	$190'74^o$
Gasto en libros	3450	$100'16^o$
Otros gastos	2380	$69'10^o$
suma	12400	$360'00^o$

- *Diagrama de barras de frecuencias absolutas*: se sitúan en el eje horizontal las categorías y sobre cada una de ellas se levanta un rectángulo de altura igual a la frecuencia absoluta, f_i.
- *Gráfico de sectores*: se divide el área de un círculo en sectores circulares de ángulos iguales a los que aparecen en la última columna de la tabla anterior.

Solución del problema 1.2. La variable estadística es el *idioma*. Es cualitativa nominal. Tiene 4 categorías, clases o modalidades. Los individuos a los que se les observa dicha variable son los estantes (se supone que en cada estante sólo hay libros en el mismo idioma; es decir, en un estante no se mezclan dos idiomas).

Categorías (Idiomas)	f_i	h_i	ángulos
Francés	78	$0'4457$	$160'452^o$
Alemán	47	$0'2686$	$96'696^o$
Ruso	20	$0'1143$	$41'148^o$
Español	30	$0'1714$	$61'704^o$
suma	175	$1'0000$	$360'000^o$

- *Diagrama de barras de frecuencias relativas*: se sitúan en el eje horizontal las categorías y sobre cada una de ellas se levanta un rectángulo de altura igual a la frecuencia relativa, h_i.
- *Gráfico de sectores*: se divide el área de un círculo en sectores circulares de ángulos iguales a los que aparecen en la última columna de la tabla anterior.

Solución del problema 1.3. La variable estadística es el *tipo de fotocopia (¿con qué fin está hecha?)*. Es cualitativa nominal. Tiene 4 categorías, clases o modalidades. Los individuos a los que se les observa dicha variable son todas y cada una de las fotocopias que se realizan en la mencionada biblioteca durante el determinado año.

Categorías (Tipos de fotocopia)	f_i	$\%_i$	ángulos
Reproducción de catálogos	16110	$12'83$	$46'188^o$
Trabajo del personal de la biblioteca	63350	$50'44$	$181'584^o$
Préstamo interbibliotecario	2600	$2'07$	$7'452^o$
Copias para usuarios de la biblioteca	43540	$34'67$	$124'812^o$
suma	125600	$100'00$	$360'000^o$

- *Diagrama de barras de porcentajes*: se sitúan en el eje horizontal las categorías y sobre cada una de ellas se levanta un rectángulo de altura igual al porcentaje, $\%_i$.
- *Gráfico de sectores*: se divide el área de un círculo en sectores circulares de ángulos iguales a los que aparecen en la última columna de la tabla anterior.

Solución del problema 1.4. Tenemos 4 variables estadísticas cualitativas nominales cuyas categorías son las mismas (Reproducción de catálogos, Trabajo del personal de la biblioteca, Préstamo interbibliotecario y Copias para usuarios de la biblioteca). Cada una de estas cuatro variables es totalmente análoga a la variable definida en el problema anterior.

Categorías (Tipos de fotocopia)	A	B	C	D
	f_i	f_i	f_i	f_i
Reproducción de catálogos	16 110	3 640	0	3 400
Trabajo del personal de la biblioteca	63 350	11 360	3 080	5 500
Préstamo interbibliotecario	2 600	1 090	560	250
Copias para usuarios de la biblioteca	43 540	58 040	1 980	0

Diagrama de barras conjunto de frecuencias absolutas: se sitúan en el eje horizontal las cuatro categorías y sobre cada una de ellas se levanta un rectángulo de altura igual a la frecuencia absoluta, f_i, con distinto color o trama de relleno para cada una de las cuatro bibliotecas.

Solución del problema 1.5. Tenemos 3 variables estadísticas cualitativas nominales cuyas categorías son las mismas (sociología, economía, política y psicología). Por ejemplo, la primera de las variables es *área de investigación de las citas que aparecen en los artículos publicados en 1970*. Los individuos a los que se les observa dicha variable son todas y cada una de las citas que aparecen en los artículos publicados en 1970. Las otras dos variables se definen de forma análoga (. . . *1980* y . . . *1990*).

Categorías (Áreas de investigación)	1970		1980		1990	
	f_i	h_i	f_i	h_i	f_i	h_i
Sociología	330	0′3075	414	0′2562	547	0′4422
Economía	299	0′2787	393	0′2432	295	0′2385
Política	115	0′1072	357	0′2209	137	0′1108
Psicología	329	0′3066	452	0′2797	258	0′2086
suma	1 073	1′0000	1 616	1′0000	1 237	1′0000

Diagrama de barras conjunto de frecuencias relativas: se sitúan en el eje horizontal las cuatro categorías y sobre cada una de ellas se levanta un rectángulo de altura igual a la frecuencia relativa, h_i, con distinto color o trama de relleno para cada uno de los tres años.

Solución del problema 1.6. La variable estadística es el *número de palabras clave por artículo*. Es cuantitativa discreta. Los individuos a los que se les observa la variable son todos y cada uno de los 72 artículos de investigación de la muestra.

x_i	3	4	5	6	7	8	9	10	11	12	13	14
f_i	5	8	12	7	9	9	10	5	3	2	1	1

Diagrama de barras de frecuencias absolutas: se sitúan en el eje horizontal los x_i y sobre cada uno de ellos se levanta un segmento rectilíneo de altura igual a la correspondiente frecuencia absoluta, f_i.

Solución del problema 1.7. La variable estadística es la *altura de los libros*. Es cuantitativa continua. Los individuos a los que se les observa la variable son los 31 libros de la muestra.

x_i	15	16	17	18	19	20	21	22	23	24	25	26	27
f_i	1	0	3	4	4	2	4	5	2	2	2	1	1
h_i	0′032	0′000	0′097	0′129	0′129	0′065	0′129	0′161	0′065	0′065	0′065	0′032	0′032

Polígono de frecuencias relativas: se sitúan los puntos que resultan de tomar en el eje horizontal los distintos valores de la variable, x_i, y en el eje vertical sus correspondientes frecuencias relativas, h_i, uniendo después los puntos mediante segmentos rectilíneos.

Solución del problema 1.8. La variable estadística es el *número de palabras por línea*. Es cuantitativa discreta. Los individuos a los que se les observa la variable son todas y cada una de las 48 líneas de la página del libro.

x_i	4	5	8	9	10	11	12	13	14	15	16	17
f_i	1	1	2	3	2	7	11	14	3	2	1	1
F_i	1	2	4	7	9	16	27	41	44	46	47	48

Gráfico de frecuencias acumuladas absolutas: es la representación gráfica de las frecuencias acumuladas absolutas, F, para todo valor numérico, x. Es una gráfica en forma de “escalera".

Solución del problema 1.9. La variable estadística es el *número de palabras por resumen*. Es cuantitativa discreta. Los individuos a los que se les observa la variable son los artículos científicos de autores españoles que han publicado en la determinada revista de investigación durante el determinado año.

x_i	f_i	h_i	F_i	H_i
10	1	0′025	1	0′025
11	1	0′025	2	0′050
12	3	0′075	5	0′125
13	2	0′050	7	0′175
14	3	0′075	10	0′250
15	6	0′150	16	0′400
16	2	0′050	18	0′450
17	6	0′150	24	0′600
18	3	0′075	27	0′675
19	4	0′100	31	0′775
20	4	0′100	35	0′875
21	2	0′050	37	0′925
22	2	0′050	39	0′975
23	1	0′025	40	1′000

- *Diagrama de barras de frecuencias absolutas*: se sitúan en el eje horizontal los x_i, y sobre cada uno de ellos se levanta un segmento rectilíneo de altura igual a la correspondiente frecuencia absoluta, f_i.
- *Polígono de frecuencias relativas*: se sitúan los puntos que resultan de tomar en el eje horizontal los distintos valores de la variable, x_i, y en el eje vertical sus correspondientes frecuencias relativas, h_i, uniendo después los puntos mediante segmentos rectilíneos.
- *Gráfico de frecuencias acumuladas relativas*: es la representación gráfica de las frecuencias acumuladas relativas, H, para todo valor numérico, x. Es una gráfica en forma de "escalera".

Solución del problema 1.10. La variable estadística es el *número de llamadas telefónicas recibidas en el servicio de información de una biblioteca pública*. Es cuantitativa discreta. Los individuos a los que se les observa la variable son los días.

$(\ell_i, \ell_{i+1}]$	(9,15]	(15,21]	(21,27]	(27,33]	(33,39]	(39,45]	(45,51]
f_i	2	4	8	14	10	6	1
x_i	12	18	24	30	36	42	48
F_i	2	6	14	28	38	44	45

- *Histograma*: se sitúan en el eje horizontal los intervalos de clase, $(\ell_i, \ell_{i+1}]$, y sobre cada uno se levanta un rectángulo de área proporcional a la frecuencia absoluta. Como los intervalos tienen la misma amplitud, basta con hacer las alturas de los rectángulos iguales a las frecuencias absolutas, f_i.
- *Polígono de frecuencias*: se sitúan los puntos que resultan de tomar en el eje horizontal las marcas de clase, x_i, y en el eje vertical sus correspondientes frecuencias absolutas, f_i, uniendo después los puntos mediante segmentos rectilíneos.
- *Gráfico de frecuencias acumuladas absolutas*: se sitúan los puntos que resultan de tomar en el eje horizontal los extremos superiores de los intervalos de clase, ℓ_{i+1}, y en el eje vertical sus correspondientes frecuencias acumuladas absolutas, F_i, uniendo después dichos puntos mediante segmentos rectilíneos.

Solución del problema 1.11. La variable estadística es el *número de socios de la biblioteca*. Es cuantitativa discreta. Los individuos a los que se les observa la variable son las bibliotecas públicas.

$(\ell_i, \ell_{i+1}]$	f_i	x_i	H_i
(675,1 375]	19	1 025	$0'2262$
(1 375,2 075]	22	1 725	$0'4881$
(2 075,2 775]	18	2 425	$0'7024$
(2 775,3 475]	14	3 125	$0'8690$
(3 475,4 175]	8	3 825	$0'9643$
(4 175,4 875]	2	4 525	$0'9881$
(4 875,5 575]	1	5 225	$1'0000$

- *Histograma*: se sitúan en el eje horizontal los intervalos de clase, $(\ell_i, \ell_{i+1}]$, y sobre cada uno se levanta un rectángulo de área proporcional a la frecuencia absoluta. Como los intervalos tienen la misma amplitud, basta con hacer las alturas de los rectángulos iguales a las frecuencias absolutas, f_i.
- *Polígono de frecuencias*: se sitúan los puntos que resultan de tomar en el eje horizontal las marcas de clase, x_i, y en el eje vertical sus correspondientes frecuencias absolutas, f_i, uniendo después los puntos mediante segmentos rectilíneos.
- *Gráfico de frecuencias acumuladas relativas*: se sitúan los puntos que resultan de tomar en el eje horizontal los extremos superiores de los intervalos de clase, ℓ_{i+1}, y en el eje vertical sus correspondientes frecuencias acumuladas relativas, H_i, uniendo después dichos puntos mediante segmentos rectilíneos.

1.4. PRÁCTICA 1: INTRODUCCIÓN A MINITAB

1.4.1. Elementos de Minitab para Windows

1.4.1.1. Introducción

Al ejecutar ***Minitab*** 15 aparece la *ventana* de la Figura 1.

Como en cualquier otra aplicación Windows, esta *ventana* puede modificarse en cuanto al tamaño y a la disposición de sus elementos. Se trata de una *ventana* típica de una aplicación Windows que consta de los siguientes elementos:

- En la primera línea aparece la **barra de título**, que contiene el nombre de la ventana y los botones de minimizar, maximizar y cerrar.
- En la segunda línea está la **barra de menús**, que consta de los 10 menús que luego comentaremos.
- Las líneas tercera y cuarta conforman la **barra de herramientas** donde, mediante botones con iconos, se representan algunas de las operaciones más habituales. Si pasamos el puntero del ratón por cualquiera de ellos, aparecerá en la pantalla un texto indicando la función que se activa.
- Después aparece la **ventana de sesión (Session)**. Es la parte donde aparecen los resultados de los análisis realizados. También sirve para escribir instrucciones, como forma alternativa al uso de los menús.
- A continuación tenemos la **hoja de datos (Worksheet)**. Tiene el aspecto de una hoja de cálculo, con filas y columnas. Las columnas se denominan $C1, C2, \ldots$, tal como está escrito, pero también se les puede dar un nombre, escribiéndolo debajo de $C1, C2, \ldots$ Cada columna es una variable y cada fila corresponde a una observación o caso.
- En la parte inferior aparece (minimizada) la **ventana de proyecto (Proyect Manager)**. En ***Minitab*** un proyecto incluye la hoja de datos, el contenido de la ventana de sesión, los gráficos que se hayan realizado, los valores de las constantes y de las matrices que se hayan creado, etc.

Para activar la ventana de sesión (**Session**) podemos hacer *clic* sobre ella o podemos hacer *clic* sobre su icono en la barra de herramientas (primer icono de la Figura 2). Para activar la hoja de datos (**Worksheet**) podemos hacer *clic* sobre ella o podemos hacer *clic* sobre su icono en la barra de herramientas (segundo icono de la Figura 2). Para activar la ventana de proyecto (**Proyect Manager**)

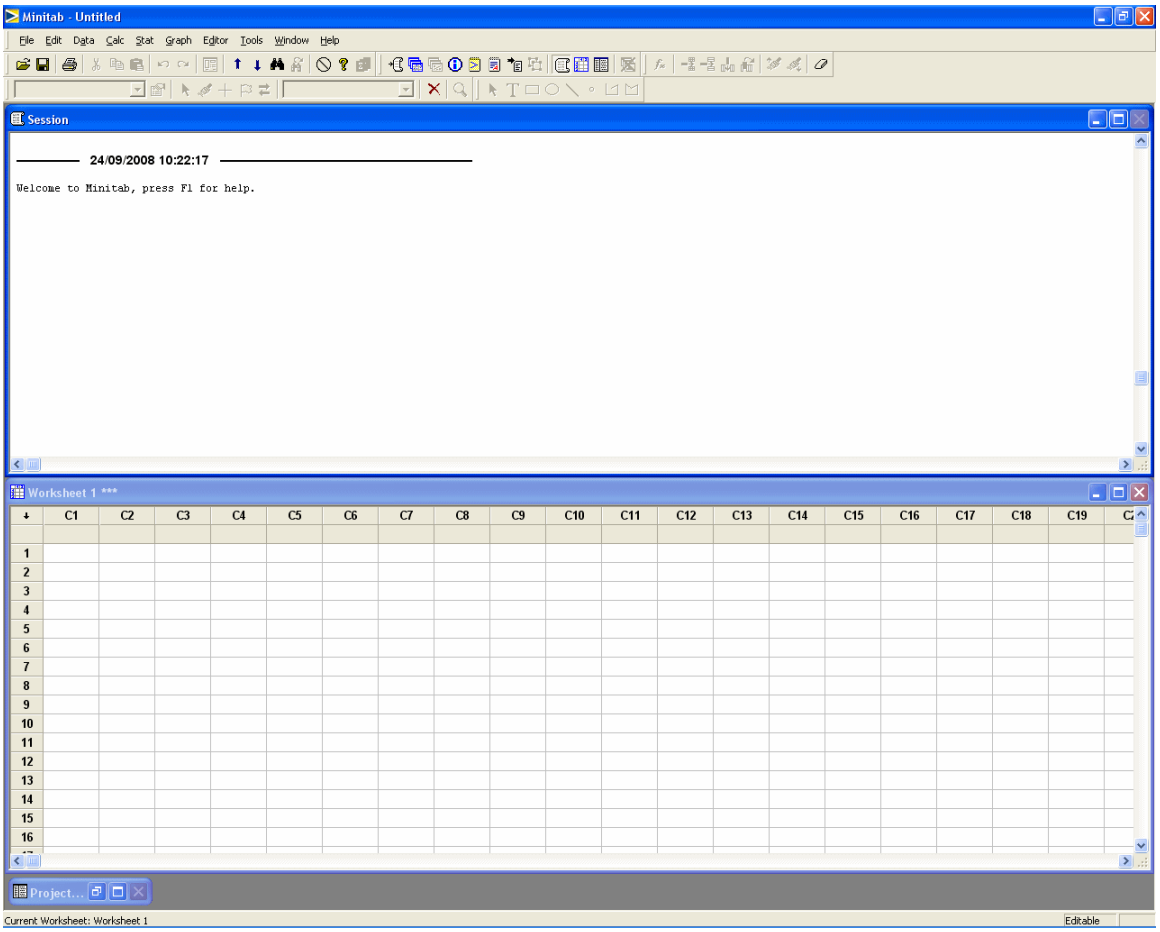


Figura 1: Ventana inicial de Minitab 15

podemos maximizarla o podemos hacer *clic* sobre su icono en la barra de herramientas (tercer icono de la Figura 2).

Figura 2: Iconos para activar las ventanas de sesión, de datos o de proyecto

Para salir del programa se selecciona la opción **File** ⇒**Exit** o se pulsa el botón de la esquina superior derecha: ☒.

1.4.1.2. Barra de menús

A continuación se da un resumen de lo que se puede encontrar en la **barra de menús**:

File: Mediante este menú se pueden abrir, crear o grabar los diferentes archivos que ***Minitab*** emplea, ya sean de datos, instrucciones, resultados o procesos. Igualmente, es posible controlar las tareas de impresión.

Edit: Permite realizar las tareas habituales de edición: modificar, borrar, copiar, pegar, seleccionar, etc.

Data: Este menú permite, entre otras cosas, efectuar modificaciones en los archivos de datos: extraer un subconjunto de datos, apilar y desapilar, ordenar, codificar, etc.

Calc: Aquí se encuentran todas las opciones relativas a la modificación y generación de nuevas variables, cálculo de los estadísticos, introducción de datos por patrón, cálculo de las distribuciones de probabilidad, etc.

Stat: Mediante este menú se accede a los diferentes análisis estadísticos que se pueden realizar con los datos.

Graph: Permite la creación y edición de diversos tipos de gráficos. Algunos de ellos son también accesibles a través de determinadas técnicas estadísticas.

Editor: Tiene distintas opciones según esté activada la ventana de sesión o la hoja de datos. Con la ventana de sesión activada permite, por ejemplo, que se pueda escribir (en dicha ventana) utilizando el *lenguaje de comandos*.

Tools: Entre otras cosas, permite personificar la barra de herramientas y la barra de menús.

Windows: Dispone de las funciones habituales para controlar las ventanas.

Help: Proporciona ayuda al usuario en el formato típico de Windows.

1.4.2. Entrada, grabación y lectura de datos

1.4.2.1. Entrada de datos

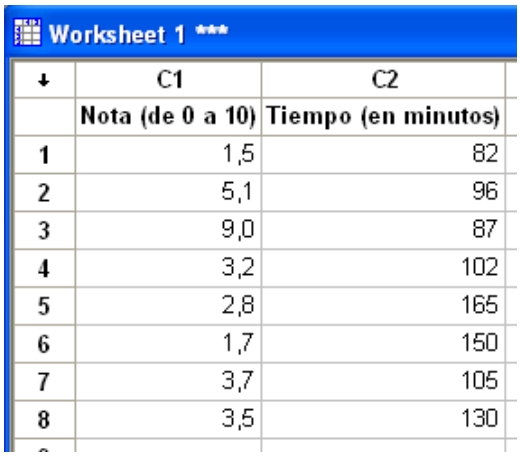
Antes de realizar ningún análisis estadístico es necesario tener un conjunto de datos en uso, para lo cual podemos proceder de cuatro formas:

- Escribirlos a través del teclado.
- Obtenerlos desde un archivo.
- Pegarlos.
- Generarlos por patrón o de forma aleatoria.

Para introducir datos a través del teclado, activamos, en primer lugar, la hoja de datos. En la parte superior aparece $C1$, $C2$, $C3$, ... y debajo un espacio en blanco para poner el nombre de cada variable. La flechita del extremo superior izquierdo de la hoja de datos señala hacia dónde se mueve el cursor al pulsar la tecla Intro. Por defecto apunta hacia abajo, ↓; si se hace *clic* sobre ella, apuntará hacia la derecha, →. Para escribir datos por columna no hay más que situarse en la casilla del caso 1, teclear el dato y pulsar la tecla Intro. La casilla activa se moverá hacia abajo. Si tecleamos datos que no son numéricos podemos observar que junto a CJ aparece un guión y la letra T (es decir, $CJ - T$), lo que significa que ***Minitab*** reconoce que la variable es cualitativa (o de texto).

Con esta versión de ***Minitab***, al introducir los resultados de una variable cuantitativa (o numérica) tenemos que recordar que la separación decimal se hace mediante una coma (en parte de abajo). Si, por ejemplo, ponemos un punto como separación decimal, entonces ***Minitab*** consideraría, automáticamente, que dicha la variable es cualitativa o de texto (junto a CJ aparece un guión y la letra T) y, por tanto, no podríamos hacer ningún cálculo matemático con los datos de esta variable.

Por ejemplo, podemos introducir los datos de la Figura 3, correspondientes a las calificaciones (de 0 a 10 puntos) en el examen de Estadística y el tiempo (en minutos) empleado en realizar dicho examen.



Worksheet 1 ***

↓	C1	C2
	Nota (de 0 a 10)	Tiempo (en minutos)
1	1,5	82
2	5,1	96
3	9,0	87
4	3,2	102
5	2,8	165
6	1,7	150
7	3,7	105
8	3,5	130

Figura 3: Ejemplo para introducir datos a través del teclado

Si el nombre de la variable (columna) no es suficientemente explicativo, podemos escribir una descripción de la variable para poder consultarla en cualquier momento. Para ello, hacemos *clic* sobre el nombre de la variable (o sobre su número de columna: CJ); pulsamos con el botón derecho del ratón y seleccionamos **Column⇒Description**. Por ejemplo, podríamos escribir etiquetas descriptivas para las variables **Nota (de 0 a 10)** y **Tiempo (en minutos)**.

Para cambiar el formato de una variable (columna) numérica, hacemos *clic* sobre el nombre de la variable (o sobre su número de columna: CJ); pulsamos con el botón derecho del ratón y seleccionamos **Format Column⇒Numeric**. Una de las utilidades de esta opción es **el cambio del número de decimales** que se muestran en la hoja de datos. Por ejemplo, podríamos hacer que ***Minitab*** mostrase 2 decimales en la columna **Nota (de 0 a 10)**.

Una hoja de datos de ***Minitab*** puede contener hasta 4 000 columnas, 1 000 constantes y hasta 10 000 000 de filas, dependiendo de la memoria que tenga el ordenador.

1.4.2.2. Grabación de datos

Una vez introducidos los datos, éstos pueden guardarse en un archivo para poder ser utilizados en cualquier otro momento.

Para guardar únicamente la hoja de datos hay que seleccionar **File⇒Save Current Worksheet As** (si vamos a grabar el archivo de datos por primera vez y, por tanto, vamos a ponerle un nombre a dicho archivo) ó **File⇒Save Current Worksheet** (si el archivo de datos ya tiene nombre pero queremos guardar los últimos cambios realizados). Por ejemplo, podemos guardar los datos de la Figura 3 en un archivo que denominaremos **Notas_Tiempo.mtw**. Para ello, elegimos la opción **File⇒Save Current Worksheet As**; en **Guardar en** seleccionamos la carpeta en la que vamos a grabar esta hoja de datos; en **Nombre** escribimos **Notas_Tiempo** (***Minitab*** le asigna automáticamente la extensión **.mtw**) y, por último, pulsamos en **Guardar**.

Si queremos grabar toda la información (la hoja de datos, el contenido de la ventana de sesión, los gráficos que se hayan realizado, los valores de las constantes y de las matrices que se hayan creado, etc.) usaremos la opción **File⇒Save Project As** (si vamos a grabar el proyecto de ***Minitab*** por primera vez y, por tanto, vamos a ponerle un nombre a dicho archivo) ó **File⇒Save Project** (si el proyecto ya tiene nombre pero queremos guardar los últimos cambios realizados). Es muy importante diferenciar entre archivos de datos (**.mtw**) y archivos de proyectos (**.mpj**).

También se puede guardar solamente la ventana de sesión. Para ello, la activamos y seleccionamos la opción **File⇒Save Session Windows As**.

1.4.2.3. Lectura de datos

Un archivo sólo puede ser recuperado de la forma en que fue grabado. Si se ha grabado como hoja de datos (**.mtw**) se recupera con la opción **File⇒Open Worksheet**. Si se ha grabado como proyecto de ***Minitab*** (**.mpj**) se recupera con la opción **File⇒Open Proyect**.

Minitab 15 lleva bastantes archivos de datos como muestra. Éstos se encuentran en **C:\Archivos de programa\Minitab 15\English\Sample Data** y, como ya sabemos, llevan la extensión **.mtw**. En las aulas de informática de la Universidad de Murcia es posible que se encuentren en **C:\Archivos de programa\UM\Minitab 15\English\Sample Data**.

Por ejemplo, podemos abrir el archivo de datos **Pulse.mtw**. Su contenido fue recogido en una clase de 92 alumnos. De cada estudiante se observó su pulso antes de correr, **Pulse1**; su pulso después de correr, **Pulse2**; si corrió o no, **Ran** (1=Sí corrió, 2=No corrió); si es fumador o no, **Smokes** (1=Sí fuma, 2=No fuma); el sexo, **Sex** (1=Hombre, 2=Mujer); su altura en pulgadas, **Height**; su peso en libras, **Weight**; y su nivel de actividad física, **Activity** (0=Ninguna actividad física, 1=Baja, 2=Media, 3=Alta). Se puede encontrar más información de este archivo de datos con la opción **Help⇒Help⇒Indice**. Bajo la frase **Escriba la palabra clave a buscar** se teclea **Pulse.mtw** y después se hace *clic* en **Mostrar** o se hace doble *clic* sobre el nombre de dicho archivo.

Con la opción **File⇒Open Worksheet** se pueden leer otros tipos de archivos de datos, como hojas de cálculo de Excel, Lotus 1-2-3, dBase, etc. Para obtener una información más detallada sobre los tipos de archivos que ***Minitab*** puede leer, se selecciona **File⇒Open Worksheet** y, en el cuadro de diálogo resultante, se hace *clic* sobre **Ayuda**.

1.4.3. Opciones principales de los menús *Data* y *Calc*

Si queremos que en la ventana de sesión (**Session**) aparezcan los comandos que va a utilizar ***Minitab*** en las opciones que vamos a explicar, activamos la ventana de sesión y luego seleccionamos **Editor⇒Enable Commands**.

1.4.3.1. Desapilamiento de columnas

La opción **Data⇒Unstack columns** permite separar los resultados de una columna en varias columnas, según los resultados de otra variable o columna (que contiene los subíndices).

Por ejemplo, de la hoja de datos **Pulse.mtw** vamos a desapilar los resultados de la variable **Pulse2** (*pulso después de correr*) según los resultados de la variable **Ran** (*1=Sí corrió, 2=No corrió*).

En primer lugar tenemos que abrir dicha hoja de datos, si no la tenemos abierta ya. Recordemos que para abrirla elegimos la opción **Open Worksheet**; en **Buscar en** seleccionamos la carpeta donde se encuentra la hoja de datos; activamos **Nombre**; seleccionamos el archivo **Pulse.mtw** y, por último, pulsamos en **Abrir**.

Para realizar el desapilamiento de los resultados de la variable **Pulse2** según los resultados de la variable **Ran** seleccionamos **Data⇒Unstack Columns**; activamos **Unstack the data in** (haciendo *clic* dentro del recuadro); seleccionamos (haciendo doble *clic* sobre su nombre) la variable o columna **Pulse2**; activamos el recuadro **Using subscripts in** (haciendo *clic* dentro del recuadro); y seleccionamos la columna

que contiene la procedencia de cada dato, que es **Ran**; en **Store unstacked data in** activamos la opción **After last column in use**; dejamos activado **Name the columns containing the unstaked data** y pulsamos en **OK**.

En la hoja de datos **Pulse.mtw** nos aparecen dos nuevas columnas: **Pulse2_1** y **Pulse2_2**. En la columna **Pulse2_1** hay 35 datos, que son los resultados del pulso después de correr (**Pulse2**) de las personas que sí corrieron (**Ran**=1); y en la columna **Pulse2_2** hay 57 datos, que son los resultados del pulso después de correr (**Pulse2**) de las personas que no corrieron (**Ran**=2).

Debemos grabar la actual hoja de datos con un nombre distinto de **Pulse.mtw** para conservar los datos originales sin transformaciones ni nuevas columnas. Para ello, elegimos la opción **File⇒Save Current Worksheet As**; en **Guardar en** seleccionamos la carpeta en la que vamos a grabar esta hoja de datos; en **Nombre** escribimos **Pulse transformada** y, por último, pulsamos en **Guardar**.

1.4.3.2. Apilamiento de columnas

Con la opción **Data⇒Stack⇒Columns** se pueden apilar varias columnas en una sola. Opcionalmente se puede indicar de qué columna procede cada valor mediante una nueva variable (subíndices). Si no se hace esta indicación no se podrá identificar la procedencia de cada dato. Esta opción es la contraria de la explicada en el apartado anterior.

Para practicar esta opción podemos apilar los datos de las columnas **Pulse2_1** y **Pulse2_2** de la hoja de datos **Pulse transformada.mtw**. En primer lugar debemos asegurarnos de que la hoja de datos activa es **Pulse transformada.mtw**. Si dicha hoja de datos no está activa, debemos activarla haciendo *clic* sobre ella o seleccionando **Window⇒Pulse transformada.mtw**. A continuación, seleccionamos la opción **Data⇒Stack⇒Columns**; activamos el recuadro **Stack the following columns** y seleccionamos (haciendo doble *clic* sobre sus nombres) las dos columnas que queremos apilar: **Pulse2_1** y **Pulse2_2**; en **Store stacked data in** activamos la opción **Column of current worksheet** y tecleamos la posición de una columna que esté vacía, por ejemplo, **C11** (o escribimos un nombre para esta nueva columna). En **Store subscripts in** tecleamos la posición de la columna en la que queremos guardar la procedencia de cada dato, por ejemplo, **C12** (o escribimos un nombre para esta nueva columna). Es conveniente dejar activada la opción **Use variable names in subscript column**.

Podemos observar que la columna **Pulse2** y la columna **C11** contienen los mismos resultados, pero no en el mismo orden.

1.4.3.3. Ordenación de datos

La opción **Data⇒Sort** ordena los datos de una columna según los resultados de una o varias columnas. Lo normal es ordenar una columna según los resultados de dicha columna. Esto es lo que vamos a explicar.

Por ejemplo, en la hoja de datos **Pulse transformada.mtw** vamos a crear una nueva variable (columna) que contenga los resultados de la variable **Pulse1** ordenados de menor a mayor. En primer lugar, activamos dicha hoja de datos (si no la tenemos activada ya). A continuación, seleccionamos **Data⇒Sort**; activamos el recuadro **Sort column**; seleccionamos (haciendo doble *clic* sobre su nombre) la variable **Pulse1**; activamos el primer recuadro **By column** que aparece y volvemos a seleccionar la misma columna, **Pulse1**. Dejamos desactivada la opción **Descending** para que la ordenación se realice de menor a mayor resultado.

En **Store sorted data in** activamos **Column of current worksheet** y tecleamos el nombre que queremos ponerle a dicha columna, por ejemplo, **'Pulse1 ordenado'**. En este cuadro de diálogo (en realidad, en

todos los cuadros de diálogo de ***Minitab***), cuando haya que escribir el nombre de una nueva variable (columna) y **el nombre contenga espacios en blanco, guiones, paréntesis, etc., entonces hay que escribirlo entre comillas simples**. La comilla simple suele estar en la misma tecla que el símbolo de cerrar interrogación.

Hay tener cuidado con la ordenación de columnas debido a que los resultados de esta nueva variable no guardan correspondencia con los casos originales. Por ejemplo, la primera persona observada tiene un pulso antes de correr (resultado de **Pulse1**) igual a 64 pulsaciones por minuto, no 48 pulsaciones por minuto, como nos ha salido en el primer lugar de la columna **Pulse1 ordenado**. Como podemos observar, el menor valor de **Pulse1** es 48 y el mayor valor es 100.

1.4.3.4. Codificación o clasificación de datos

La opción **Data⇒Code** permite la clasificación o codificación de los datos de una columna. Se puede codificar transformando datos numéricos en datos numéricos, datos numéricos en datos de texto, datos de texto en datos de texto, datos de texto en datos numéricos, etc.

Por ejemplo, con la hoja de datos **Pulse transformada.mtw** podemos codificar la variable **Pulse1** de la forma siguiente:

Resultados de Pulse1	Nueva categoría
comprendido entre 48, incluido, y 65, incluido	Pulso bajo
comprendido entre 65, sin incluir, y 83, incluido	Pulso medio
comprendido entre 83, sin incluir, y 100, incluido	Pulso alto

Para ello, seleccionamos **Data⇒Code⇒Numeric to Text**. En **Code data from columns** seleccionamos (haciendo doble *clic* sobre su nombre) la variable **Pulse1**. En **Store coded data in column** escribimos el nombre la nueva variable; por ejemplo, **'codificación de Pulse1'** (con comillas simples, al principio y al final, ya que el nombre tiene espacios en blanco). En la primera línea de **Original values** debemos escribir **48:65**, lo cual es interpretado por ***Minitab*** de la siguiente manera: todos los resultados comprendidos entre 48, incluido, y 65, incluido. En la primera línea de **New** escribimos **Pulso bajo**. En la segunda línea de **Original values** escribimos **65:83** lo cual es interpretado por ***Minitab*** de la siguiente manera: todos los resultados comprendidos entre 65, sin incluir, y 83, incluido. En la segunda línea de **New** escribimos **Pulso medio**. En la tercera línea de **Original values** escribimos **83:100** lo cual es interpretado por ***Minitab*** de la siguiente manera: todos los resultados comprendidos entre 83, sin incluir, y 100, incluido. En la tercera línea de **New** escribimos **Pulso alto**.

1.4.3.5. Transformación de variables

En este apartado vamos a ver el modo de generar nuevas variables mediante transformaciones efectuadas sobre los valores de las variables ya definidas. Para ello vamos a utilizar la opción **Calc ⇒Calculator**

En la Tabla 4 se encuentran recogidos los operadores aritméticos, relacionales y lógicos que están permitidos. Tanto las expresiones aritméticas como las lógicas se evalúan de izquierda a derecha. Todas las expresiones entre paréntesis se evalúan antes que las que están fuera de los paréntesis y ante varios operadores en el mismo nivel, el orden de preferencia (de mayor a menor) es el que figura en la Tabla 4 (de arriba hacia abajo).

()	Paréntesis
**	Exponenciación
*	Multiplicación
/	División
+	Suma
−	Resta

(a) Operadores aritméticos

<	Menor que
>	Mayor que
<=	Menor o igual que
>=	Mayor o igual que
=	Igual que
<>	No igual que

(b) Operadores relacionales

AND	Operador Y
OR	Operador O
NOT	Operador NO

(c) Operadores lógicos

Tabla 4: Operaciones aritméticas, relacionales y lógicas

Como ya hemos indicado, para construir una nueva variable mediante transformaciones de otras ya existentes, se tiene que elegir la opción **Calc ⇒Calculator**, con lo que se abre una ventana que tiene cinco partes fundamentales: arriba a la derecha está el lugar para escribir el nombre de la nueva variable (**Store result in variable**), a la izquierda aparece la lista de variables y constantes existentes, a la derecha está el lugar destinado a la definición de la nueva variable (**Expression**), debajo hay una calculadora y la lista de funciones que se pueden utilizar (**Functions**).

En primer lugar se asigna un nombre a la variable que queremos generar, escribiendo el mismo en el cuadro **Store result in variable**. Normalmente se va a tratar de una variable nueva, pero también cabe la posibilidad de especificar una de las ya existentes. En tal caso la modificación consistirá en sustituir los valores antiguos de la variable con los nuevos resultantes de la transformación numérica que se efectúe.

Una vez que se ha asignado el nombre a la variable, el siguiente paso es definir la expresión que va a permitir calcular los valores de la misma. Tal expresión se escribe en el cuadro **Expression** y puede constar de los siguientes elementos: nombres de variables del archivo original, constantes, operadores y funciones. Para escribir dicha expresión, se puede teclear directamente pero **es recomendable emplear la calculadora, la lista de variables y constantes y la lista de funciones** (haciendo *clic* dentro del recuadro **Expression** y haciendo doble *clic* sobre la variable, sobre la constante o sobre la función). Una vez que hemos terminado de escribir la expresión, pulsamos en **OK**.

Por ejemplo, del archivo de datos **Pulse transformada.mtw** vamos a calcular la media geométrica de las variables **Pulse1** y **Pulse2** (raíz cuadrada del producto de ambas variables; es decir, producto de ambas variables elevado a 1/2). Para ello, seleccionamos la opción **Calc⇒Calculator**; en **Store result in variable** tenemos que teclear la posición de la columna que contendrá los resultados (una columna, **CJ**, que esté vacía) o el nombre que queremos darle a dicha columna. Nosotros vamos a poner a la nueva variable el siguiente nombre: **'Media geométrica Pulse1 Pulse2'** (con comillas simples, al principio y al final, ya que el nombre tiene espacios en blanco). En **Expression** tenemos que colocar la operación que se realiza para determinar la media geométrica indicada: `('Pulse1' * 'Pulse2')**(1 / 2)`. Por último, pulsamos en **OK**.

1.4.3.6. Creación de datos por patrón

Con la opción **Calc⇒Make Patterned Data** se generan datos siguiendo un determinado patrón.

Por ejemplo, si queremos generar una lista de los siguientes 100 números: $0'01$, $0'02$, $0'03$, ..., 1, seguiremos los siguientes pasos:

Como estos datos no tienen nada que ver con los datos del archivo **Pulse transformada.mtw**, creamos una nueva hoja de datos con la opción **File⇒New**. En el cuadro de diálogo que aparece seleccionamos **Minitab Woorksheet**. A esta nueva hoja de datos ***Minitab*** le asignará el nombre Worksheet J, siendo J un número natural. Luego podremos cambiarle el nombre con la opción **File⇒Save Current Worksheet As**. Seleccionamos, a continuación, la opción **Calc⇒Make Patterned Data⇒Simple Set of Numbers**. En **Store patterned data in** podemos teclear $C1$ o un nombre, por ejemplo **'Patrón entre 0 y 1'** (con comillas simples, al principio y al final, ya que el nombre tiene espacios en blanco). En **From first value** tecleamos **0,01**, en **To last value** escribimos **1** y en **In steps of** ponemos **0,01**. Tanto en **List each value** como en **List the whole sequence** dejamos lo que está puesto por defecto, que es **1**.

1.4.4. Ejercicios prácticos propuestos

Ejercicio 1.1. En la Tabla 5 se muestra el número anual de usuarios de una biblioteca determinada y el número anual de préstamos durante 10 años elegidos al azar.

año	usuarios	préstamos
1	296	155
2	459	275
3	602	322
4	798	582
5	915	761
6	1145	856
7	1338	1030
8	1576	1254
9	1780	1465
10	2050	1675

Tabla 5

a) Crea un nuevo proyecto de ***Minitab***.

b) Introduce los datos (sin incluir, obviamente, la primera columna, que indica el número de caso). Pon los siguientes nombres a las dos variables: **Usuarios** y **Préstamos**. Graba la hoja de datos en un archivo denominado **Prestamos.mtw**

c) Calcula, en una nueva columna, la variable que indica el **porcentaje anual de préstamos por usuario**, resultado de multiplicar por 100 el resultado de dividir el número anual de préstamos entre el número anual de usuarios. Pon a la nueva variable el siguiente nombre: **PPU**. Haz que los resultados aparezcan con tres decimales. Pon una etiqueta descriptiva a esta variable. Vuelve a grabar la hoja de datos.

d) Ordena de mayor a menor (y coloca en una nueva columna de la actual hoja de datos) los resultados de la variable **PPU**. Pon un nombre adecuado a la nueva columna. Pon una etiqueta descriptiva a esta columna. Observa esta ordenación y escribe, a continuación, el valor mínimo y el valor máximo de dicha variable.

valor mínimo	valor máximo

Vuelve a grabar la hoja de datos.

e) Clasifica los datos de la variable **PPU** en 4 categorías o intervalos de la misma amplitud. Llama a la nueva variable **Intervalos PPU**. Las categorías han de denotarse como lo hacemos en las clases de teoría; es decir, $[a, b]$ o $(a, b]$ (sustituyendo, obviamente, a y b por los límites de los intervalos de clase). Escribe, a continuación, los cálculos previos necesarios:

Recorrido=R=
Amplitud de los intervalos=d=
$[l_1, l_2]$=
$(l_2, l_3]$=
$(l_3, l_4]$=
$(l_4, l_5]$=

Vuelve a grabar la hoja de datos.

f) Graba el proyecto con el siguiente nombre: **Ejercicio1-1.mpj**

Ejercicio 1.2. En la Tabla 6 aparece el número anual de transacciones de referencia y el número anual de transacciones de referencia finalizadas en 20 biblioteca elegidas al azar.

a) Crea un nuevo proyecto de ***Minitab***.

b) Introduce los datos (sin incluir, obviamente, la primera columna, que indica el número de caso). Pon los siguientes nombres a las variables: **Tipo**, **TR** y **TRF**. Pon una etiqueta descriptiva a cada variable. En lo que respecta a la variable **Tipo** hay que dejar claro que el valor **1** significa *biblioteca pública* y el valor **2** significa *biblioteca universitaria*. Graba la hoja de datos en un archivo denominado **Transacciones.mtw**

c) Crea una nueva variable, denominada **Tipo biblioteca**, que contenga las categorías de la variable **Tipo** designadas de la siguiente manera: *bib. pública* (en vez de **1**) y *bib. universitaria* (en vez de **2**). Vuelve a grabar la hoja de datos.

d) Calcula, en una nueva columna, la variable que indica el **porcentaje de transacciones de referencia finalizadas**, que se determina multiplicando por cien el resultado de dividir el número anual de transacciones de referencia finalizadas entre el número anual de transacciones de referencia. Pon a la nueva variable el siguiente nombre: **Porcentaje TRF**. Haz que los resultados aparezcan con 5 decimales. Pon una etiqueta descriptiva a esta variable. Vuelve a grabar la hoja de datos.

e) Desapila los resultados de la variable **Porcentaje TRF** según los resultados de la variable **Tipo biblioteca**. Vuelve a grabar la hoja de datos.

f) Ordena de menor a mayor (y coloca en una nueva columna de la actual hoja de datos) los resultados de la variable **Porcentaje TRF**. Pon un nombre adecuado a la nueva columna. Pon una etiqueta descriptiva a esta columna. Observa esta ordenación y escribe, a continuación, el valor mínimo y el valor máximo de dicha variable.

biblioteca	tipo de biblioteca	transacciones de referencia	transacciones de referencia finalizadas
1	1	11500	9400
2	1	8600	7200
3	1	20400	18100
4	1	5800	4600
5	1	6500	5800
6	1	13700	10900
7	1	12400	11200
8	1	5300	4700
9	1	6700	5600
10	1	15600	12500
11	2	1900	1700
12	2	9600	7800
13	2	8400	6900
14	2	6200	4900
15	2	7700	5900
16	2	5600	4200
17	2	6200	4900
18	2	4800	3500
19	2	3800	2600
20	2	2400	2200

Tabla 6

valor mínimo	valor máximo

Vuelve a grabar la hoja de datos.

g) Clasifica los datos de la variable **Porcentaje TRF** en 3 categorías o intervalos de la misma amplitud. Llama a la nueva variable **Intervalos Porcentaje TRF**. Las categorías han de denotarse como lo hacemos en las clases de teoría; es decir, $[a, b]$ o $(a, b]$ (sustituyendo, obviamente, a y b por los límites de los intervalos de clase). Escribe, a continuación, los cálculos previos necesarios:

Recorrido=R=
Amplitud de los intervalos=d=
$[l_1, l_2]$= $(l_2, l_3]$= $(l_3, l_4]$=

Vuelve a grabar la hoja de datos.

h) Graba el proyecto con el siguiente nombre: **Ejercicio1-2.mpj**

2

Medidas descriptivas de los datos

2.1. Desarrollo de los contenidos fundamentales

2.1.1. Medidas de posición

Son valores que nos sirven para indicar la posición alrededor de la cual se distribuyen las observaciones.

2.1.1.1. Mediana

La **mediana** es un valor que deja a su izquierda el 50 % de los datos de la muestra ordenada. La denotaremos por M_e. Su unidad de medida es la misma que la de la variable.

a) Cálculo con datos no agrupados en intervalos:

- n impar: M_e es el valor central de la muestra ordenada.
- n par: M_e es el punto medio de los dos valores centrales de la muestra ordenada.

b) Cálculo con datos agrupados en intervalos:

Llamamos *intervalo mediano* al que contiene a la mediana. Es el primer intervalo cuya frecuencia absoluta acumulada es igual o mayor que $\dfrac{n}{2}$.

Una vez determinado el intervalo mediano, la mediana se calcula por la fórmula siguiente:

$$M_e = \ell_i + \frac{\dfrac{n}{2} - F_{i-1}}{f_i} \left(\ell_{i+1} - \ell_i\right),$$

donde $(\ell_i, \ell_{i+1}]$ es el intervalo mediano, f_i es su frecuencia absoluta y F_{i-1} es la frecuencia absoluta acumulada del intervalo anterior al mediano.

2.1.1.2. Percentiles

El **percentil** al $r\,\%$ es un valor que deja por debajo el $r\,\%$ de los datos de la muestra ordenada de menor a mayor. Lo denotaremos por P_r. Su unidad de medida es la misma que la de la variable.

CASOS PARTICULARES:

- Cuartiles:

$$\begin{array}{lcl} 1^{er}\text{ cuartil} & = & Q_1 = P_{25} \\ 2^{o}\text{ cuartil} & = & Q_2 = P_{50} = M_e \\ 3^{er}\text{ cuartil} & = & Q_3 = P_{75} \end{array}$$

- Deciles:

$$\begin{array}{lcl} 1^{er}\text{ decil} & = & D_1 = P_{10} \\ 2^{o}\text{ decil} & = & D_2 = P_{20} \\ \vdots & \vdots & \vdots \\ 9^{o}\text{ decil} & = & D_9 = P_{90} \end{array}$$

Si los datos están agrupados en intervalos de clase, el intervalo que contiene a P_r es el primero cuya frecuencia acumulada absoluta es igual o mayor que

$$\frac{n\,r}{100}$$

y el percentil al $r\,\%$ se determina mediante la fórmula:

$$P_r = \ell_i + \frac{\dfrac{n\,r}{100} - F_{i-1}}{f_i}\,(\ell_{i+1} - \ell_i)\,,$$

donde $(\ell_i, \ell_{i+1}]$ es el intervalo que contiene a P_r, f_i es su frecuencia absoluta y F_{i-1} es la frecuencia absoluta acumulada del intervalo anterior.

2.1.1.3. Media

Llamaremos **media** a la **media aritmética**. (Hay otras medias, como, por ejemplo, la media geométrica, la media cuadrática y la media armónica.)

Si la variable se denota por X, la media de los datos de una muestra será denotada por $\overline{x}$. (Si tenemos los datos de toda la población, entonces representaremos la media por μ.)

a) Cálculo con datos no agrupados en intervalos:

Si $x_1, x_2, \ldots, x_n$ son los n valores de la muestra, entonces:

$$\overline{x} = \frac{\displaystyle\sum_{i=1}^{n} x_i}{n}\,.$$

Si los datos son $x_1, x_2, \ldots, x_k$, y aparecen con frecuencias absolutas respectivas $f_1, f_2, \ldots, f_k$, entonces:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n} .$$

De las fórmulas anteriores se deduce que la unidad de medida de $\overline{x}$ es la misma que la de la variable.

b) Cálculo con datos agrupados en intervalos:

La fórmula es la misma que la anterior, siendo x_i la marca de clase del intervalo $(\ell_i, \ell_{i+1}]$ y f_i su correspondiente frecuencia absoluta.

2.1.2. Medidas de dispersión

Miden el grado de separación de las observaciones entre sí o con respecto a ciertas medidas de posición, como la media o la mediana.

2.1.2.1. Recorrido

La fórmula del **recorrido** (también denominado **rango** o **amplitud total**) es:

$$R = x_{max} - x_{min} .$$

De la fórmula anterior se deduce que la unidad de medida de R es la misma que la de la variable.

El recorrido nos mide el grado de variabilidad de los datos de la muestra: cuanto más grande sea el resultado del recorrido, más dispersos están los datos.

2.1.2.2. Recorrido intercuartílico

La fórmula del **recorrido intercuartílico** es:

$$R_I = Q_3 - Q_1 = P_{75} - P_{25} .$$

De la fórmula anterior se deduce que la unidad de medida de R_I es la misma que la de la variable.

Cuanto más pequeño sea el resultado del recorrido intercuartílico, menos dispersión respecto de la mediana hay; es decir, los datos están menos alejados de la mediana y, por tanto, la mediana es más representativa. Pero, ¿cuándo podríamos decir que el valor del recorrido intercuartílico es pequeño? ...Como entre el primer cuartil, Q_1, y el tercer cuartil, Q_3, hay exactamente la mitad de los datos, podríamos comparar la mitad del recorrido con el recorrido intercuartílico, y podríamos decir que la mediana es representativa si R_I es menor o igual que $R/2$.

2.1.2.3. Varianza y desviación típica

I) Varianza

Si la variable se denota por X, la varianza de los datos procedentes de una muestra será denotada por s_x^2. (Si disponemos de los datos de toda la población, entonces representaremos la varianza por σ^2.)

a) **Cálculo con datos no agrupados en intervalos:**

Si $x_1, x_2, \ldots, x_n$ son los n valores de la muestra, entonces:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2 \,.$$

Si los datos son $x_1, x_2, \ldots, x_k$, y aparecen con frecuencias absolutas respectivas $f_1, f_2, \ldots, f_k$, entonces:

$$s_x^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i}{n} = \frac{\sum_{i=1}^{k} x_i^2 f_i}{n} - \overline{x}^2 \,.$$

De las fórmulas anteriores se deduce que la unidad de medida de s_x^2 es la misma que la de la variable elevada al cuadrado.

b) **Cálculo con datos agrupados en intervalos:**

La fórmula es la misma que la anterior, siendo x_i la marca de clase del intervalo $(\ell_i, \ell_{i+1}]$ y f_i su correspondiente frecuencia absoluta.

II) Desviación típica

Si la variable se denota por X, la desviación típica de los datos procedentes de una muestra será denotada por s_x. (Si disponemos de los datos de toda la población, entonces representaremos la desviación típica por σ.)

La fórmula de la **desviación típica** es:

$$s_x = \sqrt{\text{Varianza}} \,.$$

De la fórmula anterior se deduce que la unidad de medida de s_x es la misma que la de la variable.

Cuanto más pequeño sea el resultado de la desviación típica, menos dispersión respecto de la media hay; es decir, los datos están menos alejados de la media y, por tanto, la media es más representativa. Pero, ¿cuándo podríamos decir que el resultado de la desviación típica es pequeño? ...Como entre $\overline{x} - s$ y $\overline{x} + s$ hay, para la mayoría de las variables, más de las dos terceras partes de los datos, podríamos comparar la amplitud del intervalo $(\overline{x} - s, \overline{x} + s)$ con los dos tercios del recorrido; es decir, podríamos comparar el resultado de $2\,s$ con el resultado de $2\,R/3$, lo que es lo mismo que comparar s con $R/3$. En consecuencia, podríamos decir que la media es representativa si s es menor o igual que $R/3$.

III) Cuasi-varianza o varianza corregida

Se utiliza, sobre todo, en Estadística Inferencial.

Si la variable se denota por X, la cuasi-varianza o varianza corregida de los datos procedentes de una muestra será denotada por S_x^2.

a) **Cálculo con datos no agrupados en intervalos:**
Si $x_1, x_2, \ldots, x_n$ son los n valores de la muestra, entonces:

$$S_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2}{n-1}\,.$$

Si los datos son $x_1, x_2, \ldots, x_k$, y aparecen con frecuencias absolutas respectivas $f_1, f_2, \ldots, f_k$, entonces:

$$S_x^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i}{n-1} = \frac{\left(\sum_{i=1}^{k} x_i^2 f_i\right) - n\overline{x}^2}{n-1}\,.$$

De las fórmulas anteriores se deduce que la unidad de medida de S_x^2 es la misma que la de la variable elevada al cuadrado.

b) **Cálculo con datos agrupados en intervalos:**
La fórmula es la misma que la anterior, siendo x_i la marca de clase del intervalo $(\ell_i, \ell_{i+1}]$ y f_i su correspondiente frecuencia absoluta.

Relación entre la varianza y la cuasi-varianza:

$$n\, s_x^2 = (n-1)\, S_x^2\,.$$

IV) Cuasi-desviación típica o desviación típica corregida

Se utiliza, sobre todo, en Estadística Inferencial.

La fórmula de la **cuasi-desviación típica** es:

$$S_x = \sqrt{\text{Cuasi-varianza}}\,.$$

De la fórmula anterior se deduce que la unidad de medida de S_x es la misma que la de la variable.

2.2. Ejemplos que se van a resolver en clase

Ejemplo 2.1. Observamos la edad de 8 alumnos de clase. Calculamos la mediana. Interpretamos el resultado. Determinamos la frecuencia acumulada absoluta de la mediana y la comparamos con el valor de $n/2$.

Ejemplo 2.2. Observamos la edad de 9 alumnos de clase. Calculamos la mediana. Interpretamos el resultado. Determinamos la frecuencia acumulada absoluta de la mediana y la comparamos con el valor de $n/2$.

Ejemplo 2.3. La distribución de frecuencias de las calificaciones de 13 alumnos en un determinado examen viene dada por la tabla siguiente. Calcular la mediana. Interpretar el resultado. Determinar la frecuencia acumulada absoluta de la mediana y compararla con el valor de $n/2$.

Tabla 2.1

x_i	f_i	F_i
2	2	2
4	3	5
6	5	10
8	3	13

Ejemplo 2.4. La distribución de frecuencias de las calificaciones de 12 alumnos en un determinado examen viene dada por la tabla siguiente. Calcular la mediana. Interpretar el resultado. Determinar la frecuencia acumulada absoluta de la mediana y compararla con el valor de $n/2$.

Tabla 2.2

x_i	f_i	F_i
2	1	1
4	5	6
6	4	10
8	2	12

Ejemplo 2.5. En una biblioteca se observa el tiempo (en días) que tardan los proveedores en suministrar las peticiones de una determinada biblioteca.

Tabla 2.3

Nº de días	2	3	4	5	6	7	8	10	12	14
Nº de proveedores	5	10	15	12	8	3	3	2	1	1

a) ¿Cuál es la variable estadística que se observa? ¿De qué tipo es dicha variable? ¿Cuáles son los individuos a los que se les observa dicha variable? ¿Cuál es el tamaño muestral?

b) Calcular la mediana. Interpretar el resultado. Determinar la frecuencia acumulada absoluta de la mediana y compararla con el valor de $n/2$.

Ejemplo 2.6. En una muestra de libros se observa el número de referencias bibliográficas que contienen. Nos han proporcionado los datos agrupados en intervalos:

Tabla 2.4	
Nº de referencias	Nº de libros
(0,10]	19
(10,20]	23
(20,30]	12
(30,40]	10
(40,50]	8

a) ¿Cuál es la variable estadística que se observa? ¿De qué tipo es dicha variable? ¿Cuáles son los individuos a los que se les observa dicha variable? ¿Cuál es el tamaño muestral?

b) Calcular el valor aproximado de la mediana a partir del gráfico de frecuencias acumuladas absolutas.

c) Calcular la mediana mediante la fórmula. Interpretar el resultado.

Ejemplo 2.7. Con los datos de la Tabla 2.3 calcular: el primer decil, el primer cuartil, el tercer cuartil y el noveno decil. Interpretar los resultados.

Ejemplo 2.8. Con los datos de la Tabla 2.4 calcular: el primer decil, el primer cuartil, el tercer cuartil y el noveno decil. Interpretar los resultados.

Ejemplo 2.9. Calcular la media de los datos de la Tabla 2.3.

Ejemplo 2.10. Calcular la media de los datos de la Tabla 2.4.

Ejemplo 2.11. ¿Cuál es el grado de dispersión de los datos de la Tabla 2.3? Razonar la respuesta.

Ejemplo 2.12. ¿Cuál es el grado de dispersión de los datos de la Tabla 2.4? Razonar la respuesta.

Ejemplo 2.13. Con los datos de la Tabla 2.3 ¿cuál es el grado de representatividad de la mediana: muy fuerte, fuerte, regular, débil o muy débil? Razonar la respuesta.

Ejemplo 2.14. Con los datos de la Tabla 2.4 ¿cuál es el grado de representatividad de la mediana: muy fuerte, fuerte, regular, débil o muy débil? Razonar la respuesta.

Ejemplo 2.15. Con los datos de la Tabla 2.3 ¿cuál es el grado de representatividad de la media: muy fuerte, fuerte, regular, débil o muy débil? Razonar la respuesta.

Ejemplo 2.16. Con los datos de la Tabla 2.4 ¿cuál es el grado de representatividad de la media: muy fuerte, fuerte, regular, débil o muy débil? Razonar la respuesta.

2.3. Actividades de aplicación de los contenidos

2.3.1. Problemas propuestos

Problema 2.1. Se preguntó a varias personas, elegidas al azar, el número de periódicos distintos que leían trimestralmente, y se obtuvo las siguientes respuestas:

Nº de periódicos	0	1	2	3	4	5	6	7
Nº de lectores	7	13	18	15	11	6	4	2

a) Dibujar el gráfico de frecuencias acumuladas absolutas.

b) Calcular la mediana e interpretar su resultado.

c) ¿Cuál es el grado de representatividad de la mediana: muy poco representativa, poco, regular, bastante o muy representativa?

Problema 2.2. El número de personas que visitan diariamente una biblioteca fue observado durante 74 días elegidos al azar, y los resultados fueron:

Nº de personas	47	59	62	64	71	76	78	80
Nº de días	4	6	10	17	16	10	7	4

a) Hallar la media.

b) Determinar la mediana e interpretar su resultado.

c) Calcular la medida de dispersión adecuada para medir el grado de representatividad de la media. Interpretar su resultado.

d) Calcular la medida de dispersión adecuada para medir el grado de representatividad de la mediana. Interpretar su resultado.

Problema 2.3. La edad de las personas que aprobaron la oposición de auxiliar de biblioteca en España en un determinado año tiene la siguiente distribución:

Edad	[20,25]	(25,30]	(30,35]	(35,40]	(40,50]	(50,60]
Nº de personas	41	123	44	13	7	3

a) Dibujar el gráfico de frecuencias acumuladas absolutas. A partir de este gráfico, determinar el valor aproximado de la mediana. Calcular, después, el valor de la mediana con la fórmula estudiada.

b) ¿Cuál es el grado de representatividad de la mediana? Justificar la respuesta.

Problema 2.4. Los siguientes datos corresponden al número mensual de nuevos socios de una determinada biblioteca:

27	40	12	3	30	16	20	21	30	12
45	18	25	22	35	24	37	12	21	7
35	17	21	27	14	15	25	45	12	24

a) Determinar la distribución de frecuencias y dibujar el polígono de frecuencias absolutas.

b) Calcular la media.

c) Hallar la mediana e interpretar su resultado.

Problema 2.5. El número de veces que fueron consultados 60 artículos de investigación archivados en una hemeroteca, durante un determinado año, viene dado por la siguiente tabla:

8	25	20	4	19	3	21	2	20	22
23	9	1	24	21	22	20	2	22	21
2	24	21	9	3	21	22	3	22	3
12	6	20	2	26	46	2	4	10	37
14	9	7	25	50	26	38	46	36	1
7	1	35	23	45	36	5	65	46	37

Agrupar los datos en intervalos de la misma amplitud, y calcular, a partir de esta clasificación, el valor de la medida de posición que resulte más representativa del conjunto total de los datos.

Problema 2.6. A continuación se ofrecen los datos correspondientes al tiempo de espera (en minutos) de 50 usuarios de una biblioteca hasta que son atendidos por algún miembro del personal de ésta.

1	3	5	20	21	4	7	9	10	12
20	18	6	4	13	11	10	13	15	9
4	20	2	22	8	6	11	4	8	6
5	18	19	20	7	15	16	13	12	14
7	10	5	24	11	8	9	10	11	7

a) Determinar la distribución de frecuencias.

b) Calcular la media.

c) Hallar la mediana e interpretar su resultado.

d) Agrupar los datos en intervalos de distinta amplitud, y calcular, a partir de esta nueva clasificación, las mismas medidas descriptivas de los dos apartados anteriores. Comparar los resultados.

2.3.2. Soluciones de los problemas propuestos

Solución del problema 2.1. La distribución de frecuencias es:

x_i	f_i	F_i
0	7	7
1	13	20
2	18	38
3	15	53
4	11	64
5	6	70
6	4	74
7	2	76

a) *Gráfico de frecuencias acumuladas absolutas*: es la representación gráfica de las frecuencias acumuladas absolutas, F, para todo valor numérico, x. Es una gráfica en forma de "escalera".

b) Mediana=$M_e = 2'5$ periódicos. Su interpretación es la siguiente: El valor $2'5$ deja por debajo la mitad de los datos de la muestra; es decir, el valor $2'5$ deja por debajo 38 datos.

c) Como el recorrido intercuartílico es $R_I = 3$ periódicos y la mitad del recorrido es $R/2 = 3'5$ periódicos, entonces se cumple que R_I es un poco menor que $R/2$ y, como consecuencia, la mediana es bastante representativa.

Solución del problema 2.2.

a) Media=$\overline{x} = 67'7297$ personas.

b) Mediana=$M_e = 67'5$ personas. Su interpretación es la siguiente: El valor $67'5$ deja por debajo la mitad de los datos de la muestra; es decir, el valor $67'5$ deja por debajo 37 datos.

c) La desviación típica es $s_x = 8'1677$ personas. Como $R/3 = 11$, entonces se cumple que s_x es bastante menor que $R/3$ y, como consecuencia, la media es bastante representativa.

d) El recorrido intercuartílico es $R_I = 14$ personas. Como $R/2 = 16'5$, entonces R_I es bastante menor que $R/2$ y, como consecuencia, la mediana es bastante representativa.

Solución del problema 2.3.

a)
- *Gráfico de frecuencias acumuladas absolutas*: se sitúan los puntos que resultan de tomar en el eje horizontal los extremos superiores de los intervalos de clase, y en el eje vertical sus correspondientes frecuencias acumuladas absolutas, uniendo después dichos puntos mediante segmentos rectilíneos.
- A partir del gráfico anterior se deduce que la mediana es aproximadamente igual a 28 años.
- Con la fórmula se obtiene que la mediana es $M_e = 28'0285$ años.

b) El recorrido intercuartílico es $R_I = 5'37$ años. Como $R/2 = 20$ entonces R_I es mucho menor que $R/2$ y, como consecuencia, la mediana es muy representativa.

Solución del problema 2.4.

a)
- La distribución de frecuencias (conteniendo las columnas que posteriormente necesitaremos) es:

x_i	f_i	F_i	$x_i f_i$
3	1	1	3
7	1	2	7
12	4	6	48
14	1	7	14
15	1	8	15
16	1	9	16
17	1	10	17
18	1	11	18
20	1	12	20
21	3	15	63
22	1	16	22
24	2	18	48
25	2	20	50
27	2	22	54
30	2	24	60
35	2	26	70
37	1	27	37
40	1	28	40
45	2	30	90
suma			692

- *Polígono de frecuencias absolutas*: se sitúan los puntos que resultan de tomar en el eje horizontal los distintos valores de la variable, x_i, y en el eje vertical sus correspondientes frecuencias absolutas, f_i, uniendo después los puntos mediante segmentos rectilíneos.

b) Media=$\overline{x} = 23'0\,\widehat{6}$ socios.

c) Mediana=$M_e = 21'5$ socios. Su interpretación es la siguiente: El valor $21'5$ deja por debajo la mitad de los datos de la muestra; es decir, el valor $21'5$ deja por debajo 15 datos.

Solución del problema 2.5. La distribución de frecuencias con datos agrupados en intervalos de la misma amplitud es:

$(\ell_i, \ell_{i+1}]$	f_i	F_i
(0,10]	23	23
(10,20]	7	30
(20,30]	18	48
(30,40]	6	54
(40,50]	5	59
(50,60]	0	59
(60,70]	1	60

Como la dispersión es grande, la medida de posición más adecuada es la mediana. Con los datos agrupados en estos intervalos de clase, el valor de la mediana es $M_e = 20$ consultas en un año.

Solución del problema 2.6.

a) La distribución de frecuencias es:

x_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18	19	20	21	22	24
f_i	1	1	1	4	3	3	4	3	3	4	4	2	3	1	2	1	2	1	4	1	1	1
F_i	1	2	3	7	10	13	17	20	23	27	31	33	36	37	39	40	42	43	47	48	49	50
$x_i f_i$	1	2	3	16	15	18	28	24	27	40	44	24	39	14	30	16	36	19	80	21	22	24

b) Media=$\overline{x} = 10'86$ minutos.

c) Mediana=$M_e = 10$ minutos. Su interpretación es la siguiente: El valor 10 deja por debajo la mitad de los datos de la muestra; es decir, el valor 10 deja por debajo 25 datos.

d) Una posible agrupación de los datos en intervalos de distinta amplitud es:

$(\ell_i, \ell_{i+1}]$	f_i	x_i	$x_i f_i$	F_i
(0,4]	7	2	14	7
(4,6]	6	5	30	13
(6,8]	7	7	49	20
(8,10]	7	9	63	27
(10,12]	6	11	66	33
(12,15]	6	$13'5$	81	39
(15,19]	4	17	68	43
(19,24]	7	$21'5$	$150'5$	50
suma			$521'5$	

Con esta clasificación, los resultados de las medidas descriptivas anteriores son:

- Media=$\overline{x} = 10'43$ minutos.
- Mediana=$M_e = 9'4286$ minutos.

Los verdaderos resultados de la media y de la mediana son los calculados en los apartados **b)** y **c)**, respectivamente.

2.4. PRÁCTICA 2: ESTADÍSTICA DESCRIPTIVA

2.4.1. Distribución de frecuencias

Con ***Minitab***, para determinar la distribución de frecuencias de una (o más variables) utilizamos la opción **Stat⇒Tables ⇒Tally Individual Variables**.

Para practicar esta opción, podemos utilizar el archivo de datos (Worksheet) **Pulse.mtw**. En primer lugar tenemos que abrir dicha hoja de datos (si no la tenemos abierta ya). Recordemos que su contenido fue recogido en una clase de 92 alumnos. De cada estudiante se observó su pulso antes de correr, **Pulse1**; su pulso después de correr, **Pulse2**; si corrió o no, **Ran** (1=Sí corrió, 2=No corrió); si es fumador o no, **Smokes** (1=Sí fuma, 2=No fuma); el sexo, **Sex** (1=Hombre, 2=Mujer); su altura en pulgadas, **Height**; su peso en libras, **Weight**; y su nivel de actividad física, **Activity** (0=Ninguna actividad, 1=Baja, 2=Media, 3=Alta).

De la hoja de datos **Pulse.mtw** vamos a averiguar la distribución de frecuencias de todas las variables. Para ello, seleccionamos la opción **Stat⇒Tables⇒Tally Individual Variables**; en el recuadro **Variables** seleccionamos, de la lista de variables de la izquierda, todas las columnas. En **Display** activamos los cuatro tipos de frecuencias que aparecen: **Counts** (frecuencia absoluta), **Percents** (porcentaje), **Cumulative counts** (frecuencia acumulada absoluta) y **Cumulative percents** (porcentaje acumulado). Por último, pulsamos en **OK**.

En la ventana de sesión podemos observar, por ejemplo:

- Hay 57 personas (de las 92 que componen la muestra) que no corrieron; es decir, 57 es la frecuencia absoluta de **Ran**=2.
- Hay 64 personas (de las 92 que componen la muestra) que no fuman; es decir, 64 es la frecuencia absoluta de **Smokes**=2.
- El $38'04\,\%$ del total de personas de la muestra son mujeres; es decir, $38'04\,\%$ es el porcentaje de **Sex**=2.
- 46 personas (la mitad de las personas que componen la muestra) tienen 70 pulsaciones o menos antes de correr; es decir, 46 es la frecuencia acumulada absoluta de **Pulse1**=70.
- El $75\,\%$ de las personas (las tres cuartas partes del total) tienen 84 pulsaciones o menos después de correr; es decir, $75\,\%$ es el porcentaje acumulado de **Pulse2**=84.

2.4.2. Representaciones gráficas

En ***Minitab*** la mejor opción para hacer representaciones gráficas es usar el menú **Graph**.

Una utilidad importante de todos los gráficos creados a través del menú **Graph** es que haciendo *clic* sobre ellos con el botón derecho del ratón y activando la opción **Update Graph Automatically** del menú contextual que aparece, el gráfico cambia automáticamente al modificar los datos con que se han construido (ya sea añadiendo, modificando o eliminando datos).

2.4.2.1. Gráfico de sectores o de *pastel*

El *gráfico de sectores* se construye de la siguiente forma: se divide el área de un círculo en sectores circulares de ángulos proporcionales a las frecuencias absolutas de las clases. Se utiliza cuando la variable es cualitativa o cuantitativa discreta con pocos resultados distintos.

En ***Minitab***, este gráfico se obtiene con la opción **Graph⇒Pie Chart**.

Por ejemplo, vamos a hacer el gráfico de sectores de los datos de la columna **Activity** de la hoja de datos **Pulse.mtw**. Para ello, en el cuadro de diálogo que resulta al seleccionar **Graph⇒Pie Chart**, dejamos activada la opción **Chart counts of unique values** y seleccionamos la columna **Activity** en el recuadro **Categorical variables**. Podemos cambiar el aspecto que tendría el gráfico por defecto, pulsando en los botones que aparecen en este cuadro de diálogo: **Pie Options**, **Labels**, **Multiple Graphs** y **Data Options**. En principio, podríamos dejar todas las opciones por defecto a la hora de realizar este primer gráfico.

El gráfico obtenido podemos copiarlo en el portapapeles, haciendo *clic* sobre el gráfico con el botón derecho del ratón y seleccionando, del menú contextual que resulta, la opción **Copy Graph**. De esta manera, podríamos pegarlo en otro programa bajo Windows, por ejemplo, uno de edición de gráficos. También podemos almacenarlo en la ventana de proyecto, **Proyect Manager** (concretamente en el directorio **ReportPad**) haciendo *clic* sobre el gráfico con el botón derecho del ratón y seleccionando, del menú contextual que resulta, la opción **Append Graph to Report**. También tenemos la posibilidad de grabarlo en varios formatos (gráfico propio de Minitab, **mgf**, **jpg**, **png**, **bmp**, etc.). Para ello solo tenemos que cerrar el gráfico (botón ☒) y pulsar en **Sí** cuando ***Minitab*** nos pregunte si queremos guardar el gráfico en un archivo aparte.

Una vez obtenido el gráfico es posible cambiar su aspecto. Para ello, hacemos doble *clic* sobre la parte del gráfico que queremos cambiar. Aparece, entonces, una nueva ventana que nos permite hacer dicha transformación. Para practicar, vamos a cambiar el gráfico de sectores de los datos de la columna **Activity** de la siguiente manera:

- Que el título sea *Gráfico de sectores de la variable 'Actividad Física'*, en letra Verdana, cursiva, negrita, de color rojo oscuro y con un tamaño de 10 puntos.
- Que junto a los sectores circulares aparezca la frecuencia absoluta de cada categoría (*clic* sobre uno de los sectores circulares con el botón derecho del ratón; opción **Add**, **Slice Labels**; activamos **Frequency** y pulsamos en **OK**).

2.4.2.1.1. Diagrama de sectores cuando tenemos en una columna las categorías de una variable y en otra columna las correspondientes frecuencias

Vamos a aprender a hacer (con ***Minitab***) un diagrama de sectores cuando tenemos en una columna las categorías de una variable y en otra columna las frecuencias absolutas de dichas categorías. Por ejemplo, vamos a realizar el diagrama de sectores de los datos de la Figura 7, correspondientes a los idiomas en que están escritos los libros de los estantes de una determinada biblioteca.

idioma	nº de estantes (frecuencia)
francés	78
alemán	47
ruso	20
español	30

Figura 7: Idioma de los libros de una biblioteca

Como estos datos no tienen nada que ver con los datos del archivo **Pulse.mtw**, abrimos una nueva hoja de datos con la opción **File⇒New**. En el cuadro de diálogo que aparece seleccionamos **Minitab Woorksheet**. A esta nueva hoja de datos ***Minitab*** le asignará el nombre *Worksheet* J, siendo J un

número natural. A continuación introducimos los datos tal como se muestra en la Figura 7. Luego guardamos esta hoja de datos con el nombre **IdiomaLibros.mtw** (**File⇒Save Current Worksheet As**). Para dibujar el diagrama de sectores seleccionamos **Graph⇒Pie Chart**. En el cuadro de diálogo resultante, activamos la opción **Chart values from a table**; seleccionamos la columna **idioma** en el recuadro **Categorical Variable**; seleccionamos la columna **nº de estantes (frecuencia)** en el recuadro **Summary variables** y pulsamos en **OK**. Como ya sabemos, podemos modificar este gráfico.

2.4.2.2. Diagrama de barras simple

El *diagrama de barras simple* se construye de la siguiente manera: se sitúan en el eje horizontal las clases y sobre cada una de ellas se levanta un segmento rectilíneo (o un rectángulo) de altura igual a la frecuencia (absoluta o relativa) o al porcentaje de cada clase. Se utiliza cuando la variable es cualitativa o cuantitativa discreta con pocos resultados distintos.

En ***Minitab*** este gráfico se obtiene con la opción **Graph⇒Bar Chart**.

Por ejemplo, vamos a hacer el diagrama de barras de los datos de la columna **Activity** de la hoja de datos **Pulse.mtw**. En primer lugar, es necesario tener abierta y activada dicha hoja de datos. Para dibujar el diagrama de barras seleccionamos **Graph⇒Bar Chart**; dejamos activada la opción **Counts of unique values** del recuadro **Bars represent** y dejamos también activado el modelo **Simple** del diagrama de barras. En el cuadro de diálogo resultante, seleccionamos la columna **Activity** en el recuadro **Categorical Variables**. Como las categorías son números concretos (0, 1, 2 y 3) es más riguroso que, **en vez de barras, aparezcan solamente segmentos rectilíneos verticales** (o líneas de proyección); para hacerlo, pulsamos el botón **Data View** y en el cuadro de diálogo resultante activamos **solamente** la opción **Project lines** (las otras tres opciones deben estar desactivadas).

Igual que ocurría con el diagrama de sectores, una vez obtenido el diagrama de barras podemos copiarlo en el portapapeles o almacenarlo en el apartado **ReportPad** de la ventana **Proyect Manager** o grabarlo en un archivo aparte.

Podemos observar, además, que si hacemos *clic* sobre el gráfico (para activarlo) y luego pasamos el ratón por encima de las barras, se nos indica la frecuencia absoluta de cada categoría.

También es posible cambiar su aspecto, una vez obtenido, haciendo doble *clic* sobre la parte del gráfico que queremos cambiar. Para practicar, vamos a modificar diagrama de barras de los datos de la columna **Activity** de la siguiente manera:

- Que el título sea *Diagrama de barras de la variable 'Actividad Física'*, en letra Comic Sans MS, cursiva, negrita, de color rojo y con un tamaño de 11 puntos.
- Que las barras (líneas) sean de color rojo y de un tamaño (grosor) de 3 puntos.
- Que en el eje vertical se muestren 13 marcas (*ticks*), en letra Arial, no negrita, de color rojo y con un tamaño de 10 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, en letra Arial, cursiva, no negrita, de color rojo y con un tamaño de 9 puntos.
- Que el texto del eje horizontal sea *Actividad Física (0=Ninguna, 1=Baja, 2=Media, 3=Alta)*, en letra Arial, cursiva, no negrita, de color rojo y con un tamaño de 8 puntos.
- Que en la parte superior de cada barra aparezca la frecuencia absoluta de cada categoría (*clic* sobre una de las barras con el botón derecho del ratón, opción **Add**, **Data Labels**, dejar activado **Use y-values labels**).

2.4.2.2.1. Diagrama de barras cuando tenemos en una columna las categorías de una variable y en otra columna las correspondientes frecuencias

Vamos a aprender a hacer (con ***Minitab***) un diagrama de barras cuando tenemos en una columna las categorías de una variable y en otra columna las frecuencias absolutas de dichas categorías. Por ejemplo, vamos a realizar el diagrama de barras de los datos de la Figura 7, correspondientes a los idiomas en que están escritos los libros de los estantes de una determinada biblioteca. En primer lugar, es necesario tener abierta y activada dicha hoja de datos (**IdiomaLibros.mtw**). Para dibujar el diagrama de barras seleccionamos **Graph⇒Bar Chart**, activamos la opción **Values from a table** del apartado **Bars represent**; activamos el modelo **Simple** del apartado **One column of values** y pulsamos en **OK**. En el cuadro de diálogo resultante, seleccionamos la columna **nº de estantes (frecuencia)** en el recuadro **Graph variables**; seleccionamos la columna **idioma** en el recuadro **Categorical Variable** y pulsamos en **OK**. Como ya sabemos, podemos modificar este gráfico.

2.4.2.3. Diagrama de barras agrupado (o apilado)

Con la opción **Graph⇒Bar Chart** existe la posibilidad de seleccionar una nueva variable para determinar las barras dentro de cada grupo; esto se realiza seleccionando **Cluster** (para un diagrama de barras agrupado según los resultados de otra variable) o **Stack** (para un diagrama de barras apilado según los resultados de otra variable).

Por ejemplo, con el archivo de datos **Pulse.mtw** vamos a hacer el diagrama de barras de los datos de la columna **Activity** en grupos definidos por la variable **Sex**. En primer lugar, es necesario tener abierta y activada dicha hoja de datos. Para dibujar el citado diagrama de barras seleccionamos **Graph⇒Bar Chart**; dejamos activada la opción **Counts of unique values** del recuadro **Bars represent**; y activamos el modelo **Cluster** del diagrama de barras. En el siguiente cuadro de diálogo seleccionamos, de la lista de variables de la izquierda, las columnas **Activity** y **Sex** (en este orden) para ponerlas en el recuadro **Categorical variables**. Una vez obtenido dicho diagrama de barras es conveniente modificarlo para que sea más explicativo; por ejemplo, vamos a hacer lo siguiente:

- Que el título sea *Diagrama de barras de la variable 'Actividad Física' en grupos definidos por la variable 'Sexo'*, en letra Verdana, negrita, de color morado y con un tamaño de 9 puntos.
- Que las barras tengan distinto color según los resultados de la variable **Sex** y que aparezca una leyenda explicativa (doble *clic* sobre una de las barras, en el cuadro de diálogo resultante seleccionamos la carpeta **Groups**, en el recuadro **Assign attributes by categorical variables** seleccionamos la variable **Sex** y pulsamos en **OK**).
- Que en el eje vertical se muestren 10 marcas (*ticks*), en letra Verdana, no negrita, de color morado y con un tamaño de 10 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, en letra Verdana, no negrita, de color morado y con un tamaño de 11 puntos.
- Que en el eje horizontal todo esté escrito con la fuente Verdana, no negrita, de color morado y con un tamaño de 9 puntos. Que en dicho eje aparezcan los nombres de las variables en español: *Actividad Física* en vez de *Activity*, y *Sexo* en vez de *Sex*. Que en el mismo eje los resultados de la variable *Sex* no sean 1 y 2 sino *Hombre* y *Mujer*. Y los resultados de la variable *Activity* no sean 0, 1, 2 y 3 sino *Ninguna*, *Poca*, *Media* y *Alta*.

2.4.2.3.1. Diagrama de barras agrupado (o apilado) cuando tenemos los datos en una tabla de doble entrada

Vamos a aprender a hacer un diagrama de barras agrupado (o apilado) cuando tenemos los datos en una tabla de doble entrada. Por ejemplo, vamos a realizar el diagrama de barras agrupado de los datos de la Figura 8, correspondientes al número de citas en diferentes campos de investigación y en tres distintos años.

	Campo investigación	1970	1980	1990
1	sociología	330	414	547
2	economía	299	393	295
3	política	115	357	137
4	psicología	329	452	258

Figura 8: Citas anuales en distintos campos de investigación

En primer lugar, abrimos una nueva hoja de datos con la opción **File⇒New**. En el cuadro de diálogo que aparece seleccionamos **Minitab Woorksheet**. A continuación introducimos los datos tal como se muestra en la Figura 8. Luego guardamos esta hoja de datos con el nombre **Citas.mtw**. Para dibujar el diagrama de barras agrupado es necesario tener abierta y activada dicha hoja de datos. Luego seleccionamos **Graph⇒Bar Chart**, activamos la opción **Values from a table** del apartado **Bars represent**; activamos el modelo **Cluster** del apartado **Two-way table** y pulsamos en **OK**. En el cuadro de diálogo resultante, seleccionamos las columnas **1970**, **1980** y **1990** en el recuadro **Graph variables**; seleccionamos la columna **Campo investigación** en el recuadro **Row labels**; activamos **Rows are outermost categories and columns are innermost** y, por último, pulsamos en **OK**. Como ya sabemos, podemos modificar este gráfico.

2.4.2.4. Polígono de frecuencias

El *polígono de frecuencias* se construye de la siguiente manera: se sitúan los puntos que resultan de tomar en el eje horizontal los distintos valores de la variable y en el eje vertical sus correspondientes frecuencias (no acumuladas), uniendo después los puntos mediante segmentos rectilíneos.

En ***Minitab*** este gráfico se obtiene con la opción **Graph⇒Bar Chart**.

Por ejemplo, vamos a hacer el polígono de frecuencias de los datos de la columna **Pulse2** de la hoja de datos **Pulse.mtw**. En primer lugar, es necesario tener abierta y activada dicha hoja de datos. Para dibujar el polígono de frecuencias seleccionamos **Graph⇒Bar Chart**; dejamos activada la opción **Counts of unique values** del recuadro **Bars represent**, dejamos también activado el modelo **Simple** y pulsamos en **OK**. En el cuadro de diálogo resultante, seleccionamos la columna **Pulse2** en el recuadro **Categorical Variables**. Activamos el botón **Data View** y en el cuadro de diálogo resultante activamos **solamente** la opción **Connect line** (las otras tres opciones deben estar desactivadas).

Igual que ocurría con los gráficos anteriores, una vez obtenido el polígono de frecuencias podemos copiarlo en el portapapeles, almacenarlo en el apartado **ReportPad** de la ventana **Proyect Manager** o grabarlo en un archivo aparte. También es posible cambiar su aspecto haciendo doble *clic* sobre la parte del gráfico que queremos cambiar.

2.4.2.4.1. Polígono de frecuencias cuando tenemos en una columna las categorías de una variable y en otra columna las correspondientes frecuencias

Vamos a aprender a hacer (con ***Minitab***) un polígono de frecuencias cuando tenemos en una columna las categorías de una variable y en otra columna las frecuencias absolutas de dichas categorías.

Por ejemplo, vamos a realizar el polígono de frecuencias de los datos de la Figura 9, correspondientes al número de días que tardan los proveedores en suministrar las peticiones de una determinada biblioteca.

nº de días	nº de proveedores (frecuencia)
2	5
3	10
4	15
5	12
6	8
7	3
8	3
10	2
12	1
14	1

Figura 9: Tiempo (en días) que tardan los proveedores en suministrar las peticiones de una biblioteca

En primer lugar, abrimos una nueva hoja de datos con la opción **File⇒New**. En el cuadro de diálogo que aparece seleccionamos **Minitab Woorksheet**. A continuación introducimos los datos tal como se muestra en la Figura 9. Luego guardamos esta hoja de datos con el nombre **Proveedores.mtw**. Para dibujar el polígono de frecuencias es necesario tener abierta y activada dicha hoja de datos. Luego seleccionamos **Graph⇒Bar Chart**, activamos la opción **Values from a table** del apartado **Bars represent**; activamos el modelo **Simple** del apartado **One column of values** y pulsamos en **OK**. En el cuadro de diálogo resultante, seleccionamos la columna **nº de proveedores (frecuencia)** en el recuadro **Graph variables** y seleccionamos la columna **nº de días** en el recuadro **Categorical variable**. Activamos el botón **Data View** y en el cuadro de diálogo resultante activamos **solamente** la opción **Connect line** (las otras tres opciones deben estar desactivadas). Como ya sabemos, podemos modificar este gráfico.

2.4.2.5. Histograma

El *histograma* se construye de la siguiente manera: se sitúan en el eje horizontal los intervalos de clase y sobre cada uno se levanta un rectángulo de área igual o proporcional a la frecuencia absoluta.

En ***Minitab*** se puede obtener el histograma de una variable con la opción **Graph⇒Histogram**. Esta opción ofrece 4 tipos: **Simple**, **With Fit**, **With Outline and Groups** y **With Fit and Groups**.

Por ejemplo, podemos hacer el histograma de la variable **Weight** de la hoja de datos **Pulse.mtw**. En primer lugar, es necesario tener abierta y activada dicha hoja de datos. Para realizar el citado histograma seleccionamos la opción **Graph⇒Histogram**. De las cuatro opciones que aparecen seleccionamos **Simple**. En el cuadro de diálogo resultante seleccionamos la variable **Weight** para ponerla en el recuadro **Graph variables**. Como ya sabemos, podemos cambiar el aspecto que tendría el gráfico por defecto, pulsando en los botones que aparecen en este cuadro de diálogo: **Scale**, **Labels**, **Data View**, **Multiple Graphs**

y **Data Options**. En principio, podríamos dejar todas las opciones por defecto a la hora de realizar este primer histograma.

Como también sabemos, es posible cambiar el aspecto de este gráfico una vez obtenido. Para ello, hacemos doble *clic* sobre la parte del gráfico que queremos cambiar. Aparece, entonces, una nueva ventana que nos permite hacer dicha transformación. Los cambios más usuales son: cambio en la escala del eje horizontal, cambio en el eje vertical, aspecto de las barras, intervalos sobre los que se sitúan las barras, aspecto de la ventana del gráfico y cambio en las proporciones del gráfico. Para practicar con estas opciones, vamos a cambiar el histograma de la variable **Weight** de la siguiente manera:

- Que el título sea *Histograma de la variable 'Peso'*, en letra Arial, cursiva, negrita, de color azul oscuro y con un tamaño de 10 puntos.
- Que las barras sean de color azul claro con una trama de relleno oblicua y con los bordes de color azul oscuro.
- Que haya 7 intervalos de la misma amplitud y que en el eje horizontal aparezcan los límites de los intervalos, no los puntos medios (doble *clic* sobre el eje horizontal, seleccionamos la carpeta **Binning**, activamos **Cutpoint** en **Interval Type**, activamos **Number of intervals** en **Interval Definition**, escribimos un **7** junto a esta opción y pulsamos en **OK**).
- Que el texto del eje horizontal sea *Peso de los alumnos, en libras*, en letra Arial, cursiva, no negrita, de color azul oscuro y con un tamaño de 9 puntos.
- Que en el eje vertical se muestren 13 marcas (*ticks*), en letra Arial, de color azul oscuro y con un tamaño de 8 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, en letra Arial, cursiva, no negrita, de color azul oscuro y con un tamaño de 9 puntos.

2.4.3. Medidas descriptivas de los datos

2.4.3.1. Determinación mediante la opción *Calc⇒Column Statistics*

Con esta opción solamente podemos calcular **un** estadístico de **una** variable (cada vez que lo utilicemos). Por tanto, no podemos calcular más de un estadístico. Tampoco podemos determinar un estadístico para más de una variable. Pero una ventaja de esta opción es que se puede guardar el resultado del estadístico para luego utilizarlo, cambiar el número de decimales, etc.

Los estadísticos que se pueden determinar con esta opción son:

Sum	suma	$\sum_{i=1}^{n} x_i$
Mean	media aritmética	$\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Standard deviation	cuasi-desviación típica	$S_x = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$
Minimum	mínimo dato	x_{min}
Maximum	máximo dato	x_{max}
Range	recorrido	$R = x_{max} - x_{min}$
Median	mediana=valor que deja por debajo el 50 % de los datos	
Sum of squares	suma de cuadrados	$\sum_{i=1}^{n} x_i^2$
N total	número total de casos=N nonmissing+N missing	
N nonmissing	número de casos para los cuales sabemos el resultado de la variable = n	
N missing	número de casos para los cuales no sabemos el resultado de la variable	

El resultado del estadístico calculado se puede almacenar (opcionalmente) en una constante, si lo indicamos en **Store result in**.

Por ejemplo, del archivo de datos **Pulse.mtw** vamos a determinar la cuasi-desviación típica de los datos de la columna **Height** y vamos a guardar el resultado en una constante que vamos a denominar **cuasi-desv-Altura**. Para ello, seleccionamos **Calc⇒Column Statistics**; activamos la opción **Standard deviation**; hacemos *clic* en el recuadro que hay a la derecha de **Input variable** y seleccionamos (haciendo doble *clic* sobre su nombre) la columna **Height**; en **Store result in** tecleamos **'cuasi-desv-Altura'** (con comillas simples, al principio y al final, por llevar guiones) y pulsamos en **OK**. ***Minitab*** guarda esta constante también como $K1$ (o, en general, KJ, con $J = 1, 2, 3, \ldots$). Esta constante se puede consultar, en cualquier momento, en la ventana **Proyect Manager** (concretamente, en **Worksheets\Pulse.mtw\Constants**) y puede ser utilizada en cálculos posteriores.

Importante No es posible cambiar el número de decimales de los resultados que aparecen en la ventana de sesión. Hay una forma de **aumentar el número de decimales de un resultado** solamente en el caso en que sea posible almacenar dicho resultado en una constante; es decir, si en el cuadro de diálogo en el cual estamos solicitando a ***Minitab*** que calcule dicho resultado aparece la opción de guardar el resultado. Si, por ejemplo, tenemos guardado un resultado en la constante $K1$ y queremos tener una precisión de 6 decimales, hacemos lo siguiente: seleccionamos **Data⇒Copy⇒Constants to Column**; hacemos *clic* en el recuadro que hay debajo de **Copy from constants** y seleccionamos (haciendo doble *clic* sobre su nombre) la constante $K1$; en **In current worksheet, in column** tenemos que teclear la posición de la columna que contendrá el resultado (una columna, **CJ**, que esté vacía) o el nombre que queremos darle a dicha columna. Recordemos que si el nombre contiene espacios en blanco, guiones, paréntesis, etc., hay que escribirlo entre comillas simples. Si hemos puesto un nombre a esta columna, desactivamos **Name the column containing the copied data**. Por último, pulsamos en **OK**. Una vez que tenemos la constante $K1$ copiada en una columna, podemos cambiar su formato como hemos visto anteriormente: hacemos *clic* sobre el nombre de la variable (o sobre su número de columna: CJ);

pulsamos con el botón derecho del ratón; seleccionamos **Format Column⇒Numeric**; activamos **Fixed decimal**; en **Decimal places** tecleamos 6 y pulsamos en **OK**.

2.4.3.2. Determinación mediante la opción *Stat⇒Basic Statistics⇒Display Descriptive Statistics*

La opción **Stat⇒Basic Statistics⇒Display Descriptive Statistics** permite obtener **uno o varios** estadísticos de **una o varias** variables. Además, esta opción permite calcular los estadísticos separando los valores de una variable según el valor de otra variable.

Para practicar, vamos a calcular los estadísticos más importantes de las variables **Pulse1**, **Height** y **Weight** de la hoja de datos **Pulse.mtw**. Para ello, seleccionamos **Stat⇒Basic Statistics⇒Display Descriptive Statistics** y en el recuadro **Variables** del cuadro de diálogo resultante seleccionamos, de la lista de columnas que tenemos a la izquierda, las tres variables **Pulse1**, **Height** y **Weight**. A continuación pulsamos en **Statistics**. Nos aparece un nuevo cuadro de diálogo en el cual se pueden elegir los estadísticos que queremos determinar de las variables que hemos seleccionado en el recuadro **Variables**. Haciendo *clic* sobre el botón **Help** se obtiene información sobre el significado de cada uno de estos estadísticos. Los estadísticos que podemos seleccionar son los siguientes:

Mean	media aritmética	$\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
SE of mean	error estándar de la media	$\dfrac{S_x}{\sqrt{n}}$
Standard deviation	cuasi-desviación típica	$S_x = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}}$
Variance	cuasi-varianza	S_x^2
Coefficient of variation	coeficiente de variación media	$CV = \dfrac{s_x}{\lvert\overline{x}\rvert} \cdot 100\,\%$
First quartile	primer cuartil	Q_1
Median	mediana	$M_e = Q_2$
Third quartile	tercer cuartil	Q_3
Interquartile range	recorrido intercuartílico	$R_I = Q_3 - Q_1$
Trimmed mean	media de los datos eliminando el 5 % de los menores y el 5 % de los mayores	
Sum	suma	$\sum_{i=1}^{n} x_i$
Minimum	mínimo dato	x_{min}
Maximum	máximo dato	x_{max}
Range	recorrido	$R = x_{max} - x_{min}$
N nonmissing	número de casos para los cuales sabemos el resultado de la variable = n	

N missing	número de casos para los cuales no sabemos el resultado de la variable	
N total	número total de casos=N nonmissing+N missing	
Cumulative N	número acumulado de casos (solo cuando se ha rellenado el recuadro **By variables**)	
Percent	porcentaje de casos (solo cuando se ha rellenado el recuadro **By variables**)	
Cumulative percent	porcentaje acumulado de casos (solo cuando se ha rellenado el recuadro **By variables**)	
Sum of squares	suma de cuadrados	$\sum_{i=1}^{n} x_i^2$
Skewness	coeficiente de asimetría	$g_1 = \frac{m_3}{s_x^3}$, con $m_3 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{n}$
Kurtosis	coeficiente de apuntamiento	$g_2 = \frac{m_4}{s_x^4} - 3$, con $m_4 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{n}$
MSSD	media de los cuadrados de las sucesivas diferencias	

Siguiendo con nuestro ejemplo (cálculo de los estadísticos más importantes de las variables **Pulse1**, **Height** y **Weight**), podemos seleccionar todos los estadísticos menos **Cumulative N**, **Percent** y **Cumulative percent**. En la ventana de sesión podemos comprobar, por ejemplo, que la suma de los datos de la variable **Pulse1** es 6704 y la suma de los cuadrados de los datos de la misma variable es 499546.

Con la misma hoja de datos (**Pulse.mtw**) podemos calcular los estadísticos de la variable **Pulse2** (Pulso después de correr) separando sus resultados según los valores de la variable **Ran** (¿corrió o no corrió?). Para ello, seleccionamos **Stat⇒Basic Statistics⇒Display Descriptive Statistics**; en el recuadro **Variables** del cuadro de diálogo resultante seleccionamos la variable **Pulse2**; y en **By variables (Optional)** seleccionamos la variable **Ran**. En consecuencia, en la ventana de sesión aparecen los resultados de los mencionados estadísticos de la variable **Pulse2** separados para cada grupo de resultados de la variable **Ran**. Por ejemplo, podemos comprobar que para el grupo de personas que sí corrió (**Ran=1**) la media del pulso es $92'51$ y la mediana es 88, mientras que para el grupo de personas que no corrió (**Ran=2**) la media del pulso es $72'32$ y la mediana es 70.

2.4.4. Ejercicios prácticos propuestos

Ejercicio 2.1.

a) Crea un nuevo proyecto de ***Minitab***.

b) Abre la hoja de datos **Prestamos.mtw** (datos del Ejercicio 1.1).

c) Determina la distribución de frecuencias de la variable **Intervalos PPU**.

d) Para las variables **Usuarios**, **Préstamos** y **PPU** calcula todas las medidas descriptivas que hemos estudiado en las clases teóricas.

e) Dibuja el histograma fde la variable **PPU**. Modifícalo de la siguiente forma:

- Que haya 4 intervalos de la misma amplitud y que en el eje horizontal aparezcan los límites de los intervalos (no los puntos medios).
- Que el título sea *Histograma del 'Porcentaje anual de préstamos por usuario'*, en letra Times New Roman, negrita, de color rojo oscuro y con un tamaño de 14 puntos.
- Que las barras sean de color rojo claro con una trama de relleno horizontal y con los bordes de color rojo oscuro, de tamaño 2.
- Que el texto del eje horizontal sea *Porcentaje anual de préstamos por usuario*, en letra Times New Roman, cursiva, no negrita, de color rojo oscuro y con un tamaño de 12 puntos.
- Que en el eje vertical se muestren 7 marcas (*ticks*) y que los números sean de color rojo oscuro y con un tamaño de 12 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, en letra Times New Roman, cursiva, no negrita, de color rojo oscuro y con un tamaño de 12 puntos.

f) Dibuja el gráfico de sectores de la variable **Intervalos PPU**. Modifícalo de la siguiente forma:

- Que el título sea *Gráfico de sectores de la variable 'Intervalos PPU'*, en letra Verdana, cursiva, negrita, de color azul oscuro y con un tamaño de 12 puntos.
- Que junto a los sectores circulares aparezca la frecuencia absoluta y el porcentaje de cada categoría.
- En la leyenda, tanto la fuente de la cabecera como la fuente del cuerpo sea Verdana, de color azul oscuro y con un tamaño de 10 puntos.

g) Graba el proyecto con el siguiente nombre: **Ejercicio2-1.mpj**

Ejercicio 2.2.

a) Crea un nuevo proyecto de ***Minitab***.

b) Abre la hoja de datos **Transacciones.mtw** (datos del Ejercicio 1.2).

c) Determina la distribución de frecuencias de la variable **Intervalos Porcentaje TRF**.

d) Para las variables **TR**, **TRF** y **Porcentaje TRF** calcula las medidas descriptivas siguientes: mínimo, primer cuartil, mediana, tercer cuartil, máximo, recorrido, recorrido intercuartílico, media, cuasi-varianza, cuasi-desviación típica, suma de los datos y suma de los cuadrados de los datos.

e) Calcula la media, la mediana y la cuasi-desviación típica de la variable **Porcentaje TRF** separando sus resultados según los valores de la variable **Tipo Biblioteca**.

f) Dibuja el diagrama de barras de la variable **Intervalos Porcentaje TRF** en grupos definidos por la variable **Tipo Biblioteca**. Modifícalo de la siguiente forma:

- Que las barras tengan distinto color según los resultados de la variable **Tipo Biblioteca** y que aparezca una leyenda explicativa.
- Que el título sea *Diagrama de barras agrupado*, escrito con letra Arial, negrita, de color rojo oscuro y con un tamaño de 16 puntos.
- Que el texto del eje vertical sea *Frecuencia absoluta*, escrito con letra Arial, negrita, de color rojo oscuro y con un tamaño de 12 puntos.
- Que en el eje horizontal todo esté escrito con la fuente Arial, de color rojo oscuro y con un tamaño de 10 puntos.

g) Graba el proyecto con el siguiente nombre: **Ejercicio2-2.mpj**

Ejercicio 2.3. El gasto de una biblioteca, en euros, durante un año determinado, es:

Gasto en personal	6570
Gasto en libros	3450
Otros gastos	2380

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **GastoBiblioteca.mtw**

c) Haz un diagrama de barras y modifícalo a tu gusto.

d) Haz un gráfico de sectores y modifícalo a tu gusto.

e) Graba el proyecto con el siguiente nombre: **Ejercicio2-3.mpj**

Ejercicio 2.4. La estadística de fotocopias de 4 bibliotecas (A, B, C y D), durante un año, está recogida en la siguiente tabla:

	A	B	C	D
Reproducción de catálogos	16110	3640	0	3400
Trabajo del personal de la biblioteca	63350	11360	3080	5500
Préstamo interbibliotecario	2600	1090	560	250
Copias para usuarios de la biblioteca	43540	58040	1980	0

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **TipoFotocopias.mtw**

c) Haz un diagrama de barras agrupado y modifícalo a tu gusto.

d) Graba el proyecto con el siguiente nombre: **Ejercicio2-4.mpj**

Ejercicio 2.5. El número de palabras clave (*keywords*) de 72 artículos de investigación viene dado por:

Nº de palabras clave	3	4	5	6	7	8	9	10	11	12	13	14
Nº de artículos	5	8	12	7	9	9	10	5	3	2	1	1

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **Keywords.mtw**

c) Haz un diagrama de barras en el cual las barras sean segmentos rectilíneos verticales. Modifícalo a tu gusto.

d) Graba el proyecto con el siguiente nombre: **Ejercicio2-5.mpj**

Ejercicio 2.6. El número de palabras por línea de una página de un libro viene dado por:

Nº de palabras por línea	4	5	8	9	10	11	12	13	14	15	16	17
Nº de líneas	1	1	2	3	2	7	11	14	3	2	1	1

a) Crea un nuevo proyecto de ***Minitab***.

b) Guarda los datos en el archivo **PalabrasPorLinea.mtw**

c) Haz un polígono de frecuencias absolutas. Modifícalo a tu gusto.

d) Graba el proyecto con el siguiente nombre: **Ejercicio2-6.mpj**

3

Probabilidad

3.1. Desarrollo de los contenidos fundamentales

3.1.1. Introducción a la Probabilidad

- *Experimento*: cualquier proceso que permite asociar a cada individuo de una población un símbolo (numérico o no) entre los símbolos de un conjunto dado a priori.
  - ⋆ *Experimento determinista*: es aquel en el que los resultados están totalmente determinados una vez que se fijan las condiciones en las que se realiza el experimento.
  - ⋆ *Experimento aleatorio*: está caracterizado por las tres propiedades siguientes:
    - ○ Todos sus posibles resultados son conocidos con anterioridad.
    - ○ No se puede predecir el resultado del experimento.
    - ○ El experimento puede repetirse en condiciones idénticas.
- *Ensayo* o *prueba*: es la realización concreta de un experimento aleatorio.
- *Dato, observación* o *resultado*: es el símbolo que se ha obtenido en un ensayo de un experimento aleatorio.
- *Suceso elemental*: cada resultado de un experimento aleatorio.
- *Espacio muestral* (Ω): conjunto de todos los sucesos elementales.
- *Suceso* (A, B, ...): conjunto de sucesos elementales.
- *Suceso seguro*: es el espacio muestral.
- *Suceso imposible* ($\emptyset$): no consta de ningún suceso elemental.

3.1.2. Operaciones con sucesos

- *Suceso contrario*: Dado un suceso A, se denomina **suceso contrario** de A al suceso $\overline{A}$ que ocurre cuando no ocurre A; es decir, $\overline{A}$ consta de los sucesos elementales de Ω que no están incluidos en A.
- *Unión de sucesos*: Dados dos sucesos A y B de un mismo experimento, se entiende por **unión** de ambos, y se denota por $A \cup B$, al suceso que ocurre cuando ocurre A, cuando ocurre B o cuando ocurren ambos; es decir, al formado por todos los sucesos elementales que son de A o de B.
- *Intersección de sucesos*: Dados dos sucesos A y B de un mismo experimento, se entiende por **intersección** de ambos, y se representa por $A \cap B$, al suceso que ocurre cuando ocurren A y B a la vez; es decir, al formado por todos los sucesos elementales que pertenecen a A y a B simultáneamente.
- *Sucesos incompatibles*: A y B son dos sucesos incompatibles si no tienen ningún suceso elemental en común ($A \cap B = \emptyset$).
- *Diferencia de sucesos*: Dados dos sucesos A y B de un mismo experimento aleatorio, se entiende por **diferencia** de ambos, y se denota por $A - B$, al suceso que ocurre cuando ocurre A pero no B; es decir, al que consta de los sucesos elementales de A que no están en B.

3.1.3. Regla de Laplace

Si un experimento aleatorio da lugar a un número finito de sucesos elementales, todos ellos igualmente posibles (es decir, no se conoce razón alguna que favorezca la presentación de uno respecto de los otros), entonces la **probabilidad de un suceso** A es:

$$P(A) = \frac{\text{nº de casos favorables al suceso A}}{\text{nº de casos posibles del experimento}}.$$

3.1.4. Propiedades de la probabilidad

- **Propiedad fundamental de la probabilidad**: La probabilidad de un suceso es un número comprendido entre 0 y 1; es decir:

$$0 \leq P(A) \leq 1\,, \quad \text{para todo suceso A}\,.$$

- **Probabilidad del suceso seguro**: La probabilidad del espacio muestral es 1; es decir:

$$P(\Omega) = 1.$$

- **Probabilidad del suceso contrario**: La probabilidad del suceso contrario de A es:

$$P(\overline{A}) = 1 - P(A)\,.$$

- **Probabilidad del suceso imposible**: La probabilidad del suceso imposible es cero; es decir:

$$P(\emptyset) = 0\,.$$

- **Probabilidad de la diferencia de sucesos**: Si B está incluido en A entonces:

$$P(A - B) = P(A) - P(B)\,.$$

- **Probabilidad de la unión de dos sucesos incompatibles**: Si A y B son dos sucesos incompatibles entonces la probabilidad del suceso unión es la suma de las probabilidades de A y B; es decir:

$$P(A \cup B) = P(A) + P(B)\,, \text{ si } A \text{ y } B \text{ son incompatibles.}$$

- **Probabilidad de la unión de n sucesos incompatibles**: Si varios sucesos son incompatibles dos a dos, la probabilidad de la unión de todos ellos es la suma de sus probabilidades; es decir:

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)\,,$$
$$\text{si } A_1, A_2, \ldots, A_n \text{ son incompatibles dos a dos.}$$

- **Probabilidad de la unión de dos sucesos cualesquiera**: La probabilidad de la unión de dos sucesos cualesquiera es igual a la probabilidad del primero, más la probabilidad del segundo, menos la probabilidad de la intersección; es decir:

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

- **Probabilidad de la unión de tres sucesos cualesquiera**: Si A, B y C son tres sucesos cualesquiera entonces la probabilidad de la unión de los tres sucesos es:

$$\begin{aligned} P(A \cup B \cup C) \;=\; & P(A) + P(B) + P(C) \\ & - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ & + P(A \cap B \cap C)\,. \end{aligned}$$

3.2. Ejemplos que se van a resolver en clase

Ejemplo 3.1. Dar un ejemplo de experimento aleatorio. Determinar el espacio muestral. Poner dos ejemplos de sucesos (A y B).

Ejemplo 3.2. Determinar los sucesos contrarios de los del Ejemplo 3.1 ($\overline{A}$ y $\overline{B}$).

Ejemplo 3.3. Con los sucesos A y B del Ejemplo 3.1 determinar las siguientes uniones de sucesos: $A \cup B$ y $\overline{A} \cup \overline{B}$.

Ejemplo 3.4. Con los sucesos A y B del Ejemplo 3.1 determinar las siguientes intersecciones de sucesos: $A \cap B$ y $\overline{A} \cap \overline{B}$.

Ejemplo 3.5. ¿Son incompatibles los sucesos A y B del Ejemplo 3.1?

Ejemplo 3.6. Con los sucesos A y B del Ejemplo 3.1 determinar las siguientes diferencias de sucesos: $A - B$ y $B - A$.

Ejemplo 3.7. En una biblioteca que consta de 250 libros, 20 de ellos están escritos en inglés y el resto en español. ¿Cuál es la probabilidad de que un libro elegido al azar, entre los 250 de dicha biblioteca, esté escrito en inglés?

Ejemplo 3.8. Estamos investigando la calidad de las fotocopias hechas en una biblioteca. En una muestra de 100 copias, se observa que 2 están en blanco y manchadas, 3 están en blanco pero no están manchadas y 25 no están en blanco pero están manchadas. ¿Cuál es la probabilidad de que esta máquina fotocopiadora realice una copia que no esté en blanco ni manchada?

Ejemplo 3.9. Una biblioteca dispone de tres empleados (A, B y C) para atender a los usuarios. El 20 % de las ocasiones está disponible (para atender a cualquier usuario) el empleado A, el 30 % de las veces está disponible el empleado B y el 25 % de las ocasiones está disponible el empleado C. Además, el 10 % de las veces están disponibles A y B, el 12 % están disponibles A y C, el 14 % están disponibles B y C, y el 8 % de las ocasiones están disponibles los tres empleados. ¿Cuál es la probabilidad de que una persona sea atendida en el mismo momento en que llegue a la biblioteca?

Ejemplo 3.10. En un grupo de alumnos de una licenciatura en documentación, el 25 % suspendió la asignatura Análisis Documental, el 15 % la asignatura Documentación General y el 10 % ambas asignaturas. ¿Cuál es la probabilidad de que un alumno suspenda Análisis Documental o Documentación General?

Ejemplo 3.11. En un estudio realizado en un determinado país sobre la participación de la mujer en trabajos sobre información y documentación, antes y después de ser madre, se selecciona una muestra de 683 mujeres obteniéndose los siguientes resultados:

		Después	
		NO	SÍ
Antes	NO	169	3
	SÍ	337	174

a) Calcular la probabilidad de que una mujer participe en dicho mercado laboral antes de ser madre.

b) Calcular la probabilidad de que una mujer participe en dicho mercado laboral después de ser madre.

c) Calcular la probabilidad de que una mujer participe en dicho mercado laboral antes y después de ser madre.

d) Calcular la probabilidad de que una mujer participe en dicho mercado laboral antes o después de ser madre.

3.3. Actividades de aplicación de los contenidos

3.3.1. Problemas propuestos

Problema 3.1. Un centro de información dispone de 10 ordenadores para consultar diversas bases de datos. Se realiza el experimento que consiste en observar, en diferentes instantes del día, el número de ordenadores que no están ocupados. Determinar el espacio muestral. Poner dos ejemplos de sucesos (A y B). Hallar los sucesos contrarios ($\overline{A}$ y $\overline{B}$), el suceso unión ($A \cup B$), el suceso intersección ($A \cap B$), el suceso diferencia ($A - B$), y los sucesos $\overline{A \cup B}$, $\overline{A \cap B}$ y $\overline{A - B}$.

Problema 3.2. El número de libros por estante de una biblioteca viene dado por:

Nº de libros	19	20	21	22	23	24	25	26	27	28	29	30
Nº de estantes	2	3	7	5	14	11	12	9	6	6	3	2

Calcular la probabilidad de que un estante elegido al azar tenga:

a) exactamente 24 libros.

b) 24 o 25 libros.

c) menos de 24 libros.

Problema 3.3. Los asistentes a un acto cultural preparado por una biblioteca se clasifican de la siguiente manera:

	menos de 18 años	entre 18 y 24 años	entre 25 y 40 años	más de 40 años
Hombre	17	28	31	52
Mujer	23	39	50	75

a) Calcular la probabilidad de que un asistente al acto, elegido al azar, tenga más de 40 años.

b) Calcular la probabilidad de que un asistente al acto, elegido al azar, sea mujer y tenga más de 40 años.

c) Calcular la probabilidad de que una mujer asistente al acto, elegida al azar, tenga más de 40 años.

Problema 3.4. Se pregunta a todos los alumnos de una determinada facultad cuántas horas dedican al estudio en la biblioteca, y los resultados son:

		Curso de la licenciatura				
		1º	2º	3º	4º	5º
Nº de horas	menos de 1 hora	18	20	32	77	96
	entre 1 y 3 horas	22	35	90	83	50
	más de 3 horas	60	70	80	60	14

a) Determinar la probabilidad de que un alumno, elegido al azar, estudie más de 3 horas diarias en la biblioteca.

b) Hallar la probabilidad de que un alumno de quinto curso, elegido al azar, estudie más de 3 horas diarias en la biblioteca.

c) Calcular la probabilidad de que un alumno, elegido al azar, sea de quinto curso o estudie más de 3 horas diarias en la biblioteca.

Problema 3.5. En la siguiente tabla aparece el número de hombres y de mujeres que se han llevado prestados libros y vídeos de una biblioteca pública.

		Tipo de documento		suma
		libro	vídeo	
Sexo	hombre	195	215	410
	mujer	315	205	520
suma		510	420	930

a) Calcular la probabilidad de que un usuario de la biblioteca, elegido al azar, sea mujer.

b) Calcular la probabilidad de que un usuario de la biblioteca, elegido al azar, se lleve prestado un vídeo.

c) Calcular la probabilidad de que un usuario de la biblioteca, elegido al azar, sea mujer y se lleve prestado un vídeo.

d) Calcular la probabilidad de que un usuario de la biblioteca, elegido al azar, sea mujer o se lleve prestado un vídeo.

Problema 3.6. El porcentaje de usuarios de la biblioteca G que trabajan en Murcia es del 55 %, y el porcentaje de usuarios de dicha biblioteca que trabajan en Murcia y han nacido en Murcia es del 35 %. Elegido un usuario de dicha biblioteca al azar, ¿cuál es la probabilidad de que trabaje en Murcia pero no haya nacido en Murcia?

Problema 3.7. El 75 % de los estudiantes de la Universidad de Murcia son murcianos, el 15 % de los estudiantes de la Universidad de Murcia tienen algún hijo y el 10 % de los estudiantes de la Universidad de Murcia son murcianos y tienen algún hijo.

a) Si elegimos un estudiante de la Universidad de Murcia al azar ¿cuál es la probabilidad de que sea murciano y no tenga ningún hijo?

b) Si elegimos un estudiante de la Universidad de Murcia al azar ¿cuál es la probabilidad de que sea murciano o tenga algún hijo?

Problema 3.8. Se ha estudiado el uso de la biblioteca pública por parte de los profesores universitarios, encontrándose que 42 de 113 psicólogos, 17 de 68 biólogos, 33 de 203 ingenieros y 20 de 78 profesores de inglés son usuarios de la biblioteca pública (y el resto no).

a) Elegido un profesor universitario al azar, ¿cuál es la probabilidad de que sea profesor de inglés?

b) Elegido un profesor universitario al azar, ¿cuál es la probabilidad de que sea usuario de la biblioteca pública?

c) Elegido un profesor universitario al azar, ¿cuál es la probabilidad de que sea usuario de la biblioteca pública y profesor de inglés?

d) Elegido un profesor universitario al azar, ¿cuál es la probabilidad de que sea usuario de la biblioteca pública o profesor de inglés?

3.3.2. Soluciones de los problemas propuestos

Solución del problema 3.1.

El espacio muestral es $= \Omega = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Los sucesos A y B podrían ser:

$A = \{$el número de ordenadores no ocupados es menor que $4\} = \{0, 1, 2, 3\}$

$B = \{$el número de ordenadores no ocupados está comprendido entre 2 y $6\} = \{2, 3, 4, 5, 6\}$

Por tanto:

$\overline{A} = \{4, 5, 6, 7, 8, 9, 10\}$

$\overline{B} = \{0, 1, 7, 8, 9, 10\}$

$A \cup B = \{0, 1, 2, 3, 4, 5, 6\}$

$A \cap B = \{2, 3\}$

$A - B = \{0, 1\}$

$\overline{A \cup B} = \{7, 8, 9, 10\} = \overline{A} \cap \overline{B} \neq \overline{A} \cup \overline{B}$

$\overline{A \cap B} = \{0, 1, 4, 5, 6, 7, 8, 9, 10\} = \overline{A} \cup \overline{B} \neq \overline{A} \cap \overline{B}$

$\overline{A - B} = \{2, 3, 4, 5, 6, 7, 8, 9.10\} \neq \overline{A} - \overline{B}$

Solución del problema 3.2. **a)** $0'1375$, **b)** $0'2875$, **c)** $0'3875$.

Solución del problema 3.3. **a)** $0'403174603$, **b)** $0'238095238$, **c)** $0'401069518$.

Solución del problema 3.4. **a)** $0'351920693$, **b)** $0'0875$, **c)** $0'53283767$.

Solución del problema 3.5. **a)** $0'559140$, **b)** $0'451613$, **c)** $0'220430$, **d)** $0'790323$.

Solución del problema 3.6. $0'2$

Solución del problema 3.7. **a)** $0'65$, **b)** $0'8$.

Solución del problema 3.8. **a)** $0'168831$, **b)** $0'\widehat{24}$, **c)** $0'043290$, **d)** $0'367965$.

4

Modelos de probabilidad

4.1. Desarrollo de los contenidos fundamentales

4.1.1. Variables aleatorias discretas y continuas

4.1.1.1. Variables aleatorias

Una **variable aleatoria** es una función que asigna un número a cada suceso elemental de un experimento aleatorio.

Cualquier variable estadística cuantitativa estudiada en los temas 1 y 2 podría considerarse variable aleatoria con la condición de que esté observada en todos los individuos de una población.

La **media** de una variable aleatoria X se denota por μ_x. En el caso en el que no exista la posibilidad de confusión respecto de la variable aleatoria con la que estamos trabajando, la media se denotará solamente por μ. A la media de una variable aleatoria X también se le llama **esperanza matemática** de X, denotándola entonces por $E(X)$.

La **varianza** de una variable aleatoria X se denota por $\mathrm{Var}(X)$, por σ_x^2 o simplemente por σ^2.

Por tanto, la **desviación típica** de una variable aleatoria X se denota por σ_x o por σ.

La **función de distribución** de una variable aleatoria X se denota por F_X o simplemente por F y se define de la siguiente forma:

$$F(t) = P(X \leq t) \quad \text{para todo } t\,.$$

CLASIFICACIÓN DE LAS VARIABLES ALEATORIAS:

- ⋆ **Variable aleatoria discreta**: sólo puede tomar valores numéricos aislados (fijados dos consecutivos, no puede existir ninguno intermedio).
- ⋆ **Variable aleatoria continua**: puede tomar cualquier valor numérico dentro de un intervalo, de modo que entre cualesquiera dos de ellos siempre existe otro posible valor.

4.1.1.2. Variables aleatorias continuas

- Una variable aleatoria X queda totalmente identificada si conocemos su **función de densidad**, $f(x)$, que debe verificar:

  (1) $f(x) \geq 0$ para todo número real x.

  (2) El área total bajo la curva $y = f(x)$ vale 1.

  (3) La probabilidad de que la variable aleatoria X esté comprendida entre a y b, $P(a \leq X \leq b)$, viene determinada por el área bajo la curva $y = f(x)$ entre $x = a$ y $x = b$ (véase la Figura 10 (a)).

(a) área sombreada $= P(a < X < b)$ (b) área sombreada $= F(t)$

Figura 10: Función de densidad y función de distribución de una variable aleatoria continua

- Los valores concretos de la función de densidad no tienen ningún significado especial pues las probabilidades vienen determinadas por áreas bajo la curva determinada por la función de densidad y no por valores de la función de densidad. En todo caso, este hecho nos informa de que en las distribuciones continuas la probabilidad de que la variable aleatoria tome un valor concreto, $P(X = a)$, es cero, como corresponde al área de un rectángulo de base un punto y altura $f(a)$. Resumiendo, si X es una variable aleatoria continua, entonces:

$$P(X = a) = 0 \quad \text{para todo } a\,.$$

- La **representación gráfica de la función de densidad** de una variable aleatoria continua es equivalente al polígono de frecuencias relativas de una variable estadística continua cuando la amplitud de los intervalos es infinitesimal.

- La **media** y la **varianza** de una variable aleatoria continua se determinan mediante una operación matemática denominada *integral*.

- La **función de distribución** de una variable aleatoria continua X se define igual que para cualquier variable aleatoria; es decir:

$$F(t) = P(X \leq t) \quad \text{para todo } t\,.$$

  La interpretación gráfica de la anterior definición es la siguiente: el resultado de $F(t)$ coincide con el área bajo la curva $y = f(x)$ desde el valor más pequeño que puede tomar la variable hasta el valor t (véase la Figura 10 (b)).

- Para todas las variables aleatorias continuas importantes los resultados de la función de distribución se pueden determinar con cualquier paquete estadístico, como ***Minitab***.

- Si X es una variable aleatoria continua, entonces se cumple:

  Prop. 1: $P(X < a) = P(X \leq a) = F(a)$ para todo a.

  Prop. 2: $P(X \geq a) = P(X > a) = 1 - F(a)$ para todo a.

  Prop. 3: $P(a < X < b) = P(a \leq X \leq b) = P(a \leq X < b) = P(a < X \leq b) = F(b) - F(a)$ para todo a y b.

  Una consecuencia fundamental de estas tres propiedades es la siguiente: conociendo los resultados de la función de distribución podemos determinar los resultados de cualquier probabilidad.

- Si X es una variable aleatoria continua, el **percentil** al $100p\,\%$ es el valor x_p que verifica:

$$P(X \leq x_p) = p\,.$$

  De esta definición se deduce que los percentiles son los valores inversos de los resultados de la función de distribución (el resultado de la función de distribución es p y el valor del percentil al $100p\,\%$ es x_p).

4.1.2. La distribución Normal

4.1.2.1. Función de densidad

Una variable aleatoria continua X tiene una distribución **Normal de parámetros μ y σ** si su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\ \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \qquad \text{para todo } x\,,$$

donde μ es cualquier número, σ es cualquier número positivo y, en general, $\exp(t)$ significa e^t, siendo e la base de los logaritmos neperianos.

La representación gráfica de dicha función es la curva de la Figura 11.

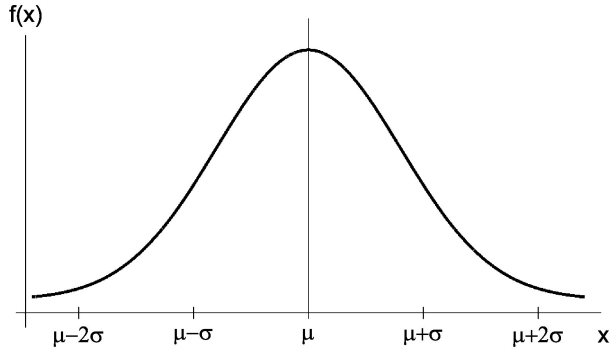


Figura 11: Función de densidad de una variable Normal de parámetros μ y σ

Son equivalentes los dos enunciados siguientes: "X tiene una distribución Normal de parámetros μ y σ" y "X es una variable aleatoria Normal de parámetros μ y σ".

La variable aleatoria Normal de parámetros μ y σ será denotada por:

$$\mathcal{N}(\mu, \sigma)\,.$$

Se cumplen las siguientes propiedades:

- La media, la mediana y la moda de una variable aleatoria $\mathcal{N}(\mu, \sigma)$ coinciden entre sí y tienen por valor al parámetro μ.
- La desviación típica de la distribución $\mathcal{N}(\mu, \sigma)$ es igual al parámetro σ.
- La curva que representa a la función de densidad de la distribución $\mathcal{N}(\mu, \sigma)$ es simétrica respecto de la recta vertical de ecuación $x = \mu$.
- El área comprendida entre el eje horizontal y la curva que representa a la función de densidad de la distribución $\mathcal{N}(\mu, \sigma)$ vale 1 (como ocurre con cualquier distribución continua).

A la variable aleatoria Normal de parámetros 0 y 1 se le llama variable aleatoria **Normal Estándar**, o **Normal Típica**, y se le denota por $\mathcal{N}(0, 1)$.

4.1.2.2. Función de distribución

Existen tablas que contienen los resultados de la función de distribución de una variable aleatoria Normal Estándar, $F(t)$, para algunos valores de t; pero nosotros vamos a utilizar el paquete estadístico ***Minitab***, que nos determina los resultados de la función de distribución, $F(t)$, de una variable aleatoria Normal de parámetros μ y σ, para cualquier valor de μ, σ y t.

4.1.2.3. Percentiles

Si X es una variable aleatoria Normal de parámetros μ y σ, entonces, en particular, es una variable aleatoria continua; por tanto, el **percentil** al $100p\,\%$ es el valor x_p que verifica:

$$P(X \leq x_p) = p\,.$$

Si Z es una variable aleatoria Normal Estándar, el **percentil** al $100p\,\%$ de Z se denota por Z_p y es el valor que verifica:

$$P(Z \leq Z_p) = p\,,$$

es decir, el área comprendida entre la curva de densidad y el eje horizontal, a la izquierda de Z_p, es igual a p (véase la Figura 12).

Otra interpretación es la siguiente: el valor Z_p deja por debajo el $100p\,\%$ de todos los resultados de una variable aleatoria Normal Estándar.

Existen tablas que contienen los resultados de los percentiles de una variable aleatoria Normal Estándar, Z_p, para algunos valores de p; pero nosotros vamos a utilizar el paquete estadístico ***Minitab***, que nos determina los resultados de los percentiles, x_p, de una variable aleatoria Normal de parámetros μ y σ, para cualquier valor de μ, σ y p.

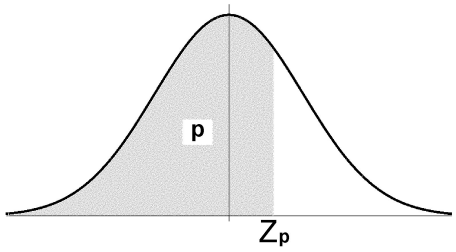


Figura 12: Percentil al $100p$ % de una distribución Normal Estándar

4.1.3. Otras distribuciones continuas importantes

4.1.3.1. Distribución chi-cuadrado de Pearson

Si $Z_1, Z_2, \ldots, Z_n$ son variables aleatorias independientes, todas ellas con distribución Normal Estándar, entonces la variable aleatoria $Z_1^2 + Z_2^2 + \cdots + Z_n^2$ sigue una distribución denominada **chi-cuadrado de Pearson con n grados de libertad**, que se denota por χ_n^2.

Si X es una variable aleatoria χ_n^2, el percentil al $100p\,\%$ de X se representa por $\chi^2_{n\,,p}$ y es el valor que verifica:

$$P(X \leq \chi^2_{n\,,p}) = p\,,$$

es decir, el área comprendida entre la curva de densidad y el eje horizontal, a la izquierda de $\chi^2_{n\,,p}$, es igual a p (véase la Figura 13).

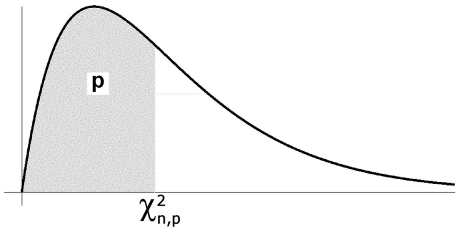


Figura 13: Percentil al $100p$ % de una distribución χ_n^2

Otra interpretación es la siguiente: el valor $\chi^2_{n\,,p}$ deja por debajo el $100p\,\%$ de todos los resultados de una variable aleatoria chi-cuadrado de Pearson con n grados de libertad.

Con Minitab podemos determinar los resultados de la función de densidad, de la función de distribución y de los percentiles de la variable aleatoria chi-cuadrado de Pearson con n grados de libertad, para todo valor de n.

4.1.3.2. Distribución t de Student

Si Z sigue una distribución Normal Estándar y χ_n^2 es independiente de Z, entonces la variable aleatoria

$$\frac{Z}{\sqrt{\dfrac{\chi_n^2}{n}}}$$

sigue una distribución denominada t **de Student con n grados de libertad**, que se denota por t_n.

Si X es una variable aleatoria t_n, el percentil al $100p\,\%$ de X se representa por $t_{n\,,p}$ y es el valor que verifica:

$$P(X \leq t_{n\,,p}) = p\,,$$

es decir, el área comprendida entre la curva de densidad y el eje horizontal, a la izquierda de $t_{n\,,p}$, es igual a p (véase la Figura 14).

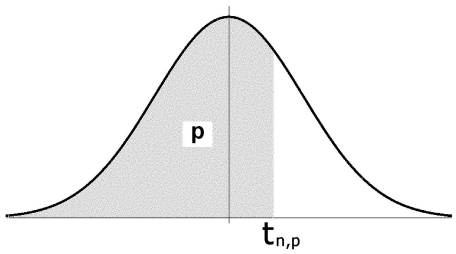


Figura 14: Percentil al $100p\,\%$ de una distribución t_n

Otra interpretación es la siguiente: el valor $t_{n\,,p}$ deja por debajo el $100p\,\%$ de todos los resultados de una variable aleatoria t de Student con n grados de libertad.

Con Minitab podemos determinar los resultados de la función de densidad, de la función de distribución y de los percentiles de la variable aleatoria t de Student con n grados de libertad, para todo valor de n.

4.1.3.3. Distribución F de Snedecor

Si tenemos dos variables aleatorias chi-cuadrado independientes, χ^2_m y χ^2_n, entonces la variable aleatoria

$$\frac{\dfrac{\chi^2_m}{m}}{\dfrac{\chi^2_n}{n}}$$

sigue una distribución denominada F **de Snedecor con** m **grados de libertad en el numerador y** n **grados de libertad en el denominador**, que se denota por $F_{m\,,n}$.

Si X es una variable aleatoria $F_{m\,,n}$, el percentil al $100p\,\%$ de X se representa por $F_{m\,,n\,,p}$ y es el valor que verifica:

$$P(X \leq F_{m\,,n\,,p}) = p\,,$$

es decir, el área comprendida entre la curva de densidad y el eje horizontal, a la izquierda de $F_{m\,,n\,,p}$, es igual a p (véase la Figura 15).

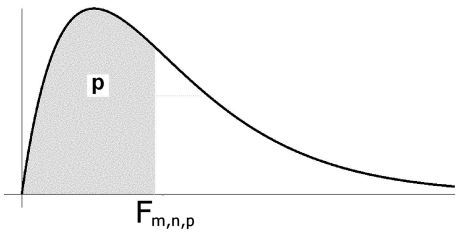


Figura 15: Percentil al $100p\,\%$ de una distribución $F_{m\,,n}$

Otra interpretación es la siguiente: el valor $F_{m,n,p}$ deja por debajo el $100p\,\%$ de todos los resultados de una variable aleatoria F de Snedecor con m grados de libertad en el numerador y n grados de libertad en el denominador.

Con Minitab podemos determinar los resultados de la función de densidad, de la función de distribución y de los percentiles de la variable aleatoria F de Snedecor con m grados de libertad en el numerador y n grados de libertad en el denominador, para cualquier valor de m y n.

4.2. Ejemplos que se van a resolver en clase

Ejemplo 4.1. Utilizando los contenidos del apartado 4.4.1, generar una muestra aleatoria de 10000 datos procedentes de una variable aleatoria $\mathcal{N}(7'65, 2'34)$. Realizar el histograma de la muestra aleatoria obtenida. Determinar el mínimo valor, el máximo valor, la media y la cuasi-desviación típica de dicha muestra.

Ejemplo 4.2. Utilizando los contenidos del apartado 4.4.2 y los resultados del ejemplo anterior, hacer la representación gráfica de la función de densidad de una variable aleatoria $\mathcal{N}(7'65, 2'34)$.

Ejemplo 4.3. Utilizando los contenidos del apartado 4.4.1, generar una muestra aleatoria de 12000 datos procedentes de una variable aleatoria $\mathcal{N}(0, 1)$. Realizar el histograma de la muestra aleatoria obtenida. Determinar el mínimo valor, el máximo valor, la media y la cuasi-desviación típica de dicha muestra.

Ejemplo 4.4. Si $Z \equiv \mathcal{N}(0, 1)$ calcular las siguientes probabilidades:

a) $P(Z < 0'321)$.

b) $P(Z \geq 1'275)$.

c) $P(Z < -2'152)$.

d) $P(Z \geq -0'456)$.

e) $P(-1'434 \leq Z \leq 1'568)$.

Ejemplo 4.5. En una determinada asignatura de un Grado en Información y Documentación se sabe que las calificaciones siguen una distribución Normal de media $5'5$ y desviación típica $1'5$. Si en un año académico hay 150 alumnos matriculados en esta asignatura, calcular el número de alumnos que obtendrán una calificación:

a) menor o igual que 3.

b) mayor o igual que 8.

c) comprendida entre 4 y 6.

Ejemplo 4.6. Si $X \equiv \mathcal{N}(7'65, 2'34)$, determinar el valor de k tal que:

a) $P(X \leq k) = 0'95$.

b) $P(X > k) = 0'01$.

Ejemplo 4.7. Si $Z \equiv \mathcal{N}(0, 1)$ determinar los siguientes percentiles e interpretar los resultados.

a) Mediana de Z.

b) Tercer cuartil de Z.

c) Primer cuartil de Z.

Ejemplo 4.8. Si $Z \equiv \mathcal{N}(0,1)$, calcular el valor de t para que se verifique:

a) $P(Z < t) = 0'05$.

b) $P(Z \geq t) = 0'99$.

Ejemplo 4.9. Utilizando los contenidos del apartado 4.4.1, generar una muestra aleatoria de 15000 datos procedentes de una variable aleatoria χ^2_{40}. Realizar el histograma de la muestra aleatoria obtenida. Determinar el mínimo valor, el máximo valor y la media de dicha muestra.

Ejemplo 4.10. Utilizando los contenidos del apartado 4.4.2 y los resultados del ejemplo anterior, hacer la representación gráfica de la función de densidad de una variable aleatoria χ^2_{40}.

Ejemplo 4.11. Si $X \equiv \chi^2_{40}$ calcular las siguientes probabilidades:

a) $P(X < 39)$.

b) $P(X \geq 33)$.

c) $P(40 \leq X \leq 45)$.

Ejemplo 4.12. Determinar los siguientes percentiles e interpretar los resultados.

a) Mediana de χ^2_{40}.

b) Tercer cuartil de χ^2_{30}.

Ejemplo 4.13. Si $X \equiv \chi^2_{40}$, calcular el valor de a para que se verifique:

a) $P(X < a) = 0'8$.

b) $P(X \geq a) = 0'8$.

Ejemplo 4.14. Utilizando los contenidos del apartado 4.4.1, generar una muestra aleatoria de 14000 datos procedentes de una variable aleatoria t_{25}. Realizar el histograma de la muestra aleatoria obtenida. Determinar el mínimo valor, el máximo valor y la media de dicha muestra.

Ejemplo 4.15. Utilizando los contenidos del apartado 4.4.2 y los resultados del ejemplo anterior, hacer la representación gráfica de la función de densidad de una variable aleatoria t_{25}.

Ejemplo 4.16. Si $X \equiv t_{25}$ calcular las siguientes probabilidades:

a) $P(X < -2)$.

b) $P(X \geq 3)$.

c) $P(-1 < X \leq 1)$.

Ejemplo 4.17. Determinar los siguientes percentiles e interpretar los resultados.

a) Tercer cuartil de t_{25}.

b) Primer cuartil de t_{60}.

Ejemplo 4.18. Si $X \equiv t_{25}$, calcular el valor de b para que se verifique:

a) $P(X < b) = 0'35$.

b) $P(X \geq b) = 0'85$.

Ejemplo 4.19. Utilizando los contenidos del apartado 4.4.1, generar una muestra aleatoria de 20000 datos procedentes de una variable aleatoria $F_{20\,,10}$. Realizar el histograma de la muestra aleatoria obtenida. Determinar el mínimo valor, el máximo valor y la media de dicha muestra.

Ejemplo 4.20. Utilizando los contenidos del apartado 4.4.2 y los resultados del ejemplo anterior, hacer la representación gráfica de la función de densidad de una variable aleatoria $F_{20\,,10}$.

Ejemplo 4.21. Si $X \equiv F_{20\,,10}$ calcular las siguientes probabilidades:

a) $P(X < 0'72)$.

b) $P(X \geq 1'05)$.

c) $P(0'7 \leq X < 1'5)$.

Ejemplo 4.22. Determinar los siguientes percentiles e interpretar los resultados.

a) Percentil al 95 % de $F_{20\,,10}$.

b) Percentil al 10 % de $F_{20\,,10}$.

Ejemplo 4.23. Si $X \equiv F_{20\,,10}$, calcular el valor de c para que se verifique:

a) $P(X < c) = 0'995$.

b) $P(X \geq c) = 0'025$.

4.3. Actividades de aplicación de los contenidos

4.3.1. Problemas propuestos

Problema 4.1. Si Z es una variable Normal Estándar, determinar:

a) $P(Z \leq 2'21)$.

b) $P(Z < 3'47)$.

c) $P(Z \leq -1'75)$.

d) $P(Z > 2'46)$.

e) $P(Z \geq 3'24)$.

f) $P(Z > -3'08)$.

g) $P(1'12 \leq Z \leq 2'68)$.

h) $P(-0'85 < Z < 1'27)$.

i) $P(-2'97 < Z \leq -1'33)$.

Problema 4.2. Si X es una variable Normal con media $8'46$ y desviación típica $1'14$, hallar:

a) $P(X \leq 9'11)$.

b) $P(X < 12'33)$.

c) $P(X \leq 6'41)$.

d) $P(X > 10'52)$.

e) $P(X \geq 12'61)$.

f) $P(X > 4'01)$.

g) $P(6'11 \leq X \leq 11'91)$.

h) $P(7'53 < X < 10'33)$.

i) $P(5'05 \leq X < 6'83)$.

Problema 4.3. Si Z denota la variable aleatoria Normal Estándar, calcular el valor de a para que se verifique:

a) $P(Z \leq a) = 0'722405$.

b) $P(Z < a) = 0'344578$.

c) $P(Z > a) = 0'284339$.

d) $P(Z \geq a) = 0'978822$.

Problema 4.4. Si X es una variable aleatoria con distribución Normal de media $3'5$ y desviación típica $0'8$, determinar el valor de a tal que:

a) $P(X \leq a) = 0'773373$.

b) $P(X < a) = 0'012224$.

c) $P(X > a) = 0'066807$.

d) $P(X \geq a) = 0'99865$.

Problema 4.5. Hallar el valor de los siguientes percentiles:

a) $Z_{0'58}$.

b) $Z_{0'42}$.

c) $Z_{0'999}$.

d) $Z_{0'001}$.

Problema 4.6. El cociente intelectual de 5.600 alumnos de la licenciatura en documentación de diversas universidades sigue una distribución Normal de media 130 y desviación típica 6. Calcular cuántos de ellos tienen un cociente intelectual:

a) mayor que 140.

b) entre 125 y 135.

c) menor que 120.

Problema 4.7. Calcular el valor de los siguientes percentiles:

a) $\chi^2_{6\,,\,0'01}$.

b) $\chi^2_{6\,,\,0'99}$.

c) $\chi^2_{72\,,\,0'975}$.

Problema 4.8. Sea X una variable aleatoria que sigue una distribución chi-cuadrado de Pearson con 10 grados de libertad. Calcular:

a) $P(X \leq 4'86518)$.

b) $P(X > 12'5489)$.

c) $P(9'34182 < X < 18'307)$.

Problema 4.9. Sea X una variable aleatoria que sigue una distribución chi-cuadrado de Pearson con 15 grados de libertad. Determinar el valor de a que verifica la siguiente igualdad:

a) $P(X \leq a) = 0'05$.

b) $P(X > a) = 0'99$.

c) $P(-a < X < a) = 0'25$.

Problema 4.10. Calcular el valor de los siguientes percentiles:

a) $t_{26\,,\,0'9}$.

b) $t_{26\,,\,0'1}$.

c) $t_{75\,,\,0'8}$.

Problema 4.11. Sea X una variable aleatoria que sigue una distribución t de Student con 7 grados de libertad. Calcular:

a) $P(X \leq 1'8946)$.

b) $P(X \geq 2'998)$.

c) $P(0'7111 \leq X \leq 3'4995)$.

Problema 4.12. Sea X una variable aleatoria que sigue una distribución t de Student con 20 grados de libertad. Determinar el valor de a que verifica la siguiente igualdad:

a) $P(X \leq a) = 0'99$.

b) $P(X \geq a) = 0'25$.

c) $P(-a < X < a) = 0'9$.

Problema 4.13. Calcular el valor de los siguientes percentiles:

a) $F_{8\,,\,6\,,\,0'975}$.

b) $F_{25\,,\,50\,,\,0'01}$.

c) $F_{45\,,\,35\,,\,0'01}$.

Problema 4.14. Sea X una variable aleatoria que sigue una distribución F de Snedecor con 12 grados de libertad en el numerador y 20 grados de libertad en el denominador. Calcular:

a) $P(X < 1'8924)$.

b) $P(X > 2'6758)$.

c) $P(2'2776 < X < 3'2311)$.

Problema 4.15. Sea X una variable aleatoria que sigue una distribución F de Snedecor con 10 grados de libertad en el numerador y 8 grados de libertad en el denominador. Determinar el valor de a que verifica la siguiente igualdad:

a) $P(X < a) = 0'9$.

b) $P(X > a) = 0'05$.

c) $P(-a < X < a) = 0'95$.

4.3.2. Soluciones de los problemas propuestos

Solución del problema 4.1. **a)** $0'986447$, **b)** $0'9997398$, **c)** $0'040059$, **d)** $0'006947$, **e)** $0'0005976$, **f)** $0'998965$, **g)** $0'127676$, **h)** $0'700295$, **i)** $0'09027$.

Solución del problema 4.2. **a)** $0'715661$, **b)** $0'9996505$, **c)** $0'03593$, **d)** $0'035148$, **e)** $0'0001363$, **f)** $0'9999519$, **g)** $0'979078$, **h)** $0'743389$, **i)** $0'074964$.

Solución del problema 4.3. **a)** $0'59$, **b)** $-0'4$, **c)** $0'57$, **d)** $-2'03$.

Solución del problema 4.4. **a)** $4'1$, **b)** $1'7$, **c)** $4'7$, **d)** $1'1$.

Solución del problema 4.5. **a)** $0'20189$, **b)** $-0'20189$, **c)** $3'09023231$, **d)** $-3'09023231$.

Solución del problema 4.6. **a)** $0'04746 \cdot 5600 = 265'776 \simeq 266$ alumnos, **b)** $0'593462 \cdot 5600 = 3323'3872 \simeq 3323$ alumnos, **c)** $0'04746 \cdot 5600 = 265'776 \simeq 266$ alumnos.

Solución del problema 4.7. **a)** $0'87209$, **b)** $16'8119$, **c)** $97'356547$.

Solución del problema 4.8. **a)** $0'1$, **b)** $0'25$, **c)** $0'45$.

Solución del problema 4.9. **a)** $7'26094$, **b)** $5'22935$, **c)** $11'0365$.

Solución del problema 4.10. **a)** $1'315$, **b)** $-1'315$, **c)** $0'844772$.

Solución del problema 4.11. **a)** $0'95$, **b)** $0'01$, **c)** $0'245$.

Solución del problema 4.12. **a)** $2'528$, **b)** $0'687$, **c)** $1'7247$.

Solución del problema 4.13. **a)** $5'5996$, **b)** $0'416684$, **c)** $0'477478$.

Solución del problema 4.14. **a)** $0'9$, **b)** $0'025$, **c)** $0'04$.

Solución del problema 4.15. **a)** $2'538$, **b)** $3'3472$, **c)** $3'3472$.

4.4. PRÁCTICA 3: MODELOS DE PROBABILIDAD

4.4.1. Muestras aleatorias de las distribuciones usuales

En ***Minitab*** podemos generar datos de distribuciones usuales utilizando la opción **Calc⇒Random Data**. Esta opción permite generar una muestra de datos de cualquier columna de la hoja de datos actualmente abierta o de una de las distribuciones de probabilidad que aparecen listadas.

En primer lugar, creamos una nueva hoja de datos con la opción **File⇒New**. En el cuadro de diálogo que aparece seleccionamos **Minitab Woorksheet**. A esta nueva hoja de datos ***Minitab*** le asignará el nombre Worksheet J, siendo J un número natural. Luego podremos cambiarle el nombre (por ejemplo, **Probabilidad.mtw**) con la opción **File⇒Save Current Worksheet As**. A continuación, vamos a crear una columna, en dicha hoja de datos, que contenga 1000 datos aleatorios procedentes de una distribución Normal de media 5 y desviación típica 2. Para ello, seleccionamos **Calc⇒Random Data⇒Normal**; en **Number of rows of data to generate** tecleamos **1000**; en **Store in column** tecleamos el nombre **'1000 datos de N(5,2)'** (con comillas simples, al principio y al final, por llevar espacios en blanco); en **Mean** tecleamos **5** y en **Standard deviation** ponemos un **2**.

A continuación vamos a hacer el histograma, con la curva Normal superpuesta, de la muestra aleatoria obtenida en la columna **'1000 datos de N(5,2)'**. Para ello, recordemos que hay que seleccionar la opción **Graph⇒Histogram**. En el cuadro de diálogo resultante elegimos **With Fit** (para que la curva Normal aparezca superpuesta). En el siguiente cuadro de diálogo, en **Graph variables** seleccionamos, de la lista de variables que tenemos a la izquierda, la columna **'1000 datos de N(5,2)'** y pulsamos en **OK**. En la representación gráfica podemos apreciar que el histograma está cerca de la curva Normal superpuesta, lo cual es lógico puesto que hemos creado una muestra de una distribución Normal. También podemos ver, en la leyenda que aparece en la parte superior derecha del gráfico, que la media de la muestra obtenida se aproxima a 5 y la desviación típica se aproxima a 2. De hecho, cuanto mayor sea el tamaño muestral, más se aproximarán las medidas descriptivas de la muestra a las medidas descriptivas, respectivas, de la variable aleatoria Normal.

4.4.2. Función de densidad y función de probabilidad

Minitab puede calcular el resultado de la función de densidad (cuando la distribución es continua) o de la función de probabilidad (cuando la distribución es discreta) para un valor concreto o para una lista de valores. Para ello hay que elegir la opción **Calc⇒Probability Distributions** y a continuación el nombre de la variable aleatoria: Chi-square (chi-cuadrado de Pearson), Normal, F (de Snedecor), t (de Student), etc.

Dentro del cuadro de diálogo que aparecerá hay que seleccionar **Probability Density** (para las distribuciones continuas) o **Probability** (para las distribuciones discretas).

Para entender mejor el interés de esta opción, vamos a determinar los resultados de la función de densidad de una distribución $\mathcal{N}(0,1)$ (Normal Estándar) para una lista de valores que vamos a crear (todos los números comprendidos entre -4 y 4, con un incremento de $0,01$). Luego haremos la representación gráfica de esta función de densidad. Para ello se procede de la siguiente manera:

a) Mediante la opción **Calc⇒Make Patterned Data⇒Simple Set of Numbers** crearemos una nueva columna que denominaremos **'x de -4 a 4'** y que contendrá todos los números comprendidos entre el -4 y el 4 con un incremento de $0,01$. Podemos comprobar que en la columna **'x de -4 a 4'** hay 801 números.

b) En otra columna se calculan los resultados de la función de densidad de la variable aleatoria Normal Estándar para cada valor de la columna **'x de -4 a 4'**. Para hacerlo, se selecciona **Calc⇒Probability Distributions⇒Normal**; se activa **Probability density**; en **Mean** y en **Standard deviation** se deja lo que aparece por defecto (cero y uno, respectivamente); en **Input column** se selecciona, de la lista de variables de la izquierda, la columna **'x de -4 a 4'** y en **Optional storage** se teclea el nombre de la columna que contendrá los resultados de la función de densidad; por ejemplo, **'f(x) N(0,1)'**.

c) Finalmente, para representar gráficamente la función de densidad de la variable aleatoria Normal Estándar se elige la opción **Graph⇒Scatterplot**, después se elige **With connect line**. En el siguiente cuadro de diálogo, en **Y variables** se selecciona, de la lista de variables de la izquierda, la columna **'f(x) N(0,1)'** y en **X variables** se selecciona la columna **'x de -4 a 4'**. Sería conveniente quitar los puntos del gráfico, dejando sólo la línea de conexión, para lo cual se hace doble *clic* sobre la curva, en **Attributes⇒Symbols** se marca la opción **Custom** y en **Type** se selecciona **None** (buscando hacia arriba). Luego se hace un *clic* dentro del gráfico, pero no sobre la curva.

4.4.3. Función de distribución

Para calcular el resultado de la función de distribución de una variable aleatoria X, $F(t) = P(X \leq t)$, hay que elegir la opción **Calc⇒Probability Distributions** y a continuación el nombre de la variable aleatoria. Dentro del cuadro de diálogo que aparece hay que seleccionar **Cumulative Probability**.

Por ejemplo, vamos a calcular la probabilidad $P(X \leq -1'36)$, siendo X una variable aleatoria Normal Estándar. Como $P(X \leq -1'36) = F(-1'36)$, para calcular su resultado seleccionamos la opción **Calc⇒Probability Distributions⇒Normal**; activamos **Cumulative Probability**; en **Mean** y en **Standard deviation** dejamos lo que aparece por defecto (cero y uno, respectivamente). No activamos la opción **Input column** sino la opción **Input constant**, en donde colocamos el valor **-1,36**. Podemos almacenar el resultado en una constante tecleando en el recuadro **Optional storage** una K seguida de un número o poniendo un nombre a dicho resultado. Nosotros no vamos a rellenar el recuadro **Optional storage**, por lo que el resultado aparecerá en la ventana de sesión. Se puede comprobar que la probabilidad pedida es $P(X \leq -1'36) = F(-1'36) = 0'086915$.

Si queremos calcular probabilidades de los tipos $P(X > a)$, $P(a < X < b)$, etc., tenemos que utilizar lápiz y papel, y aplicar las propiedades de la probabilidad para llegar a expresiones en las que sólo aparezcan probabilidades del tipo $P(X \leq x)$ (función de distribución), pues éstas son las que calcula ***Minitab***. No tenemos que olvidar, por ejemplo, que si X es una variable aleatoria continua, entonces $P(X = a) = 0$ para todo a, por lo que se cumplen las siguientes igualdades: $P(X \leq x) = P(X < x)$, $P(X \geq x) = P(X > x)$, etc.

4.4.4. Inversa de la función de distribución (percentiles)

En ocasiones, en lugar de querer calcular probabilidades de sucesos, se desea justamente lo contrario, conocer el valor t que hace que la probabilidad del suceso $(X \leq t)$ sea igual a un valor determinado p; es decir, hallar t para que se cumpla $P(X \leq t) = p$; esto no es más que calcular percentiles de variables aleatorias. Para calcular el resultado de los percentiles de una variable aleatoria hay que elegir la opción **Calc⇒Probability Distributions** y a continuación el nombre de la variable aleatoria. Dentro del cuadro de diálogo que aparece hay que seleccionar **Inverse cumulative probability**.

Por ejemplo, vamos a calcular el valor t que verifica $P(X \leq t) = 0'98$ cuando X tiene una distribución chi-cuadrado de Pearson con 20 grados de libertad. Expresado de otra manera, vamos a determinar el valor del percentil al 98 % de una variable aleatoria chi-cuadrado de Pearson con 20 grados de libertad; es decir, $\chi^2_{20,0'98}$. Para ello seleccionamos la opción **Calc⇒Probability Distributions⇒Chi-Square**. En el cuadro de diálogo activamos **Inverse cumulative probability**. Dejamos lo que aparece por defecto (cero) en **Noncentrality parameter**. En **Degrees of freedom** tecleamos **20**. No activamos la opción **Input column** sino la opción **Input constant**, en donde colocamos el valor **0,98**. Podemos almacenar el resultado en una constante tecleando en el recuadro **Optional storage** una K seguida de un número o poniendo un nombre a dicho resultado. Nosotros no vamos a rellenar el recuadro **Optional storage**, por lo que el resultado aparecerá en la ventana de sesión. Se puede comprobar que el valor t que verifica $P(X \leq t) = 0'98$ es $t = 35'0196$. Es decir, si $X \equiv \chi^2_{20}$ entonces $P(X \leq 35'0196) = 0'98$; o sea, $\chi^2_{20,0'98} = 35'0196$.

5

Tests no paramétricos en una población

5.1. Desarrollo de los contenidos fundamentales (teoría y PRÁCTICA 4)

5.1.1. Introducción a la Estadística Inferencial

Como ya sabemos, la Estadística estudia los métodos científicos para recoger, organizar, resumir y analizar datos de una o varias muestras, extrayendo, a través del cálculo de probabilidades, conclusiones válidas que nos permitan tomar decisiones sobre la población. En el bloque 1 de esta asignatura hemos estudiado la rama de la Estadística que se ocupa de describir y analizar los datos de las muestras, sin sacar conclusiones sobre un conjunto mayor de datos; es decir, hemos estudiado *Estadística Descriptiva*. En el bloque 2 se han resuelto problemas relativos al *Cálculo de Probabilidades* de ciertos sucesos relacionados con variables aleatorias que seguían determinadas distribuciones de parámetros conocidos. Sin embargo, siendo los parámetros algo característico de toda población, es usual que sean desconocidos. En el bloque 3, que ahora comienza, vamos a estudiar la rama de la Estadística que trata de sacar conclusiones o inferencias sobre un grupo grande de datos (población) a partir de un subgrupo de datos (muestra), incluyendo el problema de la determinación aproximada de los parámetros de la población. Esta rama se llama **Estadística Inferencial**.

La utilización de un método adecuado de muestreo garantiza que la muestra obtenida es representativa de la población. Esto significa que la información proporcionada por la muestra es un reflejo de la información contenida en la población. Podemos, por tanto, utilizar la información muestral para formarnos una idea sobre las propiedades de la población. Es decir, podemos servirnos de las muestras para hacer *inferencias* sobre la población.

Estas inferencias pueden adoptar diferentes formas pero las más habituales son dos: la *estimación de parámetros* y el *test de hipótesis*. Cuando la información deseada de la población es el valor de alguno de sus parámetros, la técnica a utilizar es la **estimación de parámetros**. Los **tests de hipótesis** permiten comprobar si ciertas hipótesis que se enuncian acerca de la población son correctas o no.

La estimación de parámetros puede ser:

- **Estimación puntual**: consiste en asignar un valor muestral concreto al parámetro poblacional que se desea estimar.
- **Estimación por intervalo de confianza**: consiste en atribuir al parámetro que se desea estimar, no un valor concreto, sino un rango de valores entre los que se espera que pueda encontrarse el verdadero valor del parámetro con una probabilidad alta y conocida.

5.1.2. Tests de hipótesis

- *Hipótesis estadística*: afirmación sobre la forma de una o más distribuciones, o sobre el valor de uno o más parámetros de esas distribuciones.
- *Hipótesis nula*: hipótesis estadística que se somete a contraste. Se denota por H_0.
- *Hipótesis alternativa*: es la negación de la hipótesis nula H_0, e incluye todo lo que H_0 excluye. Se denota por H_1.
- *Test (o contraste) de hipótesis*: procedimiento que nos capacita para determinar si las muestras observadas difieren significativamente de los resultados esperados, y por tanto nos ayuda a decidir si aceptamos o rechazamos la hipótesis nula.
  - *Test paramétrico*: la hipótesis nula es una afirmación sobre el valor de uno o más parámetros de la variable aleatoria observada en la población.
  - *Test no paramétrico*: la hipótesis nula no es una afirmación sobre el valor de uno o más parámetros de la variable aleatoria observada en la población.
- *Estadístico del test*: estadístico que se observa al realizar un test de hipótesis, y que nos sirve para aceptar o rechazar la hipótesis nula por poseer una distribución muestral conocida.
- *Región crítica*: zona de la distribución muestral del estadístico del test que corresponde a los valores que permiten rechazar la hipótesis nula, y por tanto aceptar la hipótesis alternativa.
- *Región de aceptación*: zona de la distribución muestral del estadístico del test que corresponde a los valores que permiten aceptar la hipótesis nula.
- *Error de tipo I*: error que se comete cuando se decide rechazar una hipótesis nula que en realidad es verdadera.
- *Nivel de significación*: probabilidad de cometer un error de tipo I al contrastar una hipótesis. Se denota por α.
- *Error de tipo II*: error que se comete cuando se decide aceptar una hipótesis nula que en realidad es falsa. La probabilidad de cometer dicho error se denota por β.
- **p-valor** o *nivel crítico*: es el nivel de significación más pequeño al que una hipótesis nula puede ser rechazada con el estadístico del test obtenido. Se rechaza H_0 si el p-valor es claramente menor que α; se acepta H_0 si el p-valor es claramente mayor que α; y se repite el test con una muestra diferente si el p-valor tiene un resultado próximo a α.
- En todos los tests de hipótesis que realicemos con ***Minitab*** nos tenemos que fijar en el p-valor ya que:

  Si p-valor $> \alpha \Rightarrow$ aceptamos la hipótesis nula.

  Si p-valor $< \alpha \Rightarrow$ rechazamos la hipótesis nula y, por tanto, aceptamos la hipótesis alternativa.

5.1.3. Test de las rachas sobre aleatoriedad de la muestra

5.1.3.1. Introducción

Con frecuencia las muestras se toman en serie temporal, cabiendo la posibilidad de que una observación dependa de la observación anterior. De ocurrir esto, la muestra no es aleatoria. Como tal propiedad es la base de la Estadística Inferencial, todos los tests de hipótesis quedarán invalidados si falla la hipótesis de aleatoriedad de la muestra. Resulta, pues, crucial dar procedimientos que permitan contrastar la hipótesis nula H_0: *la muestra es aleatoria* contra la hipótesis alternativa H_1: *la muestra no es aleatoria*. Los tests para ello son diversos, pero el más utilizado es el que describimos a continuación, que se denomina **test de las rachas**.

5.1.3.2. Hipótesis nula y alternativa del test

$$H_0 : \textit{La muestra de datos de la variable } X \textit{ es aleatoria.}$$
$$H_1 : \textit{La muestra de datos de la variable } X \textit{ no es aleatoria.}$$

5.1.3.3. Condiciones para poder realizar el test

No es necesario que se cumpla ninguna condición especial para poder realizar este test.

5.1.3.4. Resolución mediante MINITAB

Con ***Minitab*** el test de las rachas sobre aleatoriedad de una muestra se realiza mediante la opción **Stat⇒Nonparametrics ⇒Run Test**.

Este test se basa en el concepto de racha, que es una secuencia de observaciones de un mismo tipo precedida y continuada por otro tipo de observaciones o por ninguna. Esto supone que los datos son sólo de dos tipos; es decir, que la variable está dicotomizada. Si esto no sucediera, se pueden reducir los datos a dos tipos mediante lo siguiente: asignar un símbolo (por ejemplo, "+") a los datos que son mayores que la media (o la mediana) y otro símbolo (por ejemplo, "−") a los que son menores o iguales que la media (o la mediana, respectivamente).

Con los datos del archivo **Pulse.mtw** vamos a comprobar si se puede aceptar, con un nivel de significación de $0'05$, que las muestras de datos de las columnas **Pulse1**, **Pulse2**, **Height** y **Weight** son aleatorias.

En primer lugar, abrimos la hoja de datos **Pulse.mtw** (con la opción **File⇒Open Worksheet**).

A continuación, seleccionamos **Stat⇒Nonparametrics ⇒Run Test**. En el cuadro de diálogo resultante, activamos el recuadro **Variables** (haciendo *clic* dentro de él); seleccionamos (haciendo doble *clic* sobre sus nombres) las columnas **Pulse1**, **Pulse2**, **Height** y **Weight**. Como vamos a comprobar la aleatoriedad de más de una muestra, tenemos que dicotomizar mediante las respectivas medias (no podemos dicotomizar mediante las respectivas medianas). Por tanto, activamos la opción **Above and below the mean** y pulsamos en **OK**.

Si hubiésemos comprobado la aleatoriedad de una sola muestra, podríamos haber dicotomizado mediante la mediana, para lo cual habríamos calculado previamente el valor de dicha mediana; habríamos activado la opción **Above and below:** y, al lado, habríamos tecleado el resultado de dicha mediana.

En la ventana de sesión nos aparecen los resultados de los cuatro tests. Para la variable **Pulse1**, el p-valor es $0'368$, mayor que el nivel de significación elegido ($0'05$), por lo que aceptamos la hipótesis nula; es decir, aceptamos que la muestra de resultados de dicha variable es aleatoria.

Para la variable **Pulse2**, el p-valor es $0'002$, menor que el nivel de significación elegido ($0'05$), por lo que rechazamos la hipótesis nula; es decir, rechazamos que la muestra de resultados de dicha variable es aleatoria.

Para la variable **Height**, el p-valor es 0, menor que el nivel de significación elegido ($0'05$), por lo que rechazamos que la muestra de resultados de dicha variable es aleatoria.

Para la variable **Weight**, el p-valor es $0'001$, menor que el nivel de significación elegido ($0'05$), por lo que rechazamos que la muestra de resultados de dicha variable es aleatoria.

5.1.4. Tests sobre normalidad de la variable aleatoria

5.1.4.1. Introducción

El problema de comprobar la normalidad de una variable aleatoria, a partir de los datos proporcionados por una muestra, ha sido tratado a menudo debido al uso frecuente de esta hipótesis en la Estadística Inferencial. Existen diversos tests para contrastar la hipótesis nula H_0: *la variable aleatoria observada en la población es Normal* frente a la hipótesis alternativa H_1: *la variable aleatoria observada en la población no es Normal*. Algunos de ellos son: el **test de Kolmogorov-Smirnov**, el **test de Anderson-Darling**, el **test de Ryan-Joiner** y el **test de Shapiro-Wilk**.

5.1.4.2. Hipótesis nula y alternativa del test

H_0 : *La variable aleatoria X es Normal.*

H_1 : *La variable aleatoria X no es Normal.*

5.1.4.3. Condiciones para poder realizar el test

Es necesario que se verifique que la muestra de datos de la variable X sea aleatoria.

5.1.4.4. Resolución mediante MINITAB

En ***Minitab*** hay varias técnicas para comprobar el ajuste a una distribución Normal. Una de ellas es la opción **Graph⇒Probability Plot**. Con esta opción es posible comprobar la normalidad de varias variables a la vez.

Vamos a utilizar este método para comprobar qué variables de la hoja de datos **Marks.mtw** se ajustan al modelo Normal (cuando están observadas en toda la población). El archivo **Marks.mtw** es una hoja de datos que ***Minitab*** tiene de muestra y se encuentra en **C:\Archivos de programa\Minitab 15\English\Sample Data\Student9**. En las aulas de informática de la Universidad de Murcia este archivo de datos se encuentra en **C:\Archivos de programa\UM\Minitab 15\English\Sample Data\Student9**.

En primer lugar, abrimos dicha hoja de datos (**File⇒Open Worksheet**). El archivo muestra las calificaciones (puntuadas de 0 a 100) de 24 estudiantes en tres exámenes de tipo test (**Test1**, **Test2** y **Test3**).

En segundo lugar, vamos a comprobar que las muestras de los datos de las columnas **Test1**, **Test2** y **Test3** son aleatorias (**Stat⇒Nonparametrics ⇒Run Test**).

En tercer lugar, vamos a ver si se puede aceptar que las variables **Test1**, **Test2** y **Test3** son Normales. Para ello, seleccionamos **Graph⇒Probability Plot**. En el cuadro de diálogo resultante seleccionamos **Single** y pulsamos en **OK**. En **Graph variables** seleccionamos, de la lista de variables de la izquierda, las que podrían ajustarse a un modelo Normal; es decir, **Test1**, **Test2** y **Test3**. Pulsamos en **Distribution** y, en el cuadro de diálogo resultante, dejamos lo que está activado por defecto; es decir, **Normal**, y no rellenamos la opción **Historical Parameters** ya que no sabemos los resultados de las estimaciones de la media y de la desviación típica poblacionales.

Nos aparecen tres gráficos, uno para cada una de las variables seleccionadas. Además, vemos que aparecen, en la parte superior derecha de las representaciones gráficas, los resultados de un test de normalidad; concretamente, el test de Anderson-Darling.

Podemos ver que el gráfico probabilístico de la variable **Test1** se aproxima a una recta. Además, el p-valor del test de normalidad es igual a $0,232$ y, por tanto, es mayor que los usuales niveles de significación ($\alpha = 0,05$ o $\alpha = 0,01$). En consecuencia, podemos aceptar que la variable **Test1** se ajusta al modelo Normal.

Por otra parte, podemos observar que el gráfico probabilístico de la variable **Test2** también se aproxima a una recta. Además, el p-valor del test de normalidad es igual a $0,119$ y, por tanto, es mayor que los usuales niveles de significación ($\alpha = 0,05$ o $\alpha = 0,01$). En consecuencia, podemos aceptar que la variable **Test2** se ajusta al modelo Normal.

Por último, el gráfico probabilístico de la variable **Test3** no se aproxima a una recta. Además, el p-valor del test de normalidad es menor que $0,007$. Tanto si consideramos un nivel de significación de $\alpha = 0,01$ como si consideramos un nivel de significación de $\alpha = 0,05$ resulta que el p-valor es menor que α. En consecuencia, la variable **Test3** no se ajusta al modelo Normal.

5.2. Ejemplos que se van a resolver en clase

Ejemplo 5.1. En la tabla siguiente aparecen los datos de 10 bibliotecas, en las cuales se ha observado las siguientes variables: número total de títulos catalogados en un año (X), número de horas totales al año que emplea la biblioteca en catalogar sus títulos (Y) y costo, en euros, de una hora de catalogación (Z).

x_i	y_i	z_i
1550	220	15'75
1640	230	14'50
1000	140	16'40
950	135	16'70
750	110	17'10
1700	255	12'50
1650	228	14'80
1860	270	15'25
1900	280	18'50
900	130	17'30

a) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que las tres muestras son aleatorias?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que las tres variables (X, Y y Z) son Normales?

5.3. Actividades de aplicación de los contenidos

5.3.1. Problemas propuestos

Problema 5.1. En una muestra aleatoria simple de 15 individuos que consultan bases de datos, el tiempo (en minutos) que están utilizando el ordenador para realizar esta tarea es:

22	13	17	14	15	18	19	14	17	20	21	13	15	18	17

a) Crea un nuevo proyecto de ***Minitab***. Introduce los datos y grábalos con el nombre **Tiempo-Consulta.mtw**. ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la muestra es aleatoria? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la variable aleatoria X=*tiempo (en minutos) empleado en consultar bases de datos por ordenador* es Normal? ¿Por qué?

c) Si así lo deseas, puedes grabar el proyecto de ***Minitab***.

Problema 5.2. Los siguientes datos corresponden a las edades de una muestra de 10 personas que visitan una biblioteca.

19	24	83	30	17	23	33	19	68	56

a) Crea un nuevo proyecto de ***Minitab***. Introduce los datos y grábalos con el nombre con el nombre **Edad-Visitantes-Bca.mtw**. ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la muestra es aleatoria? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la variable aleatoria X=*edad de las personas que visitan la biblioteca* es Normal? ¿Por qué?

c) Si así lo deseas, puedes grabar el proyecto de ***Minitab***.

Problema 5.3. La tabla siguiente contiene el número mensual de materias buscadas por los usuarios de una biblioteca (X) y el número mensual de materias localizadas por dichos usuarios (Y):

materias buscadas (x_i)	materias localizadas (y_i)
42	22
65	30
68	35
55	30
35	20
40	25
50	30
26	15
42	22
56	38
38	15
50	34

a) Crea un nuevo proyecto de ***Minitab***. Introduce los datos de las dos variables. Guarda la hoja de datos con el nombre **Materias-Buscadas-Localizadas.mtw**. ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'03$, que las dos muestras son aleatorias? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'03$, que las dos variables (X e Y) son Normales? ¿Por qué?

c) Si así lo deseas, puedes grabar el proyecto de ***Minitab***.

Problema 5.4. En la tabla siguiente aparecen los resultados del peso, en gramos (X) y del precio, en euros (Y) de una muestra de 12 libros.

x_i	y_i
325	110
890	30
415	75
400	45
515	32
650	69
790	30
890	34
320	42
420	46
620	53
720	97

a) Crea un nuevo proyecto de ***Minitab***. Introduce los datos de las dos variables. Guarda la hoja de datos con el nombre **Peso-Precio-Libros.mtw**. ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'01$, que las dos muestras son aleatorias? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'01$, que las dos variables (X e Y) son Normales? ¿Por qué?

c) Si así lo deseas, puedes grabar el proyecto de ***Minitab***.

Problema 5.5.

a) Crea un nuevo proyecto de ***Minitab***. Abre la hoja de datos **Prestamos.mtw** (datos del Ejercicio 1.1). ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la muestra de datos de la variable **PPU** (**porcentaje anual de préstamos por usuario**) es aleatoria? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la variable **PPU** es Normal? ¿Por qué?

c) Si así lo deseas, puedes grabar el proyecto de ***Minitab***.

Problema 5.6.

a) Crea un nuevo proyecto de ***Minitab***. Abre la hoja de datos **Transacciones.mtw** (datos del Ejercicio 1.2). ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que las muestras de los datos de las variables **TR**, **TRF** y **Porcentaje TRF** son aleatorias? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que las variables **TR**, **TRF** y **Porcentaje TRF** son Normales? ¿Por qué?

c) Si así lo deseas, puedes grabar el proyecto de ***Minitab***.

5.3.2. Soluciones de los problemas propuestos

Solución del problema 5.1.

X=*Tiempo (en minutos) empleado en consultar bases de datos por ordenador.*

a) Hacemos el test de las rachas sobre aleatoriedad de la muestra.

Las hipótesis nula y alternativa son:

H_0 : *La muestra de datos de la variable X es aleatoria.*

H_1 : *La muestra de datos de la variable X no es aleatoria.*

Con la opción **Stat⇒Nonparametrics ⇒Run Test** de ***Minitab*** obtenemos un p-valor de $0'654$, mayor que el nivel de significación ($\alpha = 0'05$); por tanto, podemos aceptar H_0; es decir, la muestra de datos de la variable X es aleatoria.

b) Tenemos que hacer un test sobre normalidad de la variable aleatoria X.

Para poder realizar este test de normalidad es necesario comprobar, previamente, que la muestra de datos de la variable X es aleatoria. Efectivamente, esta condición se cumple pues lo hemos probado en el apartado anterior.

Las hipótesis nula y alternativa son:

H_0 : *La variable aleatoria X es Normal.*

H_1 : *La variable aleatoria X no es Normal.*

Con la opción **Graph⇒Probability Plot** de ***Minitab*** obtenemos un p-valor de $0'587$, que es mayor que el nivel de significación ($\alpha = 0'05$) por lo que podemos aceptar H_0; es decir, podemos aceptar que la variable aleatoria X es Normal.

Solución del problema 5.2.

X=*Edad de las personas que visitan la biblioteca.*

a) Podemos aceptar que la muestra de datos de la variable X es aleatoria pues el p-valor ($0'326$) es mayor que el nivel de significación ($0'05$).

b) No podemos aceptar que la variable aleatoria X sea Normal pues el p-valor ($0'022$) es menor que el nivel de significación ($0'05$).

Solución del problema 5.3.

X=*Número mensual de materias buscadas por los usuarios de una biblioteca.*

Y=*Número mensual de materias localizadas por los usuarios de dicha biblioteca.*

a) Podemos aceptar que las dos muestras de datos son aleatorias pues los dos p-valores ($0'545$ para X y $0'545$ para Y) son mayores que el nivel de significación ($0'03$).

b) Podemos aceptar que las dos variables, X e Y, son Normales pues los dos p-valores ($0'837$ para X y $0'544$ para Y) son mayores que el nivel de significación ($0'03$).

Solución del problema 5.4.

X=*Peso, en gramos, de los libros.*

Y=*Precio, en euros, de los libros.*

a) Podemos aceptar que las dos muestras de datos son aleatorias pues los dos p-valores ($0'545$ para X y $0'646$ para Y) son mayores que el nivel de significación ($0'01$).

b) Podemos aceptar que las dos variables, X e Y, son Normales pues los dos p-valores ($0'335$ para X y $0'064$ para Y) son mayores que el nivel de significación ($0'01$).

Solución del problema 5.5.

a) No podemos aceptar que la muestra de datos de la variable **PPU** sea aleatoria pues el p-valor ($0'009$) es menor que el nivel de significación ($0'05$).

b) Como la muestra de datos de la variable **PPU** no es aleatoria, entonces no podemos realizar el test sobre normalidad de dicha variable.

Solución del problema 5.6.

a) Podemos aceptar que las tres muestras son aleatorias pues los tres p-valores ($0'212$ para **TR**, $0'212$ para **TRF** y $0'609$ para **Porcentaje TRF**) son mayores que el nivel de significación ($0'05$).

b) Podemos aceptar que las tres variables son Normales pues los tres p-valores ($0'081$ para **TR**, $0'057$ para **TRF** y $0'363$ para **Porcentaje TRF**) son mayores que el nivel de significación ($0'05$).

6

Estimación y tests paramétricos en una población

6.1. Desarrollo de los contenidos fundamentales (teoría y PRÁCTICA 5)

6.1.1. Tests sobre la media poblacional. Intervalo de confianza para la media

6.1.1.1. Test sobre la media cuando la desviación típica poblacional es conocida

6.1.1.1.1. Introducción

Formalmente, se puede definir un intervalo de confianza de la siguiente manera. Sea ω un parámetro de una población (puede ser μ, σ, etc.). Sea α una probabilidad pequeña (habitualmente, igual a $0'01$ ó $0'05$). El intervalo (L_1, L_2) se llama un **intervalo de confianza** para ω al **nivel de confianza** $1-\alpha$, o al $100(1-\alpha)\,\%$, si la probabilidad de que dicho intervalo contenga al parámetro es mayor o igual que $1-\alpha$.

Cuando el intervalo se construye de manera que tomando muchas muestras y calculando el intervalo con cada una de ellas, el 95 % de los intervalos así construidos incluyan el valor del parámetro, decimos que son intervalos al 95 % de confianza. Por tanto, el nivel de confianza no es la probabilidad de que un intervalo concreto incluya o no el valor del parámetro, ya que al ser el parámetro un valor fijo estará o no dentro de un intervalo concreto. El nivel de confianza se refiere a la probabilidad de que, al tomar todas las muestras posibles, el intervalo contenga el parámetro; es decir, expresa el porcentaje de intervalos que efectivamente incluyen el parámetro.

El intervalo de confianza para la media poblacional con varianza poblacional conocida y el test de hipótesis sobre la media poblacional con varianza poblacional conocida están basados en lo siguiente:

Si la variable observada en la población, X, es Normal $\mathcal{N}(\mu, \sigma)$ y extraemos muestras aleatorias de tamaño n, entonces la media muestral $\overline{X}$ es una variable aleatoria Normal de media μ y desviación

típica $\sigma/\sqrt{n}$; es decir, $\overline{X} \equiv \mathcal{N}(\mu, \sigma/\sqrt{n})$. Aunque X no sea Normal, la distribución del estadístico $\overline{X}$ se aproxima a una Normal $\mathcal{N}(\mu, \sigma/\sqrt{n})$ cuando el tamaño muestral n va aumentando (en la práctica se considera válida la aproximación cuando $n \geq 30$). En los dos casos, la variable aleatoria *tipificada*

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

sigue una distribución Normal Estándar (o se aproxima a ella).

Por esta razón, a este test se le denomina **test Z sobre una media**.

6.1.1.1.2. Hipótesis nula y alternativa del test

Hay tres posibilidades:

$H_0 : \mu = \mu_0$	$H_0 : \mu \geq \mu_0$	$H_0 : \mu \leq \mu_0$
$H_1 : \mu \neq \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$

6.1.1.1.3. Condiciones para poder realizar el test

Para realizar cualquiera de los tres tipos de tests anteriores, es necesario que se verifiquen las condiciones siguientes:

- La muestra de datos de la variable X es aleatoria.
- La variable aleatoria X es Normal o el tamaño muestral, n, es grande ($n \geq 30$).
- La desviación típica poblacional, σ, es conocida.

6.1.1.1.4. Resolución mediante MINITAB

Para hacer el test sobre la media cuando la desviación típica poblacional es conocida hay que seleccionar **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**1-Sample Z**. Esta opción también nos da el intervalo de confianza para la media poblacional, μ.

Abrimos el archivo de datos **Pulse.mtw**. Vamos a suponer que conocemos el valor de la desviación típica poblacional de la variable **Pulse1** (pulso antes de correr), $\sigma = 10$ pulsaciones por minuto. Comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional antes de correr es mayor que 70 pulsaciones por minuto. Si μ denota la media poblacional de la variable X=*Pulso antes de correr*, las hipótesis nula y alternativa son $H_0 : \mu \leq 70$ y $H_1 : \mu > 70$.

En el tema anterior ya hemos comprobado que la muestra de resultados de la variable **Pulse1** es aleatoria. Además, el tamaño muestral es grande ($n = 92$). Por tanto, podemos utilizar este procedimiento estadístico.

Seleccionamos la opción **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**1-Sample Z**. En **Samples in columns** seleccionamos, de la lista de variables de la izquierda, la columna o columnas para las cuales se va a realizar este tipo de test; en nuestro caso, '**Pulse1**'. Dejamos desactivada la opción **Summarized data**. En **Standard deviation** tecleamos el valor de la desviación típica poblacional, σ, que suponemos que es **10**. Activamos **Perform hypothesis test** y en **Hypothesized mean** especificamos el valor, μ_0, con el que se compara la media

poblacional, que es **70**. Si pulsamos el botón **Options** nos aparece un nuevo cuadro de diálogo con las siguientes opciones:

Confidence level: Por defecto se muestra un intervalo de confianza al 95 % para la media poblacional μ. Se puede introducir un valor entre 1 y 99 para solicitar otro nivel de confianza. En nuestro caso, podemos dejar lo que aparece por defecto, es decir, 95.

Alternative: Aquí se especifica cuál es la hipótesis alternativa: **less than** significa que la hipótesis alternativa es $H_1 : \mu < \mu_0$, **not equal** significa que la hipótesis alternativa es $H_1 : \mu \neq \mu_0$ y **greater than** significa que la hipótesis alternativa es $H_1 : \mu > \mu_0$. Tengamos en cuenta que con la opción **less than** el intervalo de confianza para la media será del tipo $(-\infty, b)$, con la opción **not equal** el intervalo de confianza para la media será del tipo (a, b) y con la opción **greater than** el intervalo de confianza para la media será del tipo $(a, +\infty)$. En nuestro caso, tenemos que seleccionar **greater than** ya que la hipótesis alternativa es $H_1 : \mu > 70$.

Podemos comprobar, en la ventana de sesión, que el p-valor es $0'003$, claramente menor que el nivel de significación, $\alpha = 0'05$. En consecuencia, rechazamos la hipótesis nula y, por tanto, aceptamos la hipótesis alternativa; es decir, aceptamos que la media poblacional de la variable **Pulse 1** es mayor que 70 pulsaciones por minuto. El intervalo de confianza al 95 % para la media poblacional, asociado a este test de hipótesis, es $(71'15, +\infty)$.

También se puede realizar este test de hipótesis **si sabemos el tamaño muestral y el resultado de la media muestral**. Veámoslo con un ejemplo:

En el volumen de Julio de 1992 de Economics Abstracts, la media del número de palabras por resumen es $79'56$, con una varianza de $615'04$. Se extrae una muestra aleatoria simple de 30 resúmenes escritos en alemán y se observa que la media del número de palabras por resumen es $67'47$. Se quiere decidir si existe una diferencia significativa entre la media de palabras por resumen de los escritos en alemán y la media de palabras por resumen de todos los de este volumen.

Vamos a suponer que la varianza del número de palabras por resumen de los escritos en alemán coincide con la varianza del número de palabras por resumen de todos los de este volumen. Así pues, los datos que tenemos son los siguientes:

$$\begin{aligned} &\mu_0 = 79'56\,, \\ &\sigma^2 = 615'04 \Rightarrow \sigma = \sqrt{615'04} = 24'8\,, \\ &\overline{X} = 67'47\,, \\ &n = 30\,. \end{aligned}$$

La variable observada en la población no puede ser Normal pues es discreta, pero como el tamaño muestral es 30, entonces podemos aplicar esta técnica. Así pues, las hipótesis nula y alternativa son:

$$\begin{aligned} &H_0 : \mu = 79'56\,, \\ &H_1 : \mu \neq 79'56\,. \end{aligned}$$

Seleccionamos la opción **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**1-Sample Z**. Activamos la opción **Summarized data**, con lo cual se desactiva automáticamente la opción **Samples in columns**. En **Sample size** tenemos que teclear el tamaño muestral, que es **30** y en **Mean** tenemos que teclear el resultado de la media muestral,

que es **67,47**. En **Standard deviation** tecleamos el valor de la desviación típica poblacional, σ, que suponemos que es **24,8**. Activamos **Perform hypothesis test** y en **Hypothesized mean** especificamos el valor, μ_0, con el que se compara la media poblacional, que es **79,56**. Pulsamos en **Options** y, en el cuadro de diálogo resultante, en **Alternative** seleccionamos **not equal** puesto que nuestra hipótesis alternativa es $H_1 : \mu \neq 79'56$. Dentro de este cuadro de diálogo también podemos cambiar el nivel de confianza del intervalo de confianza para la media poblacional; por defecto, este nivel de confianza se fija en el 95 %; si queremos cambiarlo tenemos que modificar el valor de **Confidence level**. Nosotros dejaremos lo que está puesto por defecto: 95 %.

Podemos comprobar, en la ventana de sesión, que el p-valor es $0'008$, claramente menor que los niveles de significación usuales ($\alpha = 0'05$ ó $\alpha = 0'01$). En consecuencia, rechazamos la hipótesis nula y, por tanto, aceptamos que existe diferencia significativa entre la media del número de palabras por resumen en alemán y la media del número de palabras por resumen de todos ellos. El intervalo de confianza al 95 % para la media poblacional, asociado a este test de hipótesis, es $(58'60, 76'34)$.

6.1.1.2. Test sobre la media cuando la desviación típica poblacional es desconocida

6.1.1.2.1. Introducción

El intervalo de confianza para la media poblacional con varianza poblacional desconocida y el test de hipótesis sobre la media poblacional con varianza poblacional desconocida están basados en lo siguiente:

Si la variable observada en la población, X, es Normal $\mathcal{N}(\mu, \sigma)$ y extraemos muestras aleatorias de tamaño n, entonces la nueva variable aleatoria

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

sigue una distribución t de Student con $n - 1$ grados de libertad; es decir, $T \equiv t_{n-1}$.

Si la variable aleatoria observada en la población no es Normal, se verifica que la distribución de la variable T se aproxima a una t de Student con $n - 1$ grados de libertad cuando el tamaño muestral n va aumentando. En la práctica, es aceptable esta aproximación cuando $n \geq 30$.

Por esta razón, a este test se le denomina **test t de Student sobre una media**.

6.1.1.2.2. Hipótesis nula y alternativa del test

Hay tres posibilidades:

$H_0 : \mu = \mu_0$	$H_0 : \mu \geq \mu_0$	$H_0 : \mu \leq \mu_0$
$H_1 : \mu \neq \mu_0$	$H_1 : \mu < \mu_0$	$H_1 : \mu > \mu_0$

6.1.1.2.3. Condiciones para poder realizar el test

Para realizar cualquiera de los tres tipos de tests anteriores, es necesario que se verifiquen las condiciones siguientes:

- La muestra de datos de la variable X es aleatoria.

- La variable aleatoria X es Normal o el tamaño muestral, n, es grande ($n \geq 30$).
- La desviación típica poblacional, σ, es desconocida.

6.1.1.2.4. Resolución mediante MINITAB

Para hacer el test sobre la media cuando la desviación típica poblacional es desconocida hay que seleccionar **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**1-Sample t**.

La manera de utilizar esta opción es la misma que la explicada en el apartado 6.1.1.1.4, por lo que no vamos a repetir ahora todo el proceso.

Con el archivo de datos **Pulse.mtw**, veamos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional antes de correr es igual a 71 pulsaciones por minuto. Lo que queremos comprobar es si la media poblacional de la variable **Pulse1** es igual a 71 pulsaciones por minuto, suponiendo ahora desconocida la desviación típica poblacional (lo cual es cierto). Si μ denota la media poblacional de la variable **Pulse1**, las hipótesis nula y alternativa son $H_0 : \mu = 71$ y $H_1 : \mu \neq 71$.

Podemos comprobar, en la ventana de sesión, que el p-valor es $0'107$, claramente mayor que el nivel de significación, $\alpha = 0'05$, por lo que podemos aceptar la hipótesis nula; es decir, aceptamos que la media poblacional del número de pulsaciones por minuto antes de correr es igual a 71. El intervalo de confianza al 95 % para la media poblacional de dicha variable es $(70'59, 75'15)$.

También se puede realizar este test de hipótesis **si sabemos el tamaño muestral, el resultado de la media muestral y el resultado de la cuasi-desviación típica muestral**. Veámoslo con un ejemplo:

El número medio de libros por estante de una biblioteca es 24. Extraída una muestra de 91 estantes de libros de matemáticas se obtiene una media de 25 libros, con una cuasi-desviación típica de $1'5$. Queremos decidir si existe diferencia significativa entre el número medio de libros de matemáticas por estante y el número medio de libros por estante.

La variable $X =$ *"Número de libros de matemáticas por estante"* no puede ser Normal porque es discreta; pero como $n = 91 \geq 30$ entonces se puede utilizar este procedimiento.

Los datos conocidos son:

$$\begin{aligned} \mu_0 &= 24\,, \\ S &= 1'5\,, \\ \overline{X} &= 25\,, \\ n &= 91\,. \end{aligned}$$

Las hipótesis nula y alternativa son :

$$\begin{aligned} H_0 &: \mu = 24\,, \\ H_1 &: \mu \neq 24\,. \end{aligned}$$

Seleccionamos la opción **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**1-Sample t**. Activamos la opción **Summarized data**, con lo cual se desactiva automáticamente la opción **Samples in columns**. En **Sample size** tenemos que teclear el tamaño muestral, que es **91**, en **Mean** tenemos que teclear el resultado de la media muestral, que es **25**, y en **Standard deviation** tenemos que teclear el resultado de la cuasi-desviación típica muestral, que es **1,5**. Activamos **Perform hypothesis test** y en **Hypothesized mean** especificamos el valor, μ_0, con

el que se compara la media poblacional, que es **24**. Pulsamos en **Options** y, en el cuadro de diálogo resultante, en **Alternative** seleccionamos **not equal** puesto que nuestra hipótesis alternativa es $H_1 : \mu \neq 24$. Como ya sabemos, dentro de este cuadro de diálogo también podemos cambiar el nivel de confianza del intervalo de confianza para la media poblacional; por defecto, este nivel de confianza es 95 %; si queremos cambiarlo tenemos que modificar el valor de **Confidence level**. Nosotros dejaremos lo que está puesto por defecto: 95 %.

Podemos comprobar, en la ventana de sesión, que el p-valor es 0, el mínimo posible y, por supuesto, claramente menor que los niveles de significación usuales ($\alpha = 0'05$ ó $\alpha = 0'01$). En consecuencia, rechazamos la hipótesis nula y, por tanto, aceptamos que existe diferencia significativa entre el número medio de libros de matemáticas por estante y el número medio de libros por estante. El intervalo de confianza al 95 % para la media poblacional, asociado a este test de hipótesis, es $(24'688, 25'312)$.

6.1.2. Tests sobre la varianza poblacional

6.1.2.1. Introducción

Si desconocemos el valor de la varianza de la población, es lógico que desconozcamos también el valor de la media poblacional, por lo que vamos a desarrollar solamente este caso.

El intervalo de confianza para la varianza poblacional con media poblacional desconocida y el test de hipótesis sobre la varianza poblacional con media poblacional desconocida están basados en lo siguiente:

Si la variable observada en la población, X, es Normal, extrayendo muestras aleatorias de tamaño n se verifica que el estadístico

$$V = \frac{(n-1)S^2}{\sigma^2} = \frac{ns^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2}$$

es una variable aleatoria que sigue una distribución chi-cuadrado con $n-1$ grados de libertad.

6.1.2.2. Hipótesis nula y alternativa del test

Hay tres posibilidades:

$H_0 : \sigma^2 = \sigma_0^2$	$H_0 : \sigma^2 \geq \sigma_0^2$	$H_0 : \sigma^2 \leq \sigma_0^2$
$H_1 : \sigma^2 \neq \sigma_0^2$	$H_1 : \sigma^2 < \sigma_0^2$	$H_1 : \sigma^2 > \sigma_0^2$

6.1.2.3. Condiciones para poder realizar el test

Para realizar cualquiera de los tres tipos de tests anteriores, es necesario que se verifiquen las condiciones siguientes:

- La muestra de datos de la variable X es aleatoria.
- La variable aleatoria X es Normal.
- La media poblacional, μ, es desconocida.

6.1.2.4. Resolución mediante MINITAB

Para hacer el test de hipótesis sobre una varianza poblacional con media poblacional desconocida hay que seleccionar **Stat ⇒Basic Statistics ⇒1 Variance**. Esta opción también se utiliza para realizar un test sobre la desviación típica poblacional. Además de hacer dichos tests, ***Minitab*** también nos da el intervalo de confianza para la varianza poblacional (o para la desviación típica poblacional).

En el tema anterior ya hemos comprobado que la muestra de resultados de la variable **Pulse1** (del archivo de datos **Pulse.mtw**) es aleatoria, y que la variable **Pulse1** es Normal. Por tanto, podemos utilizar este procedimiento estadístico para comprobar si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la varianza poblacional del pulso antes de correr es menor que 130 pulsaciones al cuadrado. Si σ^2 denota la varianza poblacional de la variable X=*Pulso antes de correr*, las hipótesis nula y alternativa son $H_0 : \sigma^2 \geq 130$ y $H_1 : \sigma^2 < 130$.

Seleccionamos, por tanto, la opción **Stat ⇒Basic Statistics ⇒1 Variance**. En el cuadro de diálogo resultante, arriba a la derecha, seleccionamos **Enter variance** (si quisiéramos realizar un test sobre la desviación típica poblacional, seleccionaríamos **Enter standard deviation**); en **Samples in columns** se selecciona, de la lista de variables de la izquierda, la columna o columnas para las cuales se va a realizar este tipo de test; en nuestro caso se selecciona **'Pulse1'**. Dejamos desactivada la opción **Summarized data**. Activamos **Perform hypothesis test** y en **Hypothesized variance** se especifica el valor, σ_0^2, con el que se compara la varianza poblacional, que es **130**. Si pulsamos el botón **Options** nos aparece un nuevo cuadro de diálogo con las siguientes opciones:

Confidence level: Por defecto se muestra un intervalo de confianza al 95 % para la varianza poblacional σ^2. Se puede introducir un valor entre 1 y 99 para solicitar otro nivel de confianza. En nuestro caso, podemos dejar lo que aparece por defecto, es decir, 95.

Alternative: Aquí se especifica cuál es la hipótesis alternativa: **less than** significa que la hipótesis alternativa es $H_1 : \sigma^2 < \sigma_0^2$, **not equal** significa que la hipótesis alternativa es $H_1 : \sigma^2 \neq \sigma_0^2$ y **greater than** significa que la hipótesis alternativa es $H_1 : \sigma^2 > \sigma_0^2$. Tengamos en cuenta que con la opción **less than** el intervalo de confianza para la varianza será del tipo $(-\infty, b)$, con la opción **not equal** el intervalo de confianza para la varianza será del tipo (a, b) y con la opción **greater than** el intervalo de confianza para la varianza será del tipo $(a, +\infty)$. En nuestro caso, tenemos que seleccionar **less than** ya que la hipótesis alternativa es $H_1 : \sigma^2 < 130$.

Podemos comprobar, en la ventana de sesión, que el p-valor (para el método Standard) es $0'338$, claramente mayor que el nivel de significación, $\alpha = 0'05$. En consecuencia, aceptamos la hipótesis nula y, por tanto, no podemos aceptar la hipótesis alternativa; es decir, no podemos aceptar que la varianza poblacional del pulso antes de correr es menor que 130 pulsaciones al cuadrado. El intervalo de confianza al 95 % para la varianza poblacional, asociado a este test de hipótesis (con el método Standard), es $(-\infty, 158)$. El intervalo de confianza al 95 % para la desviación típica poblacional, asociado a este test de hipótesis (con el método Standard), es $(-\infty, 12'6)$.

También se puede realizar este test de hipótesis **si sabemos el tamaño muestral y el resultado de la cuasi-varianza muestral**. Veámoslo con un ejemplo:

Se sabe que las calificaciones en la asignatura A es una variable Normal de media y varianza desconocidas. Se extrae una muestra aleatoria simple de 81 alumnos de la asignatura A, obteniéndose una media de $6'8$ puntos, con una cuasi-varianza de $1'69$ puntos al cuadrado, en las calificaciones de dichos alumnos. Sabemos que la varianza de las calificaciones en otra asignatura B es de $2'6$ puntos al

cuadrado. Queremos saber si la verdadera varianza de las calificaciones en la asignatura A es menor que la varianza en las calificaciones en la asignatura B.

Sea la variable aleatoria X=*Calificaciones en la asignatura* A. Como siempre, denotamos la varianza poblacional de X por σ^2. Así pues, las hipótesis nula y alternativa son:

$$H_0 : \sigma^2 \geq 2'6\,,$$
$$H_1 : \sigma^2 < 2'6\,.$$

Seleccionamos la opción **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**1 Variance**. En el cuadro de diálogo resultante, arriba a la derecha, seleccionamos **Enter variance**. Activamos la opción **Summarized data**, con lo cual se desactiva automáticamente la opción **Samples in columns**. En **Sample size** tenemos que teclear el tamaño muestral, que es **81**, y en **Sample variance** tenemos que teclear el resultado de la cuasi-varianza muestral, que es **1,69**. Activamos **Perform hypothesis test** y en **Hypothesized variance** se especifica el valor, σ_0^2, con el que se compara la varianza poblacional, que es **2,6**. Pulsamos en **Options** y, en el cuadro de diálogo resultante, en **Alternative** seleccionamos **less than** puesto que nuestra hipótesis alternativa es $H_1 : \sigma^2 < 2'6$. Como ya sabemos, dentro de este cuadro de diálogo también podemos cambiar el nivel de confianza del intervalo de confianza para la varianza poblacional; por defecto, este nivel de confianza es 95 %; si queremos cambiarlo tenemos que modificar el valor de **Confidence level**. Nosotros dejaremos lo que está puesto por defecto: 95 %.

Podemos comprobar, en la ventana de sesión, que el p-valor es $0'006$, claramente menor que los niveles de significación usuales ($\alpha = 0'05$ ó $\alpha = 0'01$). En consecuencia, rechazamos la hipótesis nula y, por tanto, aceptamos que la varianza de las calificaciones en la asignatura A es menor que la varianza de las calificaciones en la asignatura B. El intervalo de confianza al 95 % para la varianza poblacional, asociado a este test de hipótesis, es $(-\infty, 2'24)$.

6.2. Ejemplos que se van a resolver en clase

Ejemplo 6.1. Volvemos a considerar los datos del Ejemplo 5.1: En la tabla siguiente aparecen los datos de 10 bibliotecas, en las cuales se ha observado las siguientes variables: número total de títulos catalogados en un año (X), número de horas totales al año que emplea la biblioteca en catalogar sus títulos (Y) y costo, en euros, de una hora de catalogación (Z).

x_i	y_i	z_i
1550	220	15'75
1640	230	14'50
1000	140	16'40
950	135	16'70
750	110	17'10
1700	255	12'50
1650	228	14'80
1860	270	15'25
1900	280	18'50
900	130	17'30

a) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el número medio poblacional de títulos catalogados en un año es igual a 1400? ¿Por qué? ¿Cuál es el intervalo de confianza al 95 % para el número medio poblacional de títulos catalogados en un año?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el número medio poblacional de horas totales al año que emplea la biblioteca en catalogar sus títulos es mayor que 190? ¿Por qué?

c) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la media poblacional del costo de una hora de catalogación es menor que 16 euros? ¿Por qué?

d) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la varianza poblacional del número total de títulos catalogados en un año es mayor que 191000? ¿Por qué?

e) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la desviación típica poblacional del número de horas totales al año que emplea la biblioteca en catalogar sus títulos es menor que 66? ¿Por qué?

f) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la desviación típica poblacional del costo de una hora de catalogación es igual a $1'7$ euros? ¿Por qué? ¿Cuál es el intervalo de confianza al 95 % para la desviación típica poblacional del costo de una hora de catalogación?

6.3. Actividades de aplicación de los contenidos

6.3.1. Problemas propuestos

Problema 6.1. Utilizamos los datos del **Problema** 5.1.

a) Crea un nuevo proyecto de ***Minitab***. Abre la hoja de datos **Tiempo-Consulta.mtw**. ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el tiempo medio poblacional empleado en consultar bases de datos por ordenador es igual a 17 minutos? ¿Por qué? ¿Cuál es el intervalo de confianza al 95 % para el tiempo medio poblacional empleado en consultar bases de datos por ordenador?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la varianza poblacional del tiempo empleado en consultar bases de datos por ordenador es igual a 8 minutos^2? ¿Por qué? ¿Cuál es el intervalo de confianza al 95 % para la varianza poblacional del tiempo empleado en consultar bases de datos por ordenador?

c) Si quieres, puedes grabar el proyecto de ***Minitab***.

Problema 6.2. Utilizamos los datos del **Problema** 5.2.

a) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la media poblacional de la edad de las personas que visitan la biblioteca es mayor que 36 años? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la varianza poblacional de la edad de las personas que visitan la biblioteca es menor que 550 años^2? ¿Por qué?

Problema 6.3. Utilizamos los datos del **Problema** 5.3.

a) Crea un nuevo proyecto de ***Minitab***. Abre la hoja de datos **Materias-Buscadas-Localizadas.mtw**. ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'01$, que la media poblacional del número mensual de materias buscadas por los usuarios de una biblioteca es igual a 45? ¿Por qué? ¿Cuál es el intervalo de confianza al 99 % para la media poblacional del número mensual de materias buscadas por los usuarios de una biblioteca?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'01$, que la media poblacional del número mensual de materias localizadas por los usuarios de dicha biblioteca es mayor que 24? ¿Por qué?

c) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'01$, que la desviación típica poblacional del número mensual de materias buscadas por los usuarios de una biblioteca es menor que 14? ¿Por qué?

d) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'01$, que la desviación típica poblacional del número mensual de materias localizadas por los usuarios de dicha biblioteca es igual a 8? ¿Por qué? ¿Cuál es el intervalo de confianza al 99 % para la desviación típica poblacional del número mensual de materias localizadas por los usuarios de dicha biblioteca?

e) Si quieres, puedes grabar el proyecto de ***Minitab***.

Problema 6.4. Utilizamos los datos del **Problema** 5.4.

a) Crea un nuevo proyecto de ***Minitab***. Abre la hoja de datos **Peso-Precio-Libros.mtw**. ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la media poblacional del peso de los libros es menor que 582 gramos? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la media poblacional del precio de los libros es menor que igual a 56 euros? ¿Por qué? ¿Cuál es el intervalo de confianza al 95 % para la media poblacional del precio de los libros?

c) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la desviación típica poblacional del peso de los libros es mayor que 205 gramos? ¿Por qué?

d) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la varianza poblacional del precio de los libros es igual a 725 euros^2? ¿Por qué? ¿Cuál es el intervalo de confianza al 95 % para la varianza poblacional del precio de los libros?

e) Si quieres, puedes grabar el proyecto de ***Minitab***.

Problema 6.5. El número medio recomendado de usuarios servidos semanalmente por cada miembro del personal de una biblioteca es de 100. En una muestra aleatoria de 81 miembros del personal de las bibliotecas de una determinada región se obtiene una media de $132'88$ usuarios servidos semanalmente, con una cuasi-desviación típica de $55'19$. ¿Las bibliotecas de dicha región siguen la recomendación mencionada? ¿Por qué?

Problema 6.6. El precio medio de los libros en rústica es de $63'4$ euros, con una desviación típica de $14'8$ euros. Una muestra aleatoria simple de 61 libros en rústica con ilustraciones en color tiene un precio medio de $69'5$ euros, con una cuasi-desviación típica de $16'6$ euros.

a) ¿Permiten los datos afirmar que los libros en rústica con ilustraciones en color son más caros que el resto de libros en rústica? ¿Por qué?

b) ¿La varianza del precio de los libros en rústica con ilustraciones en color es mayor que la del precio de los libros en rústica? ¿Por qué?

Problema 6.7. Se sabe que el número medio de veces que un artículo científico es citado durante los 5 siguientes años a su publicación es de $6'5$. Se eligen aleatoria e independientemente 71 artículos de medicina, obteniéndose una media de $7'8$ citas durante los 5 siguientes años a su publicación, con una cuasi-desviación típica de $2'3$. ¿Se puede afirmar que durante los 5 siguientes años a su publicación se citan más los artículos de medicina que el resto de artículos científicos? ¿Por qué?

6.3.2. Soluciones de los problemas propuestos

Solución del problema 6.1.

X=Tiempo (en minutos) empleado en consultar bases de datos por ordenador.

La media poblacional y la varianza poblacional de la variable aleatoria X se denotan, respectivamente, por μ y por σ^2.

a) La pregunta que se nos hace es: ¿$\mu = 17$?

Hipótesis nula y alternativa:

$$H_0 : \mu = 17$$
$$H_1 : \mu \neq 17$$

Condiciones:

- En el apartado (a) del Problema 5.1. hemos comprobado que la muestra de datos de la variable X es aleatoria.
- En el apartado (b) del Problema 5.1. hemos probado que la variable X es Normal.
- Obviamente, la desviación típica poblacional, σ, es desconocida.

Resolución con *Minitab*:

Como σ es desconocida, hay que seleccionar **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**1-Sample t**. En **Hypothesized mean** hay que teclear **17**. No olvidemos que en **Options** hay que dejar activada la opción **not equal** en **Alternative** (pues $H_1 : \mu \neq 17$).

El p-valor es $0'859$; mayor que el nivel de significación ($\alpha = 0'05$); por tanto, aceptamos H_0; es decir, aceptamos que el tiempo medio poblacional empleado en consultar bases de datos por ordenador es igual a 17 minutos.

El intervalo de confianza al 95 % para el tiempo medio poblacional empleado en consultar bases de datos por ordenador es $(15'288, 18'445)$.

b) La pregunta que se nos hace es: ¿$\sigma^2 = 8$?

Hipótesis nula y alternativa:

$$H_0 : \sigma^2 = 8$$
$$H_1 : \sigma^2 \neq 8$$

Condiciones:

- En el apartado (a) del Problema 5.1. hemos comprobado que la muestra de datos de la variable X es aleatoria.
- En el apartado (b) del Problema 5.1. hemos probado que la variable X es Normal.
- Aunque en el apartado anterior hayamos aceptado que $H_0 : \mu = 17$, esto no quiere decir que la media poblacional sea conocida. Lo que quiere decir es que no existe diferencia significativa entre la media poblacional y 17; es decir, el valor de μ está muy próximo a 17. Pero la media poblacional, μ, sigue siendo desconocida.

Resolución con *Minitab*:

Como μ es desconocida, hay que seleccionar **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**1-Variance**. Hay que elegir **Enter variance**. En **Hypothesized variance** hay que teclear **8**. No olvidemos que en **Options** hay que dejar activada la opción **not equal** en **Alternative** (pues $H_1 : \sigma^2 \neq 8$).

Como la variable X es Normal podemos utilizar el método estándar de ***Minitab*** (*Standard Method*). El p-valor para este método es $0'867$; mayor que el nivel de significación ($\alpha = 0'05$); por tanto, aceptamos H_0; es decir, aceptamos que la varianza poblacional del tiempo empleado en consultar bases de datos por ordenador es igual a 8 minutos.

El intervalo de confianza al 95 % para la varianza poblacional del tiempo empleado en consultar bases de datos por ordenador es $(4'35, 20'21)$.

Solución del problema 6.2.

X=*Edad de las personas que visitan la biblioteca.*

a) La pregunta que se nos hace es: ¿$\mu > 36$?

En el apartado (a) del Problema 5.2. hemos comprobado que la muestra de datos de la variable X es aleatoria.

En el apartado (b) del Problema 5.2. hemos comprobado que la variable X no es Normal. Además, el tamaño muestral no es mayor o igual que 30. Por Tanto, no podemos realizar el test de hipótesis sobre la media poblacional. En consecuencia, no podemos responder a la pregunta.

b) La pregunta que se nos hace es: ¿$\sigma^2 < 550$?

Como la variable X no es Normal y, además, el tamaño muestral no es mayor o igual que 30, entonces no podemos realizar el test de hipótesis sobre la varianza poblacional. En consecuencia, no podemos responder a la pregunta.

Solución del problema 6.3.

X=*Número mensual de materias buscadas por los usuarios de una biblioteca.*

Y=*Número mensual de materias localizadas por los usuarios de dicha biblioteca.*

La media poblacional y la desviación típica poblacional de la variable aleatoria X se denotan, respectivamente, por μ_x y por σ_x.

Análogamente, la media poblacional y la desviación típica poblacional de la variable aleatoria Y se denotan, respectivamente, por μ_y y por σ_y.

a) La pregunta que se nos hace es: ¿$\mu_x = 45$? Por tanto, las hipótesis nula y alternativa son:

$$H_0 : \mu_x = 45$$
$$H_1 : \mu_x \neq 45$$

Tras comprobar que se cumplen las condiciones para aplicar este test, lo resolvemos mediante ***Minitab***. Como el p-valor es $0'543$ (mayor que el nivel de significación) entonces aceptamos que la media poblacional del número mensual de materias buscadas por los usuarios de una biblioteca es igual a 45. El intervalo de confianza al 99 % para la media poblacional del número mensual de materias buscadas por los usuarios de una biblioteca es $(36'13, 58'37)$.

b) La pregunta que se nos hace es: ¿$\mu_y > 24$? Por tanto, las hipótesis nula y alternativa son:

$$H_0 : \mu_y \leq 24$$
$$H_1 : \mu_y > 24$$

Tras comprobar que se cumplen las condiciones para aplicar este test, lo resolvemos mediante ***Minitab***. Como el p-valor es $0'157$ (mayor que el nivel de significación) entonces **no** podemos aceptar que la media poblacional del número mensual de materias localizadas por los usuarios de dicha biblioteca es mayor que 24.

c) La pregunta que se nos hace es: ¿$\sigma_x < 14$? Por tanto, las hipótesis nula y alternativa son:

$$H_0 : \sigma_x \geq 14$$
$$H_1 : \sigma_x < 14$$

Tras comprobar que se cumplen las condiciones para aplicar este test, lo resolvemos mediante ***Minitab***. Como el p-valor es $0'344$ (mayor que el nivel de significación) entonces **no** podemos aceptar que la desviación típica poblacional del número mensual de materias buscadas por los usuarios de una biblioteca es menor que 14.

d) La pregunta que se nos hace es: ¿$\sigma_y = 8$? Por tanto, las hipótesis nula y alternativa son:

$$H_0 : \sigma_y = 8$$
$$H_1 : \sigma_y \neq 8$$

Tras comprobar que se cumplen las condiciones para aplicar este test, lo resolvemos mediante ***Minitab***. Como el p-valor es $0'958$ (mayor que el nivel de significación) entonces aceptamos que la desviación típica poblacional del número mensual de materias localizadas por los usuarios de dicha biblioteca es igual a 8. El intervalo de confianza al 99 % para la desviación típica poblacional del número mensual de materias localizadas por los usuarios de dicha biblioteca es $(4'92, 15'76)$.

Solución del problema 6.4.

X=Peso, en gramos, de los libros.

Y=Precio, en euros, de los libros.

La media poblacional y la desviación típica poblacional de la variable aleatoria X se denotan, respectivamente, por μ_x y por σ_x.

Análogamente, la media poblacional y la desviación típica poblacional de la variable aleatoria Y se denotan, respectivamente, por μ_y y por σ_y.

a) La pregunta que se nos hace es: ¿$\mu_x < 582$? Por tanto, las hipótesis nula y alternativa son:

$$H_0 : \mu_x \geq 582$$
$$H_1 : \mu_x < 582$$

Tras comprobar que se cumplen las condiciones para aplicar este test, lo resolvemos mediante ***Minitab***. Como el p-valor es $0'484$ (mayor que el nivel de significación) entonces **no** podemos aceptar que la media poblacional del peso de los libros es menor que 582 gramos.

b) La pregunta que se nos hace es: ¿$\mu_y = 56$? Por tanto, las hipótesis nula y alternativa son:

$$H_0 : \mu_y = 56$$
$$H_1 : \mu_y \neq 56$$

Tras comprobar que se cumplen las condiciones para aplicar este test, lo resolvemos mediante ***Minitab***. Como el p-valor es $0'925$ (mayor que el nivel de significación) entonces aceptamos que la media poblacional del precio de los libros es menor que igual a 56 euros. El intervalo de confianza al 95 % para la media poblacional del precio de los libros es $(38'16, 72'34)$.

c) La pregunta que se nos hace es: ¿$\sigma_x > 205$? Por tanto, las hipótesis nula y alternativa son:

$$H_0 : \sigma_x \leq 205$$
$$H_1 : \sigma_x > 205$$

Tras comprobar que se cumplen las condiciones para aplicar este test, lo resolvemos mediante ***Minitab***. Como el p-valor es $0'4$ (mayor que el nivel de significación) entonces **no** podemos aceptar que la desviación típica poblacional del peso de los libros es mayor que 205 gramos.

d) La pregunta que se nos hace es: ¿$\sigma_y^2 = 725$? Por tanto, las hipótesis nula y alternativa son:

$$H_0 : \sigma_y^2 = 725$$
$$H_1 : \sigma_y^2 \neq 725$$

Tras comprobar que se cumplen las condiciones para aplicar este test, lo resolvemos mediante ***Minitab***. Como el p-valor es $0'89$ (mayor que el nivel de significación) entonces aceptamos que la varianza poblacional del precio de los libros es igual a 725 euros^2. El intervalo de confianza al 95 % para la varianza poblacional del precio de los libros es $(363, 2086)$.

Solución del problema 6.5. Sea X=*Número de usuarios servidos semanalmente por cada miembro del personal de la biblioteca*. Hacemos un test sobre μ con σ desconocida. Comprobamos que

se cumplen las condiciones para poder aplicar este test. La hipótesis nula es $H_0 : \mu = 100$. En ***Minitab*** tenemos que activar **Summarized data** (por tanto, se nos desactivará **Samples in columns**) y rellenar las tres medidas descriptivas (muestrales) que nos solicitan. El p-valor es igual a cero; menor que cualquier nivel de significación. En consecuencia, rechazamos H_0 y, por tanto, las bibliotecas de dicha región **no** siguen la recomendación.

Solución del problema 6.6. Sea X=*Precio de los libros en rústica con ilustraciones color*.

a) Hacemos un test sobre μ con σ desconocida. Comprobamos que se cumplen las condiciones para poder aplicar este test. La hipótesis nula es $H_0 : \mu \leq 63'4$. En ***Minitab*** tenemos que activar **Summarized data**. El p-valor es $0'003$; menor que el habitual nivel de significación ($0'05$); en consecuencia, rechazamos H_0 y, por tanto, los libros en rústica con ilustraciones en color son más caros (tienen un precio medio mayor) que el resto de los libros en rústica.

b) Hacemos un test sobre σ^2 con μ desconocida. Comprobamos que se cumplen las condiciones para poder aplicar este test. La hipótesis nula es $H_0 : \sigma^2 \leq (14'8)^2$. En ***Minitab*** tenemos que activar **Summarized data**. El p-valor es $0'086$; mayor que el habitual nivel de significación ($0'05$); en consecuencia, aceptamos H_0 y, por tanto, **no** se puede aceptar que la varianza del precio de los libros en rústica con ilustraciones en color sea mayor que la varianza del precio de todos los libros en rústica.

Solución del problema 6.7. Sea X=*Número de veces que los artículos de medicina son citados durante los cinco siguientes años a su publicación*. Hacemos un test sobre μ con σ desconocida. Comprobamos que se cumplen las condiciones para poder aplicar este test. La hipótesis nula es $H_0 : \mu \leq 6'5$. En ***Minitab*** tenemos que activar **Summarized data**. El p-valor es igual a cero; menor que cualquier nivel de significación. En consecuencia, rechazamos H_0 y, por tanto, se citan más los artículos de medicina que el resto de artículos científicos (la media del número de citas de los artículos de medicina es mayor que la del resto de artículos científicos).

7

Estimación y tests paramétricos en dos poblaciones

7.1. Desarrollo de los contenidos fundamentales (teoría y PRÁCTICA 6)

En todo este tema trataremos con una variable aleatoria X observada en dos poblaciones distintas, que podemos llamar *población 1* y *población 2*. Denotaremos por X_1 a la variable aleatoria X observada en la *población 1* y por X_2 a la variable aleatoria X observada en la *población 2*. Como es habitual, denotaremos por μ_1 y por σ_1^2 a la media poblacional y a la varianza poblacional, respectivamente, de la variable X_1. Análogamente, denotaremos por μ_2 y por σ_2^2 a la media poblacional y a la varianza poblacional, respectivamente, de la variable X_2. Tendremos dos muestras aleatorias (una de cada población) de tamaños n_1 y n_2, respectivamente. Las medias muestrales se denotarán por $\overline{X}_1$ y $\overline{X}_2$, respectivamente; y las cuasi-varianzas muestrales se denotarán por S_1^2 y S_2^2, respectivamente.

7.1.1. Comparación de dos varianzas poblacionales con muestras independientes y medias poblacionales desconocidas

7.1.1.1. Introducción

En un apartado posterior vamos a estudiar el problema de la comparación de μ_1 con μ_2 en el caso en que las dos muestras sean independientes. Veremos en dicho apartado que necesitamos saber si σ_1^2 y σ_2^2 (que serán desconocidas) son iguales o distintas. Por este motivo estudiamos ahora el test de comparación de varianzas en el caso en que μ_1 y μ_2 sean desconocidas.

El test que vamos a explicar está basado en lo siguiente:

En las condiciones generales que se han dado al principio del tema, si las dos muestras son independientes y las dos variables, X_1 y X_2, son Normales entonces consideramos el estadístico:

$$F = \frac{S_1^2}{S_2^2}$$

Si es cierta la hipótesis nula $H_0 : \sigma_1^2 = \sigma_2^2$ entonces se puede demostrar que el estadístico F sigue una distribución F de Snedecor con $n_1 - 1$ grados de libertad en el numerador y $n_2 - 1$ grados de libertad en el denominador (es decir, $F \equiv F_{n_1-1\,,\,n_2-1}$) por lo que podemos utilizar esta distribución para conocer las probabilidades asociadas a los diferentes valores del estadístico F. Precisamente el conocimiento de esas probabilidades es el que nos permite tomar decisiones respecto al parámetro σ_1^2/σ_2^2 y, por tanto, respecto de la igualdad o diferencia de las varianzas poblacionales.

Por esta razón, a este test se le denomina **test F de Snedecor sobre comparación de dos varianzas**.

7.1.1.2. Hipótesis nula y alternativa del test

Hay tres posibilidades:

$H_0 : \sigma_1^2 = \sigma_2^2$	$H_0 : \sigma_1^2 \geq \sigma_2^2$	$H_0 : \sigma_1^2 \leq \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$H_1 : \sigma_1^2 < \sigma_2^2$	$H_1 : \sigma_1^2 > \sigma_2^2$

pero ***Minitab*** solamente resuelve la primera de ellas, pues es la que realmente se necesita para realizar, posteriormente, el test de comparación de dos medias poblacionales con muestras independientes.

7.1.1.3. Condiciones para poder realizar el test

Para realizar cualquiera de los tres tipos de tests anteriores, es necesario que se verifiquen las condiciones siguientes:

1. Las dos muestras son aleatorias.
2. Las dos muestras son independientes entre sí.
3. Las dos variables aleatorias, X_1 y X_2, son Normales.
4. Las dos medias, μ_1 y μ_2, son desconocidas.

Si se cumplen todas las condiciones menos la tercera (es decir, las variables no son Normales) se pueden aplicar otros tests de hipótesis, como, por ejemplo, el test de Levene.

7.1.1.4. Resolución mediante MINITAB

Para realizar el test de comparación de dos varianzas poblacionales con muestras independientes y medias poblacionales desconocidas hay que seleccionar **Stat** ⇒**Basic Statistics** ⇒**2 Variances**.

Ejemplo A. Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la varianza poblacional del pulso de los hombres antes de correr es

igual a la varianza poblacional del pulso de las mujeres antes de correr. Lo que se quiere es comparar la varianza poblacional de la variable **Pulse1** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). Las hipótesis nula y alternativa son $H_0 : \sigma_1^2 = \sigma_2^2$ y $H_1 : \sigma_1^2 \neq \sigma_2^2$, siendo X_1=*"Pulso de los hombres antes de correr"* y X_2=*"Pulso de las mujeres antes de correr"*. Como no hay relación alguna entre el grupo de hombres y el grupo de mujeres, podemos afirmar que las muestras son independientes. Por tanto, nos encontramos ante un test de comparación de dos varianzas poblacionales, con muestras independientes y medias poblacionales desconocidas. Ya comprobamos (en el tema anterior al anterior a éste) que las dos muestras son aleatorias y que las dos variables, X_1 y X_2, son Normales.

Para hacer este test seleccionamos **Stat** ⇒**Basic Statistics** ⇒**2 Variances**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different columns** y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna **'Pulse1'**; en **Subscripts** seleccionamos, de la lista de la izquierda, la columna **'Sex'**. Si pulsamos el botón **Options** nos aparece un nuevo cuadro de diálogo con las siguientes opciones:

Confidence level: Por defecto se muestra un intervalo de confianza al 95 % para la diferencia de desviaciones típicas poblacionales, $\sigma_1 - \sigma_2$. Se puede introducir un valor entre 1 y 99 para solicitar otro nivel de confianza. En nuestro ejemplo, podemos dejar lo que aparece por defecto, es decir, 95.

Title: Aquí se puede escribir un título para el resultado del test. En nuestro ejemplo, podemos dejarlo en blanco.

Como resultado de este test obtenemos una nueva ventana que contiene dos gráficos y los resultados de dos tests de hipótesis sobre comparación de dos varianzas (el test F de Snedecor y el test de Levene). El test F de Snedecor es el que hemos explicado. **El test de Levene se utiliza cuando las variables no son Normales**.

Podemos comprobar que el p-valor para el test F de Snedecor es $0'299$; claramente mayor que el nivel de significación, $\alpha = 0'05$, por lo que podemos aceptar la hipótesis nula; es decir, podemos aceptar que la varianza poblacional del pulso de los hombres antes de correr es igual a la varianza poblacional del pulso de las mujeres antes de correr. Con el test de Levene también aceptaríamos la hipótesis nula pues el p-valor es igual a $0'148$.

Ejemplo B. Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la varianza poblacional del pulso de los hombres después de correr es igual a la varianza poblacional del pulso de las mujeres después de correr. Lo que se quiere es comparar la varianza poblacional de la variable **Pulse2** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). Las hipótesis nula y alternativa son $H_0 : \sigma_1^2 = \sigma_2^2$ y $H_1 : \sigma_1^2 \neq \sigma_2^2$, siendo X_1=*"Pulso de los hombres después de correr"* y X_2=*"Pulso de las mujeres después de correr"*.

Para hacer este test seleccionamos **Stat** ⇒**Basic Statistics** ⇒**2 Variances**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different columns** y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna **'Pulse2'**; en **Subscripts** seleccionamos, de la lista de la izquierda, la columna **'Sex'**.

En el tema anterior al anterior a éste hemos comprobado que la variable **Pulse2** no es Normal. Por tanto, vamos a utilizar el test de Levene en lugar del test F de Snedecor.

El p-valor para el test de Levene es $0'011$, menor que el nivel de significación, $\alpha = 0'05$, por lo que tenemos que rechazar la hipótesis nula y, por tanto, aceptar que la varianza poblacional del pulso de los hombres después de correr es distinta de la varianza poblacional del pulso de las mujeres después de correr.

También se puede realizar este test de hipótesis **si sabemos los dos tamaños muestrales y las dos cuasi-varianzas muestrales**. Pero, en este caso, ***Minitab*** no realiza el test de Levene por lo que es necesario que las dos variables sean Normales. Veámoslo con un nuevo ejemplo:

Ejemplo C. Supongamos que, de una muestra aleatoria de 21 personas que son socias de una biblioteca, la media del número de horas por semana que pasan en la biblioteca es 10, con una cuasi-varianza de 9. Y para una muestra aleatoria independiente de la primera, de 16 personas que no son socias de la biblioteca, la media es 6, con una cuasi-varianza de 4. ¿Existe diferencia significativa entre las varianzas del número de horas semanales que pasan en la biblioteca los socios y los no socios?

Sean X_1=*"Tiempo semanal que permanecen en la biblioteca los socios"* y X_2=*"Tiempo semanal que permanecen en la biblioteca los no socios"*. Hemos de suponer que las variables aleatorias X_1 y X_2 son Normales.

Así pues, se tienen los siguientes datos:

$$n_1 = 21\,, \quad S_1^2 = 9\,,$$
$$n_2 = 16\,, \quad S_2^2 = 4\,.$$

Las hipótesis nula y alternativa son:

$$H_0 : \sigma_1^2 = \sigma_2^2\,,$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2\,.$$

Seleccionamos la opción **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**2 Variances**. Activamos la opción **Summarized data**, con lo cual se desactivan automáticamente las opciones **Samples in one column** y **Samples in different columns**. Dentro de **First**, en **Sample size** tenemos que teclear el tamaño muestral de la primera muestra, que es **21**, y en **Variance** tenemos que teclear el resultado de la cuasi-varianza de la primera muestra, que es **9**. Dentro de **Second**, en **Sample size** tenemos que teclear el tamaño muestral de la segunda muestra, que es **16**, y en **Variance** tenemos que teclear el resultado de la cuasi-varianza de la segunda muestra, que es **4**.

Tanto en la ventana de sesión como en el gráfico generado comprobamos que el p-valor para el test F de Snedecor es $0'114$, mayor que los niveles de significación usuales ($\alpha = 0'05$ ó $\alpha = 0'01$) y, por tanto, aceptamos la hipótesis nula. En consecuencia, aceptamos que no existe diferencia significativa entre las varianzas del número de horas semanales que pasan en la biblioteca los socios y los no socios.

7.1.2. Comparación de dos medias poblacionales. Intervalo de confianza para la diferencia de dos medias

7.1.2.1. Comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas pero iguales

7.1.2.1.1. Introducción

En general, un test para decidir sobre la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la hipótesis alternativa $H_1 : \mu_1 \neq \mu_2$ es muy frecuente y constituye uno de los primeros objetivos de cualquier investigador que se inicia en Estadística.

El test que vamos a explicar está basado en lo siguiente:

En las condiciones generales que se han dado al principio del tema, si las dos variables, X_1 y X_2, son Normales; las dos muestras son independientes y las dos varianzas poblacionales son desconocidas pero iguales ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), entonces se verifica que el estadístico:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

sigue una distribución t de Student con $n_1 + n_2 - 2$ grados de libertad; es decir, $T \equiv t_{n_1+n_2-2}$. Si las variables aleatorias no son Normales, pero se verifica que los tamaños muestrales son grandes (en la práctica, $n_1 \geq 30$, $n_2 \geq 30$), entonces el estadístico T se aproxima a una variable t de Student con $n_1 + n_2 - 2$ grados de libertad.

Por esta razón, a este test se le denomina **test t de Student sobre comparación de dos medias con muestras independientes**.

7.1.2.1.2. Hipótesis nula y alternativa del test

Hay tres posibilidades:

$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 \geq \mu_2$	$H_0 : \mu_1 \leq \mu_2$
$H_1 : \mu_1 \neq \mu_2$	$H_1 : \mu_1 < \mu_2$	$H_1 : \mu_1 > \mu_2$

7.1.2.1.3. Condiciones para poder realizar el test

Para realizar cualquiera de los tres tipos de tests anteriores, es necesario que se verifiquen las condiciones siguientes:

- Las dos muestras son aleatorias.
- Las dos muestras son independientes entre sí.
- Las dos variables aleatorias, X_1 y X_2, son Normales o los dos tamaños muestrales son grandes ($n_1, n_2 \geq 30$).
- Las dos varianzas poblacionales son desconocidas e iguales ($\sigma_1^2 = \sigma_2^2$).

7.1.2.1.4. Resolución mediante MINITAB

Para realizar el test de comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas pero iguales hay que seleccionar **Stat** ⇒**Basic Statistics**⇒**2-Sample t**.

Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional de los hombres antes de correr es igual al pulso medio poblacional de las mujeres antes de correr. Lo que se quiere es comparar la media poblacional de la variable **Pulse1** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). Las hipótesis nula y alternativa son $H_0 : \mu_1 = \mu_2$ y $H_1 : \mu_1 \neq \mu_2$, siendo X_1=*"Pulso de los hombres antes de correr"* y X_2=*"Pulso de las mujeres antes de correr"*.

En el **Ejemplo A** de la sección 7.1.1.4 hemos comprobado que la varianza poblacional del pulso de los hombres antes de correr es igual a la varianza poblacional del pulso de las mujeres antes de correr. Por tanto, nos encontramos ante un test de comparación de dos medias poblacionales, con muestras independientes y varianzas poblacionales desconocidas pero iguales. Aunque las variables aleatorias X_1 y X_2 no fuesen Normales (que sí lo son, pues lo hemos comprobado en el tema anterior al anterior a éste), se puede aplicar este test debido a que los tamaños muestrales son suficientemente grandes: $n_1 = 57$ y $n_2 = 35$.

Para hacer este test seleccionamos **Stat** ⇒**Basic Statistics** ⇒**2-Sample t**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different columns** y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna '**Pulse1**'; en **Subscripts** seleccionamos, de la lista de la izquierda, la columna '**Sex**'; y activamos **Assume equal variances** ya que hemos comprobado que las varianzas poblacionales son desconocidas pero iguales. Si pulsamos el botón **Options** nos aparece un nuevo cuadro de diálogo con las siguientes opciones:

Confidence level: Por defecto se muestra un intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$. Se puede introducir un valor entre 1 y 99 para solicitar otro nivel de confianza. En nuestro ejemplo, podemos dejar lo que aparece por defecto, es decir, 95.

Test difference: Aquí se pone el valor con el que se compara la diferencia de medias poblacionales, μ_0. La hipótesis nula $H_0 : \mu_1 = \mu_2$ es equivalente a $H_0 : \mu_1 - \mu_2 = 0$, por lo que el valor con el que se compara la diferencia de medias poblacionales, en este ejemplo, es cero; es decir, $\mu_0 = 0$. En consecuencia, nosotros dejamos lo que aparece por defecto (cero).

Alternative: Aquí se especifica cuál es la hipótesis alternativa: **less than** significa que la hipótesis alternativa es $H_1 : \mu_1 - \mu_2 < \mu_0$, **not equal** significa que la hipótesis alternativa es $H_1 : \mu_1 - \mu_2 \neq \mu_0$ y **greater than** significa que la hipótesis alternativa es $H_1 : \mu_1 - \mu_2 > \mu_0$. Tengamos en cuenta que con la opción **less than** el intervalo de confianza para $\mu_1 - \mu_2$ será del tipo $(-\infty, b)$, con la opción **not equal** el intervalo de confianza será del tipo (a, b) y con la opción **greater than** el intervalo de confianza será del tipo $(a, +\infty)$. En nuestro ejemplo, tenemos que dejar lo que aparece por defecto, que es **not equal**, ya que la hipótesis alternativa es $H_1 : \mu_1 \neq \mu_2$, que es equivalente a $H_1 : \mu_1 - \mu_2 \neq 0$.

Podemos comprobar, en la ventana de sesión, que el p-valor es $0'006$, claramente menor que el nivel de significación, $\alpha = 0'05$, por lo que debemos rechazar la hipótesis nula y, por tanto, aceptar la hipótesis alternativa. Aceptamos que el pulso medio poblacional de los hombres antes de correr es distinto del pulso medio poblacional de las mujeres antes de correr. Como la media muestral del pulso de las mujeres antes de correr $(76'9)$ es mayor que la media muestral del pulso de los hombres antes de correr $(70'42)$ podríamos, incluso, aceptar que la media poblacional del pulso de las mujeres antes

de correr es mayor que la media poblacional del pulso de los hombres antes de correr. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$, es $(-10'96, -1'91)$.

También se puede realizar este test de hipótesis **si sabemos los dos tamaños muestrales, las dos medias muestrales y las dos cuasi-desviaciones típicas muestrales**. Veámoslo con un nuevo ejemplo:

Con los datos del **Ejemplo C** (de la sección 7.1.1.4) queremos decidir si existe diferencia significativa entre el número medio de horas semanales que permanecen en la biblioteca los socios y los no socios.

Como en dicho ejemplo hemos decidido aceptar que no existe diferencia significativa entre las varianzas poblacionales, entonces nos encontramos ante un test de comparación de dos medias poblacionales, con muestras independientes y varianzas poblacionales desconocidas pero iguales. Las hipótesis nula y alternativa son:

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

Los datos son:

$$n_1 = 21 \,, \quad \overline{X}_1 = 10 \,, \quad S_1 = 3 \,,$$
$$n_2 = 16 \,, \quad \overline{X}_2 = 6 \,, \quad S_2 = 2 \,.$$

Seleccionamos la opción **Stat** ⇒**Basic Statistics** ⇒**2-Sample t**. Activamos la opción **Summarized data**, con lo cual se desactivan automáticamente las opciones **Samples in one column** y **Samples in different columns**. Dentro de **First**, en **Sample size** tenemos que teclear el tamaño muestral de la primera muestra, que es **21**, en **Mean** tenemos que teclear el resultado de la media de la primera muestra, que es **10**, y en **Standard deviation** tenemos que teclear el resultado de la cuasi-desviación típica de la primera muestra, que es **3**. Dentro de **Second**, en **Sample size** tenemos que teclear el tamaño muestral de la segunda muestra, que es **16**, en **Mean** tenemos que teclear el resultado de la media de la segunda muestra, que es **6**, y en **Standard deviation** tenemos que teclear el resultado de la cuasi-desviación típica de la segunda muestra, que es **2**. Activamos **Assume equal variances** ya que hemos comprobado (en el **Ejemplo C**, como ya hemos dicho) que las varianzas poblacionales son desconocidas pero iguales. Pulsamos en **Options** y en el cuadro de diálogo resultante dejamos lo que aparece por defecto (**Confidence level: 95**, **Test difference: 0**, **Alternative: not equal**).

Podemos comprobar, en la ventana de sesión, que el p-valor es 0, el mínimo posible y, por supuesto, menor que los niveles de significación usuales ($\alpha = 0'05$ ó $\alpha = 0'01$), por lo que debemos rechazar la hipótesis nula. Aceptamos, en consecuencia, que existe diferencia significativa entre el número medio de horas semanales que permanecen en la biblioteca los socios y los no socios. Como la media muestral del número de horas semanales que permanecen en la biblioteca los socios (10) es mayor que la media muestral del número de horas semanales que permanecen en la biblioteca los no socios (6) podríamos, incluso, aceptar que la media poblacional del número de horas semanales que permanecen en la biblioteca los socios es mayor que la media poblacional del número de horas semanales que permanecen en la biblioteca los no socios. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$, es $(2'326, 5'674)$.

7.1.2.2. Comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas y distintas

7.1.2.2.1. Introducción

El test que vamos a explicar está basado en lo siguiente:

En las condiciones generales que se han dado al principio del tema, si las dos variables, X_1 y X_2, son Normales; las dos muestras son independientes y las dos varianzas poblacionales son desconocidas y distintas, consideramos el estadístico:

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} .$$

Este estadístico no sigue una distribución t de Student con $n_1 + n_2 - 2$ grados de libertad. Pero se trata de un problema poco importante, pues disponemos de algunos procedimientos que nos permiten conocer de forma aproximada la distribución muestral de T.

El matemático Welch propuso una aproximación que acapara las preferencias de muchos investigadores. En esta aproximación, T se concibe como una variable aleatoria distribuida según una t de Student, pero con un número desconocido de grados de libertad. La solución pasa por determinar los grados de libertad (g) que corresponden a la distribución mediante la expresión:

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

Luego redondeamos el valor de g para que no tenga ningún decimal, y le llamamos de la misma manera (g). Esto se hace necesario ya que g va a ser los grados de libertad de una distribución t de Student, y tiene que ser un número natural. Se obtienen así unos grados de libertad comprendidos entre un mínimo y un máximo conocidos: el mínimo es el valor más pequeño entre $n_1 - 1$ y $n_2 - 1$; el máximo es $n_1 + n_2 - 2$. Comparando el valor de T con los correspondientes percentiles de la distribución t de Student con g grados de libertad podemos tomar decisiones respecto a $\mu_1 - \mu_2$.

7.1.2.2.2. Hipótesis nula y alternativa del test

Hay tres posibilidades:

$H_0 : \mu_1 = \mu_2$	$H_0 : \mu_1 \geq \mu_2$	$H_0 : \mu_1 \leq \mu_2$
$H_1 : \mu_1 \neq \mu_2$	$H_1 : \mu_1 < \mu_2$	$H_1 : \mu_1 > \mu_2$

7.1.2.2.3. Condiciones para poder realizar el test

Para realizar cualquiera de los tres tipos de tests anteriores, es necesario que se verifiquen las condiciones siguientes:

- Las dos muestras son aleatorias.

- Las dos muestras son independientes entre sí.

- Las dos variables aleatorias, X_1 y X_2, son Normales o los dos tamaños muestrales son grandes ($n_1, n_2 \geq 30$).

- Las dos varianzas poblacionales, σ_1^2 y σ_2^2, son desconocidas y distintas.

7.1.2.2.4. Resolución mediante MINITAB

Para realizar el test de comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas y distintas hay que seleccionar, igual que antes, **Stat ⇒Basic Statistics ⇒2-Sample t**. Hay que rellenar el cuadro de diálogo de manera similar al apartado anterior, con la salvedad de que, en este caso, hay que desactivar la opción **Assume equal variances**.

Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional de los hombres después de correr es igual al pulso medio poblacional de las mujeres después de correr. Queremos comparar la media poblacional de la variable **Pulse2** para los grupos en los que la variable **Sex** vale **1** (Hombre) y **2** (Mujer). Las hipótesis nula y alternativa son $H_0 : \mu_1 = \mu_2$ y $H_1 : \mu_1 \neq \mu_2$, siendo X_1=*"Pulso de los hombres después de correr"* y X_2=*"Pulso de las mujeres después de correr"*.

En el **Ejemplo B** de la sección 7.1.1.4 hemos comprobado que se puede aceptar que la varianza poblacional del pulso de los hombres después de correr es distinta de la varianza poblacional del pulso de las mujeres después de correr. Por tanto, nos encontramos ante un test de comparación de dos medias poblacionales, con muestras independientes y varianzas poblacionales desconocidas y distintas. Aunque las variables aleatorias X_1 y X_2 no son Normales, se puede aplicar este test debido a que los tamaños muestrales son suficientemente grandes: $n_1 = 57$ y $n_2 = 35$.

Para hacer el test seleccionamos **Stat ⇒Basic Statistics ⇒2-Sample t**. Activamos la opción **Samples in one column**, con lo cual se desactivan automáticamente las opciones **Samples in different columns** y **Summarized data**. En **Samples** seleccionamos, de la lista de variables de la izquierda, la columna **'Pulse2'**; y en **Subscripts** seleccionamos, de la lista de la izquierda, la columna **'Sex'**. Si se pulsa el botón **Options** aparece un cuadro de diálogo similar al ejemplo anterior. En este cuadro de diálogo dejamos lo que aparece por defecto (**Confidence level: 95**, **Test difference: 0**, **Alternative: not equal**).

Podemos comprobar, en la ventana de sesión, que el p-valor es $0'007$, claramente menor que el nivel de significación, $\alpha = 0'05$, por lo que debemos rechazar la hipótesis nula y, por tanto, aceptar la hipótesis alternativa. Aceptamos que el pulso medio poblacional de los hombres después de correr es distinto del pulso medio poblacional de las mujeres después de correr. Como la media muestral del pulso de las mujeres después de correr ($86'7$) es mayor que la media muestral del pulso de los hombres después de correr ($75'9$) podríamos, incluso, aceptar que la media poblacional del pulso de las mujeres después de correr es mayor que la media poblacional del pulso de los hombres después de correr. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, $\mu_1 - \mu_2$, es $(-18'65, -3'02)$.

7.1.2.3. Comparación de dos medias con muestras dependientes

7.1.2.3.1. Introducción

Una ocasión en que tenemos muestras apareadas es cuando un grupo de individuos es evaluado dos veces. Por ejemplo, si queremos comparar el tiempo medio que pasan en la biblioteca un grupo de alumnos antes de los exámenes y después de los exámenes, podemos evaluar el tiempo que pasa cada alumno en la biblioteca antes y después de los exámenes, y comparar las medias obtenidas; así tendremos dos muestras apareadas (dependientes) porque ambas pertenecen a los mismos individuos.

Pero esta no es la única forma que tenemos para generar muestras apareadas. También tenemos muestras apareadas cuando, en lugar de medir a los mismos individuos en dos ocasiones, utilizamos *pares* de individuos. Por ejemplo, podría interesarnos preguntar a una muestra de parejas (hombre y mujer) que conviven juntos el tiempo semanal que dedica a la lectura. Aquí, a cada individuo sólo le tomamos una medida, pero cada pareja, como una unidad, contribuye con un par de observaciones. Parece razonable asumir que existe una relación entre las dos muestras y, por tanto, que las muestras están apareadas.

Muchos diseños experimentales utilizan muestras relacionadas, y todos ellos tienen una cosa en común: el conocimiento de una de las observaciones de un par nos proporciona alguna información sobre la otra observación del mismo par. Cuando éste es el caso, puesto que las observaciones de cada par pertenecen al mismo individuo o a dos individuos emparejados, podemos transformar las observaciones originales en *diferencias*, $D_i = X_{1i} - X_{2i}$, haciendo así que a cada par de individuos le corresponda una sola observación. Estas observaciones D_i informan sobre el cambio producido entre las observaciones de cada par. Tendremos así una única variable $D = X_1 - X_2$, con media $\overline{D}$ de la que podremos servirnos para efectuar inferencias sobre la diferencia de las medias poblacionales: $\mu_D = \mu_1 - \mu_2$.

Como las muestras son apareadas, el tamaño de las dos muestras ha de ser el mismo; es decir, $n_1 = n_2 = n$. En las condiciones generales que se han dado al principio del tema, si la variable aleatoria $D = X_1 - X_2$ es Normal, entonces el estadístico

$$T = \frac{\overline{D} - (\mu_1 - \mu_2)}{S_D/\sqrt{n}} \tag{7.1}$$

se distribuye según el modelo t de Student con $n - 1$ grados de libertad. Si la variable aleatoria D no es Normal pero se verifica que el tamaño muestral es grande ($n \geq 30$) entonces el estadístico T se aproxima a una variable t de Student con $n - 1$ grados de libertad.

Por esta razón, a este test se le denomina **test t de Student sobre comparación de dos medias con muestras apareadas**.

7.1.2.3.2. Hipótesis nula y alternativa del test

Hay tres posibilidades:

$H_0 : \sigma_1^2 = \sigma_2^2$	$H_0 : \sigma_1^2 \geq \sigma_2^2$	$H_0 : \sigma_1^2 \leq \sigma_2^2$
$H_1 : \sigma_1^2 \neq \sigma_2^2$	$H_1 : \sigma_1^2 < \sigma_2^2$	$H_1 : \sigma_1^2 > \sigma_2^2$

7.1.2.3.3. Condiciones para poder realizar el test

Para realizar cualquiera de los tres tipos de tests anteriores, es necesario que se verifiquen las condiciones siguientes:

- Las dos muestras son aleatorias.
- Las dos muestras son dependientes (relacionadas o apareadas).
- La variable diferencia $D = X_1 - X_2$ es Normal o el tamaño muestral común ($n = n_1 = n_2$) es grande ($n \geq 30$).

7.1.2.3.4. Resolución mediante MINITAB

Para realizar el test de comparación de dos medias con muestras dependientes (relacionadas o apareadas) hay que seleccionar **Stat** ⇒**Basic Statistics** ⇒**Paired t**.

Con el archivo de datos **Pulse.mtw**, comprobemos si se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que el pulso medio poblacional antes de correr es igual al pulso medio poblacional después de correr. Lo que se quiere es comparar la media poblacional de la variable **Pulse1** con la media poblacional de la variable **Pulse2**. Las hipótesis nula y alternativa son $H_0 : \mu_1 = \mu_2$ y $H_1 : \mu_1 \neq \mu_2$, siendo X_1=*"Pulso antes de correr"* y X_2=*"Pulso después de correr"*. Como las dos variables están observadas en los mismos individuos, podemos afirmar que las muestras están relacionadas; es decir, son apareadas o dependientes. Por tanto, nos encontramos ante un test de comparación de dos medias poblacionales con muestras apareadas. Aunque la variable aleatoria diferencia, $D = X_1 - X_2$, no fuese Normal, se puede aplicar este test debido a que los tamaños muestrales son suficientemente grandes: $n_1 = n_2 = n = 92$.

Para hacer este test seleccionamos **Stat** ⇒**Basic Statistics** ⇒**Paired t**. Activamos la opción **Samples in columns**; en **First sample** seleccionamos, de la lista de variables de la izquierda, la columna '**Pulse1**'; en **Second sample** seleccionamos, de la lista de variables de la izquierda, la columna '**Pulse2**'. Si pulsamos el botón **Options** nos aparece un cuadro de diálogo similar al de la opción anterior (**2-Sample t**⇒**Options**). En este cuadro de diálogo dejamos lo que aparece por defecto (**Confidence level: 95**, **Test difference: 0**, **Alternative: not equal**).

Podemos comprobar, en la ventana de sesión, que el p-valor es igual a 0, el mínimo posible y, por supuesto, menor que el nivel de significación, $\alpha = 0'05$, por lo que debemos rechazar la hipótesis nula y, por tanto, aceptar la hipótesis alternativa. Aceptamos, por tanto, que el pulso medio poblacional antes de correr es distinto del pulso medio poblacional después de correr. Como la media muestral del pulso después de correr ($80'00$) es mayor que la media muestral del pulso antes de correr ($72'87$) podríamos, incluso, aceptar que la media poblacional del pulso después de correr es mayor que la media poblacional del pulso antes de correr. El intervalo de confianza al 95 % para la diferencia de medias poblacionales, en este caso, es $(-9'92, -4'34)$.

7.2. Ejemplos que se van a resolver en clase

Ejemplo 7.1. En una determinada biblioteca se observa el precio, en euros, de los libros. Los libros se clasifican en dos grupos o poblaciones: los libros que se han prestado pocas veces en el

último año y los libros que se han prestado muchas veces en el último año. Sean las variables X_1=*precio, en euros, de los libros que se han prestado pocas veces en el último año* y X_2=*precio, en euros, de los libros que se han prestado muchas veces en el último año*. Para dos muestras independientes tenemos los resultados de X_1 y X_2:

x_{1i}	x_{2i}
75	76
32	30
30	45
34	69
42	46
57	53
51	97
36	43
82	42
45	37
58	48
66	45
40	82
35	61
51	57

a) ¿Se puede aceptar, con un nivel de significación de $0'05$, que las dos muestras son aleatorias? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $0'05$, que las dos variables, X_1 y X_2, son Normales? ¿Por qué?

c) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la varianza poblacional del precio de los libros que se prestan poco es igual a la varianza poblacional del precio de los libros que se prestan mucho? ¿Por qué?

d) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional del precio de los libros que se prestan poco es igual a la media poblacional del precio de los libros que se prestan mucho? ¿Por qué?

Ejemplo 7.2. Sean las dos variables X_1=*número de palabras que contienen los resúmenes (abstracts) de los artículos científicos escritos en español* y X_2=*número de palabras que contienen los resúmenes (abstracts) de los artículos científicos escritos en inglés*. Para dos muestras independientes tenemos los resultados de X_1 y X_2:

x_{1i}	70	65	68	74	79	67	75	80	62	69
	61	57	71	74	82	91	70	64	72	67
	74	70	81	85	70	74	75	71	69	54
x_{2i}	80	47	59	67	89	57	72	78	74	72
	104	118	89	87	79	78	101	120	107	95
	85	87	90	98	89	75	90	101	85	94

a) ¿Se puede aceptar, con un nivel de significación de $0'05$, que las dos muestras son aleatorias? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la varianza poblacional de la longitud de los resúmenes de artículos escritos en español es igual a la varianza poblacional de la longitud de los resúmenes de artículos escritos en inglés? ¿Por qué?

c) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional de la longitud de los resúmenes de artículos escritos en español es igual a la media poblacional de la longitud de los resúmenes de artículos escritos en inglés? ¿Por qué?

Ejemplo 7.3. Dos expertos califican una muestra aleatoria de 30 libros según su calidad (1=muy mala, 2=mala, 3=regular, 4=buena, 5=muy buena). En la tabla siguiente aparece la opinión del primer experto (X_1) y la opinión del segundo experto (X_2).

x_{1i}	x_{2i}	x_{1i}	x_{2i}
2	1	4	4
5	4	4	3
4	5	5	4
2	3	5	3
3	3	1	2
1	5	2	5
3	3	2	3
1	3	3	2
4	2	4	1
2	5	4	2
3	2	1	3
4	3	2	4
3	3	1	2
1	3	5	5
2	5	5	2

a) ¿Se puede aceptar, con un nivel de significación de $0'05$, que las dos muestras son aleatorias? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional de los resultados de la opinión del primer experto es igual a la media poblacional de los resultados de la opinión del segundo experto? ¿Por qué?

7.3. Actividades de aplicación de los contenidos

7.3.1. Problemas propuestos

Problema 7.1.

a) Crea un nuevo proyecto de ***Minitab***. Abre la hoja de datos **Transacciones.mtw** (datos del Ejercicio 1.2). ¿Se puede aceptar, con un nivel de significación de $0'05$, que la varianza poblacional del número anual de transacciones de referencia de las bibliotecas públicas es igual a la varianza poblacional del número anual de transacciones de referencia de las bibliotecas universitarias? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional del número anual de transacciones de referencia de las bibliotecas públicas es mayor que la media poblacional del número anual de transacciones de referencia de las bibliotecas universitarias? ¿Por qué?

c) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la varianza poblacional del porcentaje de transacciones de referencia finalizadas de las bibliotecas públicas es igual a la varianza poblacional del porcentaje de transacciones de referencia finalizadas de las bibliotecas universitarias? ¿Por qué?

d) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional del porcentaje de transacciones de referencia finalizadas de las bibliotecas públicas es mayor que la media poblacional del porcentaje de transacciones de referencia finalizadas de las bibliotecas universitarias? ¿Por qué?

e) Si quieres, puedes grabar el proyecto de ***Minitab***.

Problema 7.2. En la siguiente tabla aparece el número de citas de los artículos del área de Información y Documentación (X_1) para una muestra aleatoria de 10 artículos de dicho área y el número de citas de los artículos del área de Periodismo (X_2) para una muestra aleatoria de 12 artículos de dicho área, independiente de la anterior muestra:

x_{1i}	x_{2i}
21	18
16	13
14	11
27	24
30	27
15	12
10	7
18	11
20	17
14	11
	18
	13

a) Crea un nuevo proyecto de ***Minitab***. Guarda los datos en el archivo **Citas-Articulos.mtw**. ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la varianza del número de citas en la población de todos los artículos del área de Información y Documentación es igual a la varianza del número de citas en la población de todos los artículos del área de Periodismo? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $\alpha = 0'05$, que la media del número de citas en la población de todos los artículos del área de Información y Documentación es igual a la media del número de citas en la población de todos los artículos del área de Periodismo? ¿Por qué?

c) Si quieres, puedes grabar el proyecto de ***Minitab***.

Problema 7.3. Elegimos al azar 30 parejas (hombre y mujer) que conviven juntos y observamos el número de veces que los hombres han visitado alguna biblioteca en los tres últimos meses (X_1) y el número de veces que las mujeres han visitado alguna biblioteca en los tres últimos meses (X_2). Los resultados se muestran en la siguiente tabla.

x_{1i}	x_{2i}	x_{1i}	x_{2i}	x_{1i}	x_{2i}
12	8	8	10	25	14
30	11	14	15	12	16
10	12	20	12	8	10
20	16	13	19	23	20
15	10	11	6	14	17
14	9	7	7	8	10
11	12	6	7	12	23
9	10	8	6	27	10
7	7	15	20	32	27
5	4	42	35	14	18

a) Crea un nuevo proyecto de ***Minitab***. Guarda los datos en el archivo **Visitas-Biblioteca-Parejas.mtw**. ¿Podemos afirmar que hay diferencia significativa entre los hombres y las mujeres de las parejas en cuanto al número de veces que van a la biblioteca? ¿Por qué?

b) Si quieres, puedes grabar el proyecto de ***Minitab***.

Problema 7.4. En la siguiente tabla aparece el número de usuarios diarios de la biblioteca A (variable X_1) y el número de usuarios diarios de la biblioteca B (variable X_2) en 10 días elegidos al azar.

x_{1i}	x_{2i}
51	45
72	58
35	32
70	56
75	68
98	76
100	88
80	69
72	57
90	75

a) Crea un nuevo proyecto de ***Minitab***. Guarda los datos en el archivo **Usuarios-Diarios-2-Bcas.mtw**. Calcula, en una nueva columna, los resultados de la variable diferencia $D = X_1 - X_2$. ¿Se puede aceptar, con un nivel de significación de $0'05$, que la variable diferencia, $D = X_1 - X_2$, es Normal? ¿Por qué?

b) ¿Se puede aceptar, con un nivel de significación de $0'05$, que la media poblacional del número de usuarios diarios de la biblioteca A es igual a la media poblacional del número de usuarios diarios de la biblioteca B? ¿Por qué?

c) Si quieres, puedes grabar el proyecto de ***Minitab***.

Problema 7.5. Se nos ha señalado la posibilidad de que se paguen sueldos distintos a documentalistas según el sexo. Presumiblemente, a los hombres se les ha pagado más que a las mujeres. En una muestra aleatoria de 50 mujeres documentalistas hemos obtenido un sueldo medio anual de 118851 euros con una cuasi-desviación típica de 2259 euros. En una muestra aleatoria de 35 hombres documentalistas hemos obtenido un sueldo medio anual de 135675 euros con una cuasi-desviación típica de 1807 euros. A la vista de estos datos, y utilizando un nivel de significación de $0'05$, ¿podemos afirmar que el sueldo de los hombres documentalistas es mayor que el de las mujeres documentalistas?

7.3.2. Soluciones de los problemas propuestos

Solución del problema 7.1.

En el **Problema** 5.6. ya hemos comprobado que las muestras de los datos de las variables **TR**, **TRF** y **Porcentaje TRF** son aleatorias y que las variables **TR**, **TRF** y **Porcentaje TRF** son Normales.

a)

X_1=*número anual de transacciones de referencia de las bibliotecas públicas=variable* ***TR*** *en la población de las bibliotecas públicas= variable* ***TR*** *para el grupo en el que la variable* ***Tipo*** *es igual a* ***1****.*

X_2=*número anual de transacciones de referencia de las bibliotecas universitarias=variable* ***TR*** *en la población de las bibliotecas universitarias= variable* ***TR*** *para el grupo en el que la variable* ***Tipo*** *es igual a* ***2****.*

La pregunta que se nos hace es: ¿$\sigma_1^2 = \sigma_2^2$?

Tenemos que realizar un test de comparación de dos varianzas poblacionales con muestras independientes y medias poblacionales desconocidas.

Hipótesis nula y alternativa:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Condiciones:

- Como la muestra total de datos de la variable **TR** es aleatoria, entonces las dos muestras son aleatorias.
- Las dos muestras son independientes entre sí porque no existe relación entre las bibliotecas públicas y las bibliotecas universitarias.
- Como la variable **TR** es Normal, entonces las dos variables aleatorias, X_1 y X_2, son Normales.
- Las dos medias, μ_1 y μ_2, son desconocidas.

Resolución con *Minitab*:

Como las muestras son independientes y las medias poblacionales son desconocidas, hay que seleccionar **Stat** ⇒**Basic Statistics** ⇒**2 Variances**. Hay que activar **Samples in one column** pues las dos muestras están, realmente, en una misma columna. En **Samples** hay que seleccionar (de las variables de la izquierda) la columna **TR** y en **Subscripts** hay que seleccionar (de las variables de la izquierda) la columna **Tipo**.

El p-valor, para el test F, es $0'055$; un poco mayor que el nivel de significación ($\alpha = 0'05$); por tanto, aceptamos H_0; es decir, aceptamos que las dos varianzas son iguales.

b)

Utilizamos las dos mismas variables, X_1 y X_2, del apartado anterior.

La pregunta que se nos hace es: ¿$\mu_1 > \mu_2$?

Tenemos que realizar un test de comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas.

Hipótesis nula y alternativa:

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

Condiciones:

- Como ya sabemos, las dos muestras son aleatorias.
- Como ya sabemos, las dos muestras son independientes.
- Como ya sabemos, las dos variables aleatorias, X_1 y X_2, son Normales.
- En el apartado anterior hemos comprobado que las dos varianzas poblacionales son desconocidas pero iguales.

Resolución con *Minitab*:

Como las muestras son independientes y las varianzas poblacionales son desconocidas, hay que seleccionar **Stat** ⇒**Basic Statistics** ⇒**2-Sample t**. Hay que activar **Samples in one column**. En **Samples** hay que seleccionar (de las variables de la izquierda) la columna **TR**

y en **Subscripts** hay que seleccionar (de las variables de la izquierda) la columna **Tipo**. Hay que activar la opción **Assume equal variances** (pues las dos varianzas poblacionales son desconocidas pero iguales). En **Options** hay que seleccionar **greater than** en **Alternative** (pues $H_1 : \mu_1 > \mu_2$).

El p-valor es $0'005$; menor que el nivel de significación ($\alpha = 0'05$); por tanto, rechazamos H_0 y, por tanto, aceptamos H_1; es decir, la media poblacional del número anual de transacciones de referencia de las bibliotecas públicas es mayor que la media poblacional del número anual de transacciones de referencia de las bibliotecas universitarias.

c)

X_1=*número anual de transacciones de referencia finalizadas de las bibliotecas públicas=variable* ***TRF*** *en la población de las bibliotecas públicas= variable* ***TRF*** *para el grupo en el que la variable* ***Tipo*** *es igual a* ***1****.*

X_2=*número anual de transacciones de referencia finalizadas de las bibliotecas universitarias=variable* ***TRF*** *en la población de las bibliotecas universitarias= variable* ***TRF*** *para el grupo en el que la variable* ***Tipo*** *es igual a* ***2****.*

La pregunta que se nos hace es: ¿$\sigma_1^2 = \sigma_2^2$?

Tenemos que realizar un test de comparación de dos varianzas poblacionales con muestras independientes y medias poblacionales desconocidas.

Hipótesis nula y alternativa:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Condiciones:

- Como la muestra total de datos de la variable **TRF** es aleatoria, entonces las dos muestras son aleatorias.
- Las dos muestras son independientes entre sí porque no existe relación entre las bibliotecas públicas y las bibliotecas universitarias.
- Como la variable **TRF** es Normal, entonces las dos variables aleatorias, X_1 y X_2, son Normales.
- Las dos medias, μ_1 y μ_2, son desconocidas.

Resolución con *Minitab*:

Como las muestras son independientes, hay que seleccionar **Stat** ⇒**Basic Statistics** ⇒**2 Variances**. Hay que activar **Samples in one column** pues las dos muestras están, realmente, en una misma columna. En **Samples** hay que seleccionar (de las variables de la izquierda) la columna **TRF** y en **Subscripts** hay que seleccionar (de las variables de la izquierda) la columna **Tipo**.

El p-valor, para el test F, es $0'034$; menor que el nivel de significación ($\alpha = 0'05$); por tanto, rechazamos H_0; es decir, aceptamos que las dos varianzas son distintas.

d)

Utilizamos las dos mismas variables, X_1 y X_2, del apartado anterior.

La pregunta que se nos hace es: ¿$\mu_1 > \mu_2$?

Tenemos que realizar un test de comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas.

Hipótesis nula y alternativa:

$$H_0 : \mu_1 \leq \mu_2$$
$$H_1 : \mu_1 > \mu_2$$

Condiciones:

- Como ya sabemos, las dos muestras son aleatorias.
- Como ya sabemos, las dos muestras son independientes.
- Como ya sabemos, las dos variables aleatorias, X_1 y X_2, son Normales.
- En el apartado anterior hemos comprobado que las dos varianzas poblacionales son desconocidas y distintas.

Resolución con *Minitab*:

Como las muestras son independientes, hay que seleccionar **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**2-Sample t**. Hay que activar **Samples in one column**. En **Samples** hay que seleccionar (de las variables de la izquierda) la columna **TRF** y en **Subscripts** hay que seleccionar (de las variables de la izquierda) la columna **Tipo**. Hay que desactivar la opción **Assume equal variances**. En **Options** hay que seleccionar **greater than** en **Alternative** (pues $H_1 : \mu_1 > \mu_2$).

El p-valor es $0'005$; menor que el nivel de significación ($\alpha = 0'05$); por tanto, rechazamos H_0 y, por tanto, aceptamos H_1; es decir, la media poblacional del número anual de transacciones de referencia finalizadas de las bibliotecas públicas es mayor que la media poblacional del número anual de transacciones de referencia finalizadas de las bibliotecas universitarias.

Solución del problema 7.2.

a)

X_1=*Número de citas en la población de todos los artículos del área de Información y Documentación*.

X_2=*Número de citas en la población de todos los artículos del área de Periodismo*.

La pregunta que se nos hace es: ¿$\sigma_1^2 = \sigma_2^2$?

Tenemos que realizar un test de comparación de dos varianzas poblacionales con muestras independientes y medias poblacionales desconocidas.

Hipótesis nula y alternativa:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Condiciones:

- Hay que probar que las dos muestras son aleatorias. Para ello, realizamos el test de las rachas sobre aleatoriedad de las dos muestras. Los p-valores son $0'888$ (para X_1) y $0'466$ (para X_2). Ambos son mayores que el nivel de significación ($0'05$). Por tanto, las dos muestras son aleatorias.
- Las dos muestras son independientes entre sí porque no existe relación entre los artículos del área de Información y Documentación y los artículos del área de Periodismo.

- Hay que probar que las dos variables aleatorias, X_1 y X_2, son Normales. Los p-valores del test de normalidad *AD* (*de Anderson-Darling*) son $0'413$ (para X_1) y $0'137$ (para X_2). Ambos son mayores que el nivel de significación ($0'05$). Por tanto, las dos variables aleatorias son Normales.
- Las dos medias, μ_1 y μ_2, son desconocidas.

Resolución con *Minitab*:

Como las muestras son independientes, hay que seleccionar **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**2 Variances**. Hay que activar **Samples in different columns**. En **First** hay que seleccionar la variable **X**$_1$ y en **Second** hay que seleccionar la variable **X**$_2$.

El p-valor, para el test F, es $0'844$; mayor que el nivel de significación ($\alpha = 0'05$); por tanto, aceptamos H_0; es decir, aceptamos que las dos varianzas son iguales.

b)

Utilizamos las dos mismas variables, X_1 y X_2, del apartado anterior.

La pregunta que se nos hace es: ¿$\mu_1 = \mu_2$?

Tenemos que realizar un test de comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas.

Hipótesis nula y alternativa:

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

Condiciones:

- Como ya sabemos, las dos muestras son aleatorias.
- Como ya sabemos, las dos muestras son independientes.
- Como ya sabemos, las dos variables aleatorias, X_1 y X_2, son Normales.
- En el apartado anterior hemos comprobado que las dos varianzas poblacionales son desconocidas pero iguales.

Resolución con *Minitab*:

Como las muestras son independientes, hay que seleccionar **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**2-Sample t**. Hay que activar **Samples in different columns**. En **First** hay que seleccionar la variable **X**$_1$ y en **Second** hay que seleccionar la variable **X**$_2$. Hay que activar la opción **Assume equal variances**. En **Options** hay que seleccionar **not equal** en **Alternative**.

El p-valor es $0'209$; mayor que el nivel de significación ($\alpha = 0'05$); por tanto, aceptamos H_0. En consecuencia, la media del número de citas en la población de todos los artículos del área de Información y Documentación es igual a la media del número de citas en la población de todos los artículos del área de Periodismo.

Solución del problema 7.3.

X_1=*Número de veces que los hombres (de las parejas) han visitado alguna biblioteca en los tres últimos meses*.

X_2=*Número de veces que las mujeres (de las parejas) han visitado alguna biblioteca en los tres últimos meses*.

La pregunta que se nos hace es equivalente a la siguiente: ¿$\mu_1 \neq \mu_2$?

Como los hombres y las mujeres conviven juntos, puede influir el resultado de una de las variables en el resultado de la otra variable. En consecuencia, las dos muestras son dependientes (relacionadas o apareadas).

Tenemos que realizar un test de comparación de dos medias con muestras dependientes.

Hipótesis nula y alternativa:

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

Condiciones:

- Debemos comprobar que las dos muestras son aleatorias. Para ello, realizamos el test de las rachas sobre aleatoriedad de las dos muestras. Los p-valores son $0'898$ (para X_1) y $0'587$ (para X_2). Ambos son mayores que el nivel de significación usual ($0'05$). Por tanto, las dos muestras son aleatorias.
- Las dos muestras son dependientes.
- No es necesario comprobar que variable diferencia $D = X_1 - X_2$ es Normal puesto que el tamaño muestral común es grande; concretamente, $n = 30$.

Resolución con *Minitab*:

Como las muestras son dependientes o apareadas, hay que seleccionar **Stat** ⇒**Basic Statistics** ⇒**Paired t**. Hay que activar **Samples in columns**. En **First sample** hay que seleccionar la variable $\mathbf{X_1}$ y en **Second sample** hay que seleccionar la variable $\mathbf{X_2}$. En **Options** hay que seleccionar **not equal** en **Alternative**.

El p-valor es $0'156$; mayor que el nivel de significación usual ($0'05$); por tanto, aceptamos H_0. En consecuencia, la media del número de veces que los hombres (de las parejas) han visitado alguna biblioteca en los tres últimos meses es igual a la media del número de veces que las mujeres (de las parejas) han visitado alguna biblioteca en los tres últimos meses; es decir, no hay diferencia significativa entre los hombres y las mujeres de las parejas en cuanto al número de veces que van a la biblioteca.

Solución del problema 7.4.

X_1=*Número de usuarios diarios de la biblioteca A*.

X_2=*Número de usuarios diarios de la biblioteca B*.

a)

Tenemos que realizar un test de normalidad para la variable aleatoria $D = X_1 - X_2$.

Una de las condiciones para hacer este test es que la muestra de datos de la variable D ha de ser aleatoria. Por tanto, en primer lugar tenemos que aplicar el test de las rachas sobre aleatoriedad de la muestra de datos de la variable D. El p-valor para el test de las rachas es $0'122$; mayor que el nivel de significación ($0'05$); por tanto, aceptamos que la muestra de datos de la variable D es aleatoria.

Ahora ya podemos aplicar el test de normalidad para la variable D. El p-valor de test *AD* es $0'508$; mayor que el nivel de significación ($0'05$); por tanto, aceptamos que la variable D es Normal.

b)

La pregunta que se nos hace es: ¿$\mu_1 \neq \mu_2$?

Los individuos a los que se les observa ambas variables son los días: En el día 1 se observan X_1 y X_2, en el día 2 se observan X_1 y X_2, en el día 3 se observan X_1 y X_2, etc. Por tanto, los individuos de las dos muestras son los mismos. En consecuencia, las dos muestras son dependientes (relacionadas o apareadas).

Tenemos que realizar un test de comparación de dos medias con muestras dependientes.

Hipótesis nula y alternativa:

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

Condiciones:

- Debemos comprobar que las dos muestras son aleatorias. Para ello, realizamos el test de las rachas sobre aleatoriedad de las dos muestras. Los p-valores son $0'18$ (para X_1) y $0'18$ (para X_2). Ambos son mayores que el nivel de significación usual ($0'05$). Por tanto, las dos muestras son aleatorias.
- Las dos muestras son dependientes.
- En el apartado anterior ya se ha comprobado que variable diferencia $D = X_1 - X_2$ es Normal.

Resolución con *Minitab*:

Como las muestras son dependientes o apareadas, hay que seleccionar **Stat** ⇒**Basic Statistics** ⇒**Paired t**. Hay que activar **Samples in columns**. En **First sample** hay que seleccionar la variable $\mathbf{X_1}$ y en **Second sample** hay que seleccionar la variable $\mathbf{X_2}$. En **Options** hay que seleccionar **not equal** en **Alternative**.

El p-valor es 0; menor que cualquier nivel de significación; por tanto, rezhazamos H_0. En consecuencia, la media poblacional del número de usuarios diarios de la biblioteca A es distinta de la media poblacional del número de usuarios diarios de la biblioteca B.

Solución del problema 7.5.

X_1=*Sueldo anual de las mujeres documentalistas*.

X_2=*Sueldo anual de los hombres documentalistas*.

La pregunta que se nos hace es: ¿$\mu_1 < \mu_2$?

Tenemos que realizar un test de comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas. Antes de realizar este test es necesario hacer un test de comparación de las dos varianzas poblacionales pues éstas son desconocidas pero no sabemos si son iguales o si son distintas.

1)

Vamos a responder, en primer lugar, a la siguiente pregunta: ¿$\sigma_1^2 = \sigma_2^2$?

Tenemos que realizar un test de comparación de dos varianzas poblacionales con muestras independientes y medias poblacionales desconocidas.

Como $S_1 = 2259$ euros entonces $S_1^2 = 5103081$ euros^2.

Como $S_2 = 1807$ euros entonces $S_2^2 = 3265249$ euros^2.

Hipótesis nula y alternativa:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

Condiciones:

- En el enunciado del problema se nos dice que las dos muestras son aleatorias.
- Por el enunciado también se deduce que las dos muestras son independientes entre sí.
- Suponemos que las dos variables son Normales (no podemos demostrarlo, con los datos que nos dan).
- Las dos medias, μ_1 y μ_2, son desconocidas.

Resolución con *Minitab*:

Como las muestras son independientes, hay que seleccionar **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**2 Variances**. Hay que activar **Summarized data**. En **First** hay que poner **50** en **Sample size** y **5103081** en **Variance**. En **Second** hay que poner **35** en **Sample size** y **3265249** en **Variance**.

El p-valor, para el test F, es $0'173$; mayor que el nivel de significación ($\alpha = 0'05$); por tanto, aceptamos H_0; es decir, aceptamos que las dos varianzas son iguales.

2)

Utilizamos las dos mismas variables, X_1 y X_2, del apartado anterior.

La pregunta que se nos hace es: ¿$\mu_1 < \mu_2$?

Tenemos que realizar un test de comparación de dos medias con muestras independientes y varianzas poblacionales desconocidas.

Hipótesis nula y alternativa:

$$H_0 : \mu_1 \geq \mu_2$$
$$H_1 : \mu_1 < \mu_2$$

Condiciones:

- Como ya sabemos, las dos muestras son aleatorias.
- Como ya sabemos, las dos muestras son independientes.
- Como los dos tamaños muestrales son grandes ($n_1 = 50 \geq 30$, $n_2 = 35 \geq 30$) no es necesario demostrar que las dos variables son Normales.
- En el apartado anterior hemos comprobado que las dos varianzas poblacionales son desconocidas pero iguales.

Resolución con *Minitab*:

Como las muestras son independientes, hay que seleccionar **Stat** $\Rightarrow$**Basic Statistics** $\Rightarrow$**2-Sample t**. Hay que activar **Summarized data**. En **First** hay que poner **50** en **Sample size**, **118851** en **Mean** y **2259** en **Standard deviation**. En **Second** hay que poner **35** en **Sample size**, **135675** en **Mean** y **1807** en **Standard deviation**. Hay que activar **Assume equal variances** (pues las varianzas poblacionales son desconocidas pero iguales). En **Options** hay que seleccionar **less than** en **Alternative** (pues $H_1 : \mu_1 < \mu_2$).

El p-valor es cero; menor que cualquier nivel de significación; por tanto, rechazamos H_0 y, en consecuencia, aceptamos H_1; es decir, la media del sueldo anual de la mujeres documentalistas es menor que la media del sueldo anual de los hombres documentalistas.