Guía de estudio - GES1

ESTADÍSTICA DESCRIPTIVA

- Presentación de la Guía de Estudio (GES)
- Objetivos
- Contenidos
- Bibliografía
- Fe de erratas

Presentación

Esta primera Guía de Estudio (**GES_1**) pretende orientar el estudio de los contenidos de los Módulos 1, 2, 3 y 4 de la asignatura *"estadística*". Con este objetivo, esta **GES** contiene el siguiente material:

1. Una breve introducción a la Estadística
2. Las propiedades de la **varianza** y la media aritmética, la media ponderada, la moda y el coeficiente de variación.
3. Un ejemplo de cálculo de la **varianza** y la **desviación estándar.**

Materiales: para trabajar esta **GES_1** necesitáis los materiales básicos de la asignatura.

Calendario: la temporalización de la **GES_1** será la prevista en el Plan Docente.

Objetivos

Con el estudio de esta **GES** se pretende que el estudiante abarque los siguientes objetivos:

1. Servir de material de apoyo al estudiante en ésta primera parte de la estadística descriptiva a fin de que se puedan alcanzar los objetivos fijados en los módulos que la integran.
2. Conocer el tipo de datos existentes y su representación gráfica.
3. Saber realizar los cálculos y entender la aplicación de los parámetros estadísticos (medidas de centralización y dispersión)

Contenidos

1. Una breve introducción a la Estadística

Las estadísticas y la estadística

Distinguir entre **estadísticas** y **estadística** supone el primer paso para enteder de qué trataremos en esta asignatura.

Si consultamos las *Estadísticas de la Sociedad de la Información 2001* producidas por la Generalitat de Catalunya veremos un ejemplo de **estadísticas**, pero la forma en la cual estas estadísticas han sido construidas y la interpretación de las mismas precisa de un buen conocimiento de **estadística**.

Así, en este documento nos informan de que un 45,7% de los hogares en Cataluña en el año 2001 tenían algún ordenador en casa y que esta proporción en el año 2000 era de un 44,8%.

Para llegar a estos porcentajes hemos tenido que construir un cuestionario donde una de las preguntas era si había o no ordenador en casa, pasarlo a todos los hogares catalanes o a una muestra representativa de ellas y calcular la relación entre el número de hogares de Cataluña que tienen ordenador sobre el total de hogares en Cataluña. Realizar todo este proceso con cuidado necesita conocer estadística. Además, de estos dos porcentajes (45,7% en el año 2001 y 44,8% en el año 2000) no se deriva *necesariamente*, por más que lo pueda parecer a primera vista, que la presencia de ordenadores en los hogares catalanes se ha incrementado entre el 2000 y el 2001. La comparación de estas dos proporciones tiene que tener en cuenta como se han recogido los datos, en cuántos hogares se ha preguntado y otros elementos que estudiaremos a lo largo de este curso.

En general, la estadística se divide en dos categorías:

- *Estadística descriptiva*, que es la parte de la estadística encargada de extraer y organizar los datos procedentes de un determinado conjunto de observaciones.

- *Estadística inferencial*, que pretende predecir una información en torno a un conjunto de datos, a partir de los resultados extraídos de un subconjunto de ellas.

Así, denominaremos **población** al conjunto de objetos, individuos o acontecimientos cuyas propiedades queremos analizar. Una **muestra** será un subconjunto de la población objeto de estudio.

El conjunto de los datos recogidos para realizar un estudio estadístico recibirá el nombre de **variable aleatoria,** que normalmente se denotará por X. Los datos correspondientes pueden ser básicamente de dos tipos: cuantitativos y cualitativos.

Datos cuantitativos y datos cualitativos

En las estadísticas nos podemos encontrar con que estamos tratando con dos tipos de datos: los **cuantitativos** (que expresan una cantidad[1]) y las **cualitativos** (que expresan una calidad[2]). Por ejemplo, cuántos ordenadores hay en una casa es un dato **cuantitativo** (que puede ir desde 0 hasta cualquier cifra), mientras que el saber si en un hogar hay ordenador o no es un dato **cualitativo** (que divide los hogares en dos grupos, aquéllos que tienen ordenador y aquéllos que no tienen). En efecto, los datos **cualitativos** son aquéllos que dividen la población en grupos diferentes; como el sexo, que divide en la población en hombres y mujeres. Por otra parte, los datos **cuantitativos** dan una cifra numérica a la variable estudiada; como la edad, que nos cuantifica a todo el mundo según los años que llevamos de vida.

Dentro del grupo de los datos cuantitativos podemos distinguir otras dos categorías: las **variables discretas**, que serían aquéllas que sólo pueden asumir ciertos valores (p.e., número de estudiantes en la asignatura de Estadística) y por otra parte, las **variables continuas** que pueden coger cualquier valor dentro de un rango específico (p.e., altura de los alumnos de una clase).

Una vez tenemos recogidos los datos, agrupamos éstos de forma excluyente dando a cada uno de ellos el número de observaciones, es decir, el número a veces que se repite cada valor. Éste valor tomará el nombre de **frecuencia**.

Recoger información: observar y experimentar

Para recoger datos estadísticos podemos optar por la **observación** o por la **experimentación**.

[1] Son los que en vuestro manual llama datos numéricos.
[2] Son los que en vuestro manual llama datos categóricos.

La **observación** implica **recoger directamente los datos** que se van dando de un fenómeno determinado; por ejemplo, ver la altura de una niña o el número de coches que pasan por delante de un determinado colegio.

Por otra parte, cuando llevamos a cabo un **experimento** no nos limitaremos a observar directamente la realidad, si no que **diseñamos unas condiciones específicas** para averiguar los efectos de esta *situación experimental* sobre el fenómeno estudiado.

Un tipo específico de técnica de recogida de datos, entre la observación y la experimentación, es el **cuestionario**; se trata de preguntar a una persona sobre una serie de variables: su edad, el sexo, el nivel de instrucción, quien votará en las próximas elecciones, si está a favor o en contra de la despenalización total de las drogas, etc.

Representación gráfica de una distribución

Según el tipo de datos, utilizaremos uno u otro tipo de representación gráfica. Si los datos son **cuantitativos**, utilizaremos un gráfico de **tallos y hojas** o un **histograma**. Si los datos son **cualitativos**, utilizaremos un **diagrama de barras**.

Medias y proporciones

Según el tipo de datos que contenga una variable utilizaremos un tipo de medida determinada: para las variables **cuantitativas** utilizamos la **media** y para las variables **cualitativas** utilizamos la **proporción**. Así, por ejemplo, dentro de una población, podemos decir cuál es la proporción de individuos que son hombres y cuál es la proporción de individuos que son mujeres, así como calcular la media de la edad del colectivo analizado.

2. Medidas de centralización y dispersión

Las **medidas de centralización** son aquéllas, el objetivo de las cuales es explicar mediante un valor numérico, cuál es la tendencia mayoritaria de las observaciones de la colección de datos que se analizan. Estos parámetros serán, entre otros, la media, la mediana y la moda.

Las **medidas de dispersión** corresponden a aquellos parámetros, el objetivo de los cuales es detectar el grado de proximidad de los datos con respecto a los valores centrales. Estos parámetros son, entre otros, el rango, los cuartiles, la varianza y la desviación estándar.

Relación de la media aritmética y la mediana

Tanto la media aritmética como la mediana miden el centro de la distribución, pero lo hacen de formas diferentes. En el caso en que la distribución sea simétrica ambas medidas son iguales. Si la distribución es asimétrica, la media aritmética se desplaza hacia la cola de la distribución.

Hace falta tener en cuenta también que si hay valores extremos, la media se verá mucho más afectada que la mediana.

Propiedades de la varianza y de la media aritmética

De vez en cuando, puede ser útil conocer algunos aspectos adicionales de la estadística descriptiva que no han sido incorporados al manual de referencia, bien sea por simplificar un

poco el trabajo (apartados 1, 2 y 3), o bien porque aportan información adicional (apartados 4 y 5). Es por ello que os hacemos llegar estas notas aclaratorias adicionales.

1. Propiedades de la varianza

a) La **varianza** no puede ser nunca negativa.

b) La **varianza** de una constante es cero.

c) Si a los valores de una distribución añadimos o restamos una constante, la **varianza** no se ve afectada y se mantiene constante.

d) La **varianza** del producto de una variable por un número es igual al cuadrado del número por la **varianza** de la variable

2. Propiedades de la media aritmética

a) La suma de las desviaciones respecto de la media suman cero.

b) La media de una constante es la propia constante.

c) Si a los valores de una distribución añadimos o restamos una constante, la media varía en el mismo sentido y en la cuantía de la constante.

d) La media del producto de una constado por una variable es igual a la constante por la media de la variable, de manera, que si multiplican una variable por una constante la media también queda multiplicada por la constante.

e) La media de un conjunto de medias es la media ponderada de este conjunto de medias.

3. La media aritmética ponderada

A veces nos encontramos que no todos los valores de la variable tienen la misma importancia para el estudio de un colectivo. En estos casos, la media no se calcula de la forma habitual, si no que se hace utilizando cada valor de la variable multiplicado por un *coeficiente de ponderación* o *peso*, sumando los mencionados productos y dividiéndolos por la suma de los *pesos* o *ponderaciones*, según la fórmula siguiente:

$$\overline{X}_p = \frac{x_1 p_1 + x_2 p_2 + ...x_n p_n}{p_1 + p_2 + ...p_n}$$

p1 = pesos o ponderaciones

x1 = cada uno de los valores de la distribución

Eso nos permite resolver más fácilmente ciertos ejercicios y también nos facilita establecer discriminaciones de unos colectivos frente a otros para obtener más información sobre los mismos o primar ciertas características respecto de otros.

4. Moda

La Moda (**Mo**) es el valor de la variable que más veces se repite dentro de una distribución o un conjunto de valores. Si trabajamos con datos agrupados sería la clase o intervalo que presenta una frecuencia mayor.

Cuando en una distribución o conjunto de valores no se repite ningún valor, no podemos decir que no hay Moda, al contrario, cuando determinados valores se repitan el mismo número de veces, tendremos distribuciones *multimodales* (*bimodales, trimodales ...*).

La Moda no ofrece mucha información dado que no utiliza toda la información contenida en la muestra, dará más información en la medida en que la muestra sea más simétrica.

5. Coeficiente de Variación (de Karl Pearson)

Cuando necesitamos comparar las dispersiones de dos o más distribuciones, no podemos confrontar simplemente las varianzas o las desviaciones estándares respectivas, dado que estos coeficientes de dispersión vienen afectados por la escala de medida de las respectivas variables. En estos casos es necesario, por lo tanto, eliminar esta influencia convirtiendo estas medidas en números abstractos o adimensionales (sin unidades de medida) y, para conseguir eso, se puede utilizar el coeficiente de variación (CV) de Pearson.

El Coeficiente de Variación de Pearson se define como:

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

$CV = \frac{s}{\bar{x}} 100$ => Si se quiere trabajar con datos porcentuales.

$CV = \frac{\sigma}{\mu}$ => Cuando en lugar de una muestra, se trata de una población.

$\bar{x} =$ Media aritmética, que se cogerá siempre en valor absoluto.

$s =$ **Desviación estándar** o desviación típica.

Y cumple perfectamente esta función, dado que, al dividir la **desviación estándar** por la media aritmética (teniendo en cuenta que ambos estadísticos utilizan las mismas unidades de medida), se elimina la influencia de la escala de medida, convirtiéndose en una medida susceptible de comparación por ser abstracta o adimensional. El inconveniente de este coeficiente reside en que no se puede utilizar cuando la media es cero.

Hay que destacar que dadas dos o más distribuciones, es más homogénea aquélla que tiene un CV más pequeño. Cuando se da en %, se asume que si el CV > 60%, la media deja de ser un estadístico significativo.

3. Un ejemplo de cálculo de la varianza y desviación estándar EN UNA ACTIVIDAD.

El objetivo de la **desviación estándar** es informarnos de hasta qué punto los valores observados se alejan de la media. En el caso hipotético en el que todos los valores observados fueran iguales, la **desviación estándar** sería igual a 0. Es como crear un resumen de todas las diferencias de todos los valores con la media que hemos calculado como medida de distribución central.

Por ejemplo, si miráis la actividad que hay en la página 33 del manual (así es más fácil explicar los pasos a seguir para calcular la **desviación estándar**). Se trata del tiempo que tardan unas pilas al descargarse completamente. Los valores son los dados por la siguiente tabla:

Número de la pila	**minutos que tardan en descargarse**
1	65,1
2	58,4
3	64,9
4	76,0
5	67,8
6	75,1
7	76,7
8	64,2
9	74,9
10	77,6
11	58,0
12	68,0
13	73,3
14	75,4
15	76,0
16	59,4
17	65,4
18	74,7
19	76,6
20	81,3
Número de pilas	20
Total minutos	1408,8
Media	70,44

Según esta distribución, sabiendo que hay 20 pilas y que el tiempo que han tardado todas las pilas al descargarse ha sido de 1408'8 minutos, podemos decir que las pilas se han descargado, como media, después de 70'44 minutos de funcionamiento.

¿Qué buscamos a partir de aquí con la **desviación estándar**? Querríamos saber si todas las pilas se acabaron en torno a estos 70'44 minutos, o si hubo muchas diferencias entre unas y otras. Se trata, en definitiva, de conocer racionalmente si vale la pena comprar la pila del conejillo de Duracel, o si da igual la pila que compras, ya que todas poco más o menos duran igual. Para averiguarlo, utilizamos la **desviación estándar**.

Para llegar a la **desviación estándar** hay que encontrar la **varianza** que, la verdad, no es nada por ella misma sino que todo su sentido se basa en ser el camino imprescindible para el cálculo de la **desviación estándar**. Fijaos que en la fórmula de la **varianza** que hay en la página 32 hay un pequeño error, la correcta es la siguiente:

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

Primero calculamos los factores que hay entre paréntesis. Se trata de darle a cada uno de los valores encontrados (x_i) el valor de la media ($\overline{x}$), tal como se observa en la siguiente tabla:

Número de la pila	minutos que tardan al descargarse	Minutos - media
1	65,1	-5,3
2	58,4	-12,0
3	64,9	-5,5
4	76,0	5,6
5	67,8	-2,6
6	75,1	4,7
7	76,7	6,3
8	64,2	-6,2
9	74,9	4,5
10	77,6	7,2
11	58,0	-12,4
12	68,0	-2,4
13	73,3	2,9
14	75,4	5,0
15	76,0	5,6
16	59,4	-11,0
17	65,4	-5,0
18	74,7	4,3
19	76,6	6,2

20	81,3	10,9
Sumatorio	1408,8	0,0

Si sumáis la columna donde se ha calculado la diferencia entre cada valor con la media os tenéis que encontrar NECESARIAMENTE con el valor 0. Ésta es una propiedad universal de las distribuciones.

El siguiente paso es elevar al cuadrado a todos y cada uno de los valores obtenidos en esta última columna. Así tendréis:

Número de la pila	**minutos que tardan al descargarse**	**Minutos – media**	**Cuadrado**
1	65,1	-5,3	28,5156
2	58,4	-12,0	144,9616
3	64,9	-5,5	30,6916
4	76,0	5,6	30,9136
5	67,8	-2,6	6,9696
6	75,1	4,7	21,7156
7	76,7	6,3	39,1876
8	64,2	-6,2	38,9376
9	74,9	4,5	19,8916
10	77,6	7,2	51,2656
11	58,0	-12,4	154,7536
12	68,0	-2,4	5,9536
13	73,3	2,9	8,1796
14	75,4	5,0	24,6016
15	76,0	5,6	30,9136
16	59,4	-11,0	121,8816
17	65,4	-5,0	25,4016
18	74,7	4,3	18,1476
19	76,6	6,2	37,9456
20	81,3	10,9	117,9396
Sumatorio	1408,8	0,0	958,7680

El sumatorio de la última columna da 958'7680 y éste es el valor que hay que poner al final de la fórmula de la **varianza**. Ahora sólo es necesario que dividamos este 958'7680 por **n-1**, es decir, por el número de observaciones (que en este caso son 20) menos 1 (19). Así pues, la **varianza** será 958'7680 partido por 19 = 50'4615.

Para encontrar la **desviación estándar**, que es lo que queríamos conseguir con todo este proceso matemático, encontramos la raíz cuadrada de este último número, y eso da, finalmente, 7'10 minutos.

Y eso nos indica que de la media de 70 '44 minutos las pilas se separan unos 7'10 minutos. Moraleja: tanto da comprar Duracell como cualquier otra, ya que el conejillo durará poco más o menos lo mismo (claro que siempre puedes decir que estas pilas de nuestra distribución no son representativas de la población total de pilas).

Esperamos que con estos comentarios os ayudemos a comprender el sentido de la **desviación estándar** y cómo se calcula. En la unidad de la distribución normal descubriréis otras interesantes

Bibliografía, materiales complementarios y enlaces de interés

- Como bibliografía complementaria se puede consultar la que figura en el Plan Docente.

Fe de erratas

Error pg. 32: A la fórmula de la varianza en la **página 32** hay un pequeño error (en la fórmula de la varianza hay una $\bar{x}$ con un subíndice 2). Éste 2 es incorrecto, podéis trabajar con la fórmula que hay en esta **GES_1**.

Error Unidad 6 de vídeo: aproximadamente en torno al minuto 7'26, cuando habla de las propiedades de **desviación estándar**, dice que tiene que ser igual o MENOR de 0; aunque gráficamente está bien puesto: $S \geq 0$.

Guía de estudio - GES2

DATOS AGRUPADOS

Guía de estudio 2

Presentación

Esta segunda Guía de Estudio (**GES_2**) pretende orientar el estudio de los datos cuando éstos, una vez obtenidos, ya han sido transformados para hacer algún tipo de agrupación que facilite su estudio. **Esta GES_2 es un material añadido a los Módulos 1 al 4.**

Contiene el siguiente material:

1. Datos agrupados: frecuencias, intervalos y *marcas de clase*
2. El cálculo de la media, moda, mediana y varianza con datos agrupados.
3. Ejemplos de cálculo de los estadísticos básicos de dispersión con datos agrupados

Calendario: la temporización de la **GES_2** será la prevista en el Plan Docente.

Objetivos

Con el estudio de esta **GES_2** se pretende que el estudiante alcance los siguientes objetivos:

1. Introducirse en el estudio de los datos agrupados.

2. Ser capaz de utilizar los datos agrupados dentro de los diferentes apartados del curso.

Contenidos

1. DATOS AGRUPADOS: FRECUENCIAS, INTERVALOS Y *MARCAS DE CLASE*

1.1 Frecuencias

Dado que la estadística tiene múltiples aplicaciones, en especial en el campo de las ciencias sociales y las ciencias de la observación; a la hora de analizar una cierta variable estadística nos podemos encontrar con variables de diferentes tipos, que nos aporten más o menos información y con un tratamiento analítico casi bien intuitivo o un poco más complejo.

Si de la variable que pretendemos analizar no se dispone de demasiadas observaciones y el número de valores que toma la variable tampoco es muy elevado, el tratamiento estadístico se simplifica mucho, es muy intuitivo y sigue el procedimiento que hemos visto hasta ahora.

Pero a menudo las cosas no son tan sencillas. Cuando el investigador quiere estudiar la realidad, se encuentra muy a menudo con variables, que si bien toman un número reducido de valores, se dispone de un gran número de observaciones. Eso pasa cuando queremos conocer determinadas características de la población: número de hijos por pareja en una cierta ciudad, número de idiomas que habla una persona, número de bibliotecas en cada municipio de menos de 25.000 h, etc. En estas situaciones, lo que se hace es agrupar los resultados de las observaciones en una tabla de frecuencias del tipo:

x_i	$\mathbf{n_i}$
x_1	n_1
x_2	n_2
x_3	n_3

--	--
--	--
xk	nk

x_i = valor que toma la variable observada x.

n_i = frecuencia absoluta de x.

En estas distribuciones, como podremos comprobar, el número de valores diferentes de la variable x es inferior al número total de observaciones de la variable.

Frecuencia absoluta (n_i)

Denominamos frecuencia absoluta de un valor x_i de la variable estadística observada, al número de veces que aparece repetido este valor en el conjunto de las observaciones.

Las frecuencias relativas

A la hora de trabajar con datos estadísticos, a veces nos encontramos con que a la distribución de frecuencias anterior se le añade otra columna que recoge las frecuencias relativas de la variable, adoptando esta forma:

x_i	**n_i**	**fin**
x_1	n_1	f_1
x_2	n_2	f_2
x_3	n_3	f_3
--	--	--
--	--	--
xk	nk	fk

Frecuencia relativa (f_{in})

Denominamos frecuencia relativa de un valor x_i de la variable estadística observada, al cociente entre la frecuencia absoluta de un valor y el número total de observaciones realizadas. A menudo, este valor se da en tantos por ciento.

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

Lógicamente, y como veremos más adelante, la suma de todas las frecuencias relativas será igual a la unidad.

Frecuencia absoluta acumulada (N_i)

Si tenemos una distribución en donde los valores han sido ordenados de forma creciente, denominamos frecuencia absoluta acumulada hasta un valor x_i de la variable estadística observada, al resultado de sumar las frecuencias absolutas de los valores anteriores o iguales a él.

$$N_i = \sum_{j=1}^{j=i} n_j$$

Frecuencia relativa acumulada (F_{in})

Denominamos frecuencia relativa acumulada en el punto x_i de la variable estadística observada, al cociente entre la frecuencia absoluta acumulada y el número total de observaciones realizadas. A menudo, este valor se da en tantos por ciento.

$$F_i = \frac{N_i}{N}$$

También podríamos decir que la frecuencia acumulada de un valor s_{ha} es la suma de todas las frecuencias desde el primer valor x_i hasta el valor sh_a incluido.

$$F_{in} = \sum f_i$$

Como decíamos más arriba, la suma de las frecuencias relativas será igual a la frecuencia acumulada final y ambas serán iguales a 1.

N1 $_{= n1}$
N2 = $_{n1 + n2}$
N3 = n1 $_{+ n2 +n3}$
.....
.....
Nk = n1 + n2 $_{+.....+ nk =}$ N

Pero como: $\sum f_i = \sum \frac{n_i}{N} = \frac{\sum n_i}{N} = \frac{N}{N} = 1$

Tenemos que: $F_{in} = \sum f_i = 1$

Así pues, podemos formar una tabla con dos columnas más, quedando los datos -en este caso ya en forma de tabla de cinco columnas-, de la siguiente manera:

$\mathbf{x_i}$	$\mathbf{n_i}$	$\mathbf{N_i}$	$\mathbf{f_{in}}$	$\mathbf{F_{in}}$
x_1	n_1	N_1	f_1	F_1
x_2	n_2	N_2	f_2	F_2
x_3	n_3	N_3	f_3	F_3
--	--	--	--	--
xk	nk	N_k	fk	F_k

x_i = valor que toma la variable observada.
n_i = frecuencia absoluta de x.
N_i = frecuencia absoluta acumulada.
f_{in} = frecuencia relativa de x.
F_{in} = frecuencia relativa acumulada.

Ejemplo: Suponemos que la siguiente tabla recoge los datos sobre el nivel de conocimiento de idiomas del personal que trabaja en las bibliotecas públicas de una cierta comunidad autónoma y queremos calcular las frecuencias relativas, absolutas y acumuladas

Número de idiomas que conoce	**Número de personas**

conoce	
2	130
3	65
4	28
5 o más	15

Tabla de cálculos:

x_i	n_i	N_i	f_{in}	F_{in}
2	130	130	0'5462	0'5462
3	65	195	0'2731	0'8193
4	28	223	0'1176	0'9369
5	15	238	0'0630	≈ 1
	Total = 238		Total = 1	

A partir de estos cálculos podemos decir que:

- el 11 '76% conoce 4 idiomas
- el 81 '93% conoce 2 o 3 idiomas
- que 223 personas conocen hasta 4 idiomas
- etc., etc.

1.2. Los Intervalos

A veces, trabajamos con variables que, indistintamente de que el número de observaciones sea muy grande o muy pequeño, la variable coge un número importante de valores, lo cual hace necesario llevar a término algún tipo de agrupamiento de estos valores que reduzca el tamaño de la distribución y permita un tratamiento más cómodo de los datos observados.

En estos casos el procedimiento más habitual es agrupar los datos en un número razonable de intervalos. La amplitud de este intervalo la puede fijar el investigador, sabiendo que cuanto mayores sean los intervalos, menos información nos dan y más fácil es su tratamiento.

Amplitud del intervalo = (Valor mayor de los datos no agrupados - valor más pequeño de los datos no agrupados)/número de clases deseado.

Ejemplo 1: A un grupo de 24 personas, les preguntamos cuántas veces al año van al cine. De esta consulta obtenemos las siguientes respuestas:

5 6 30, 38, 40 42 60 72 19 26 18 28

30 9 37 20 37 36 50 28 20 56 79 18

Si queremos trabajar con un número de 8 intervalos o clases, haríamos:

Intervalo $= \frac{79-5}{8} = 9'25$.

De manera que podemos coger intervalos de 10 unidades. Y tendríamos:

Las veces que van al cine al año	Número de personas
0-9	3
10-19	3
20-29	5
30-39	6
40-49	2
50-59	2
60-69	1
70-79	2
	Total = 24

Cuando nos encontramos delante de un número importante de datos o de observaciones y éstos están muy dispersos, nos puede interesar que los intervalos no sean regulares para tener un número limitado de intervalos y que todos ellos contengan alguna observación.

En estas situaciones, los intervalos más anchos o mayores se utilizan por aquellos rangos de valores con, relativamente, pocas observaciones.

Ejemplo 2: Observamos las edades de 36 víctimas de accidentes de tráfico durante un cierto periodo y encontramos los siguientes datos:

22	24	11	5	0'8	24	26	28	23	22	40	49
25	27	35	27	29	20	26	45	70	19	20	24
80	72	4	25	17	30	26	20	38	40	18	25

si miramos el rango de esta distribución, vemos que en los extremos hay muy pocas observaciones, de manera que parece razonable que los intervalos no tengan todos la misma anchura, que no sean regulares, sino que los de los extremos sean mayores. Una posibilidad sería dividir el rango entre los siguientes intervalos:

Edad de las víctimas	Número de personas
0-14	4
15-19	3
20-24	9
25-29	10

30-39	3
40-49	4
50-64	-
65-79	2
80 o más	1
	Total = 36

Y de esta manera tendríamos unos intervalos en los extremos que son más anchos que los que hay en medio y que recogen las observaciones situadas en los extremos de la distribución.

1.3. Marcas de clase

El hecho de trabajar con intervalos hace necesario la introducción del concepto de *marca de clase*, la cual, sería el punto medio del intervalo.

Marca de clase = $x_i = \frac{L_{i-1} + L_i}{2}$

L_e: límite superior del intervalo o frontera superior
L_{i-1}: límite inferior del intervalo o frontera inferior

Cuando se trabaja con *marcas de clase* se puede cometer un error de agrupamiento, ya que la marca de clase aparecerá repetida n_i veces. En la realidad, esto no sólo no es cierto, sino que es posible que la *marca de clase* ni siquiera sea un valor de los observados. Esta pérdida de información, hace que si, por ejemplo, se quiere calcular la moda (Mo) no se pueda decir que este estadístico toma un determinado valor, sino que tenemos que decir que cae dentro de un cierto intervalo.

Una vez que hemos establecido las marcas de clase, el tratamiento de la información con el fin de analizar los datos de que disponemos, se hace esencialmente igual que cuando trabajábamos con frecuencias, cogiendo como valor de todo el intervalo su *marca de clase*.

Los intervalos en que se han agrupado los datos para simplificar el trabajo a la hora de analizarlos, pueden ser construidos por el propio investigador o puede que le vengan dados con unas ciertas características que él no puede variar. Así, nos podemos encontrar a veces con unos intervalos que pueden ser abiertos por alguno de los extremos. Cuando se produce esto, habitualmente son abiertos por la izquierda.

Si, por ejemplo, forman intervalos por una variable como podría ser la edad de las personas, es habitual dejar de establecer intervalos por encima de una cierta edad: p.e. 80 años; en este caso, se cerrará la serie con un intervalo del tipo: (80+) o (80 o más), que recogerá a todas las personas de más de 80 años.

Cuando los intervalos se solapan, algunos autores hablan *de intervalos semiabiertos*, representados por la notación [a - b), donde el intervalo así establecido, contendrá siempre el valor coincidente con su límite inferior pero no contendrá el valor coincidente con su límite superior.

Ejemplo:

Intervalo	Valores dentro del intervalo
[10-20)	11, 18, 15, 10, 19
[20-30)	20, 23, 28, 27, 27, 21
[30-40)	30, 38, 33, 32, 32,
[40-50)	41, 45, 48

En el anterior ejemplo que hemos propuesto, los intervalos no se solapaban y cada intervalo contenía todos los valores comprendidos entre sus dos límites.

2. EL CÁLCULO DE LA MEDIA, MODA, MEDIANA Y VARIANZA CON DATOS AGRUPADOS.

2.1 Media aritmética

La media aritmética de un conjunto de observaciones es la suma de los valores que toman las observaciones dividido por el total de observaciones.

$$\overline{x} = \frac{x_1 n_1 + x_2 n_2 + ... + x_k n_k}{N} = \frac{\sum_{i=1}^{i=k} x_i n_i}{N} = \sum_{i=1}^{i=k} x_i f_i$$

2.2 Moda

La moda en un conjunto de números ordenados o de observaciones, es el valor que está afectado con una frecuencia mayor o el valor que más se repite, pero si ya nos dan los datos agrupados, sólo podremos decir dentro de qué intervalo se encuentra el valor modal. Como ya hemos visto antes, en las distribuciones puede no existir ninguna moda o haber más de una, en este último caso decimos que las distribuciones son *bimodales* o *multimodales*.

$$Mo = L_{i-1} + \frac{d_1}{d_1 + d_2}.a_i$$

Siendo:

$\mathbf{L_{i-1}}$ el límite inferior del intervalo que contiene la moda.
$\mathbf{a_y}$ amplitud del intervalo modal.
$\mathbf{d_1}$ la frecuencia de la clase modal menos la frecuencia de la clase inmediata precedente.
$\mathbf{d_2}$ la frecuencia de la clase modal menos la frecuencia de la clase inmediata posterior.

Nota

También se podría encontrar la Moda partiendo de las *marcas de clase* y tratando los datos como si no hubiera intervalos. Si seguimos este procedimiento perdemos alguna información a cambio de una simplificación del proceso. Para encontrar la moda de esta manera sólo haría falta una inspección ocular de los datos para ver qué valor tiene la frecuencia mayor.

2.3. Mediana

Si bien previamente habíamos dicho que la mediana, en un conjunto de observaciones o de números ordenados, es el valor central o la media aritmética de los dos valores centrales si el número de observaciones es par. Cuando tenemos los datos agrupados en frecuencias o intervalos, tenemos que incorporar nuevos elementos para poder encontrar la mediana y que el resultado encontrado sea representativo.

Así, si los datos están frecuenciados, para encontrar la mediana podemos seguir el siguiente procedimiento:

- Dividimos el número de observaciones N entre 2.
- Comprobamos que el número obtenido como resultado de la división, N/2, se encuentre en la tabla de frecuencias absolutas no acumuladas, N_i.
- Si no está, sería necesario comparar entre dos números de la referida tabla y la mediana sería aquel valor de la variable que se corresponde con el mayor de estos dos números. Eso pasa cuando tenemos un número impar de datos.
- Si el valor N/2 está dentro de la columna de las N_i, quiere decir que coincide con la frecuencia absoluta observada de algún valor x_i y en este caso cogeremos como mediana el punto medio. Eso pasa cuando tenemos un número par de datos.

En el caso de que los datos estuvieran agrupados en intervalos, la mediana se puede calcular según la siguiente fórmula

$$Me = L_{i-1} + \frac{\frac{N}{2} - N_{i-1}}{n_i}.c_1$$

Siendo:

L_{i-1} el límite inferior del intervalo que contiene la mediana
N el número total de observaciones
N_{i-1} el número total de observaciones acumuladas hasta el intervalo inmediatamente anterior a aquél que contiene la mediana.
n_i el número de observaciones incluidas en el intervalo que contiene la mediana.
c_i la anchura del intervalo que contiene la mediana
*** Sería el mismo procedimiento que seguiríamos para encontrar el segundo cuartil.**

2.4. Cuartiles

Los cuartiles son los valores que dividen en la muestra en cuatro partes iguales. Los cuartiles coincidirán con los percentiles*: P_{25}, P_{50} y P_{75}. El segundo cuartil (P_{50}) coincide con la mediana.

(*) Los deciles dividen la muestra en 10 partes iguales y el percentiles en 100 partes iguales.

Para calcular los cuartiles realizamos el mismo proceso que para encontrar la Mediana, pero en lugar de dividir el total de las observaciones por 2, lo hacemos por 4. Tal como decíamos antes, el segundo cuartil (Q_2) coincide con la Mediana.

$$Q_1 = L_{i-1} + \frac{1.\frac{N}{4} - N_{i-1}}{n_i}.c_1$$

$$Q_2 = L_{i-1} + \frac{2.\frac{N}{4} - N_{i-1}}{n_i}.c_1$$

$$Q_3 = L_{i-1} + \frac{3.\frac{N}{4} - N_{i-1}}{n_i}.c_1$$

Siendo:

L_{i-1} el límite inferior del intervalo que contiene el cuartil
N el número total de observaciones
N_{i-1} el número total de observaciones acumuladas hasta el intervalo inmediatamente anterior a aquél que contiene el cuartil.
n_i el número de observaciones incluidas en el intervalo que contiene el cuartil.
c_i la anchura del intervalo que contiene el cuartil

Rango intercuartílico (RI)

RI = $Q_{3/4}$ - $Q_{1/4}$

La varianza

La varianza, como sabemos, es la media aritmética de los cuadrados de las desviaciones respecto de la media

$$s^2 = \frac{\sum_{i=1}^{1=k} (x_i - \bar{x})^2 n_i}{N}$$

O también podemos utilizar ésta otra fórmula que simplifica mucho los cálculos:

$$s^2 = \frac{\sum_{i=1}^{i=k} x_i^2 - N\bar{x}^2}{N} = \frac{\sum_{i=1}^{i=k} x_i^2}{N} - \bar{x}^2$$

que nos permite encontrar la varianza como: la suma de los valores de la variable al cuadrado partido por el total de observaciones, menos la media al cuadrado.

** Aunque estamos acostumbrados a dividir por N-1 y no por N, con el fin de tener unas varianzas más esmeradas, en éste podemos dividir para N dado que si estamos trabajando con intervalos ya estamos perdiendo una parte de la información tal como hemos comentado más arriba y las varianzas, a pesar de ser correctas y válidas, ya no son tan exactas como si trabajáramos con datos desagrupados.

3. ALGUNOS EJEMPLOS

La mediana (y los cuartiles): datos agrupados en intervalos

La fórmula para calcular la mediana cuando los datos están agrupados en intervalos es la siguiente:

$$Me = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i}(L_{i+1} - L_i)$$

Donde $(L_{i+1} - L_i)$ es el intervalo donde se encuentra la mediana, f_{in} su frecuencia absoluta y F_{i-1} es la frecuencia acumulada del intervalo anterior a aquél donde se encuentra la mediana.

Vemos en un **ejemplo** de cómo se aplica esta fórmula.

Supongamos que tenemos la siguiente distribución del sueldo mensual de 25 trabajadoras de una biblioteca (en miles de pesetas: ¡sin haber entrado aún en los euros!):

Salario	f_{in}	F_{in}
[90, 100)	12	12
[100,110)	7	19
[110,129)	4	23
[120,130)	2	25

Primero vemos cuál es **el intervalo donde se encuentra la mediana**: será aquél cuya frecuencia acumulada es la primera que iguala o supera en n/2.

En nuestro caso, n/2 = 25/2 = 12'5

La primera frecuencia acumulada que supera este valor es F_i=19, que corresponde al intervalo (100,110]

Aplicamos la fórmula: $Me = L_i + \frac{\frac{n}{2} - F_{i-1}}{f_i}(L_{i+1} - L_i) = 100 - \frac{12'5 - 12}{7}(110 - 100) = 100'714$

…miles de pesetas

La mediana es el cuartil 2 (Q_2), podemos ver en las fórmulas del cuartil 1 y del cuartil 3 cómo se calculan éstos estadísticos cuándo tenemos datos agrupados.

Vemos un ejemplo:

A partir de la siguiente tabla:

xi	fin	Fin
(0-10]	8	8
(10-20]	12	20
(20-30]	10	30
(30-40]	14	44
(40-50]	21	65
(50-60]	16	81

(60-70] 9 90

Calculamos los cuartiles:

Tenemos 90 datos, un 25% de 90 se 22'5. El primer cuartil estará dentro del intervalo 20-30.

$$Q_1 = L_i + \frac{1.\frac{n}{4} - F_{i-1}}{f_i}.(L_{i+1} - L_i) = 20 + \frac{1.\frac{90}{4} - 20}{10}.10 = \mathbf{22'5}$$

Tenemos 90 datos, un 50% de 90 se 45. El segundo cuartil estará dentro del intervalo 40-50.

$$Q_2 = L_i + \frac{2.\frac{n}{4} - F_{i-1}}{f_i}.(L_{i+1} - L_i) = 40 + \frac{2.\frac{90}{4} - 44}{21}.10 = \mathbf{40'5}$$

Tenemos 90 datos, un 75% de 90 se 67'5. El tercer cuartil estará dentro del intervalo 50-60.

$$Q_3 = L_i + \frac{3.\frac{n}{4} - F_{i-1}}{f_i}.(L_{i+1} - L_i) = 50 + \frac{3.\frac{90}{4} - 65}{16}.10 = \mathbf{51'6}$$

La media: datos agrupados en intervalos

En muestras numerosas es habitual que los datos aparezcan repetidos. En este caso, para cada dato, desde x_1 hasta x_n, tendríamos la frecuencia absoluta de cada uno de estos datos, desde f_1 hasta f_k (de manera que f_{1+} $f_{2+...}$+ f_k = n).

La media aritmética de un conjunto de observaciones es la suma de los valores que toman las observaciones dividido por el total de observaciones.

$$\overline{x} = \frac{x_1 f_1 + x_2 f_2 + ... + x_k f_k}{n} = \frac{\sum_{i=1}^{i=k} x_i f_i}{n}$$

Ejemplo:

Suponemos que tenemos el siguiente número de libros prestados durante 10 días en una biblioteca:

Días	1	2	3	4	5	6	7	8	9	10
Libros	12	14	15	14	13	15	15	16	17	16

Si calculamos la media sin hacer una tabla de frecuencias, tendremos:

$$\overline{x} = \frac{12+14+15+14+13+15+15+16+17+16}{10} = 14\text{ '7 libros}$$

Si transformamos estos datos en una tabla de frecuencias absolutas, tendremos que

x_i	f_{in}	x_i f_{in}
12	1	12
13	1	13

14	2	28
15	3	45
16	2	32
17	1	17

$$\bar{x} = \frac{12\times1+14\times2+15\times3+16\times2+17\times1}{10} = 14'7$$

La varianza: datos agrupados en intervalos

Si los datos de la muestra son x_1, x_2...x_k y aparece cada uno de estos datos con sus respectivas frecuencias absolutas, f_1, f_2..., f_k (en que $\sum_{i=1}^{k} f_i$ = n), la expresión de la varianza se simplifica como sigue:

$$s^2 = \frac{\sum_{i=1}^{1=k} (x_i - \bar{x})^2 f_i}{N}$$

Ejemplo:

Si tenemos la siguiente tabla de frecuencias

x_i	f_{in}	$(x_i - \bar{x})^2 f_i$
12	7'29	
13	2'89	
14	0'98	
15	0'27	
16	3'38	
17	5'29	
Sumatorio	10	20'10

Entonces,

$s^2 = 2'10$

$s = \sqrt{2'10} = 1'42$

Bibliografía, materiales complementarios y enlaces de interés

- Como bibliografía complementaria se puede consultar la que figura en el Plan Docente.

ESTADÍSTICA

Módulos 5 y 6

Guía de estudio - GES3

LA DISTRIBUCIÓN NORMAL

- Presentación de la Guía de Estudio (GES)
- Objetivos
- Contenidos
- Bibliografía

Guía de estudio 3

Presentación

Esta tercera Guía de Estudio (**GES3**) pretende orientar el estudio de los contenidos de los Módulos 5 y 6. Contiene el siguiente material:

1. Una breve idea intuitiva del concepto y uso de la distribución normal.
2. Definición de variable aleatoria continua.
3. Definición de distribución de probabilidad normal y función de densidad de probabilidad.
4. La distribución normal estándar.

Materiales: para trabajar esta GES3 necesitáis los materiales básicos de la asignatura (módulos 5 y 6).

Calendario: la temporización de la **GES3** será la prevista en el Plan Docente.

Objetivos

Con el estudio de esta **GES** se pretende que el estudiante alcance los siguientes objetivos:

1. Servir de material de soporte en la introducción al concepto, uso y aplicación de la distribución normal.
2. Ser capaz de utilizar la distribución normal en su forma estandarizada.

Contenidos

1. Una breve introducción a la distribución normal

La distribución de probabilidad conocida como distribución normal es, por la cantidad de fenómenos que explica, la más importante de las distribuciones estadísticas.

A la distribución normal también se la conoce con el nombre de Campana de Gauss, porque si representamos las frecuencias de los valores de una variable en un diagrama, la curva resultante tendrá forma de campana.

Guía de estudio 3

Definición de variable aleatoria continua

Una **variable aleatoria continua** es aquélla que puede asumir un número infinito de valores dentro de un determinado rango.

Por ejemplo, en el intervalo [80,85], el peso de una persona podría ser 80.5, 80.52, 80.525 ... dependiendo de la precisión de la báscula.

2. Definición de distribución de probabilidad normal

La **Normal** es la distribución de probabilidad más importante. Multitud de variables aleatorias continuas siguen una distribución normal o aproximadamente normal.

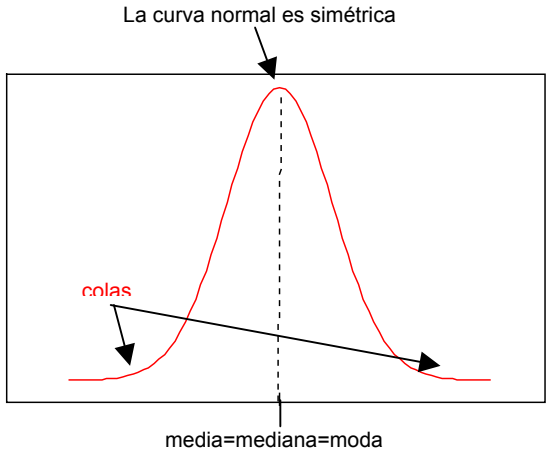
Una de sus características más importantes es que cualquier distribución de probabilidad, tanto discreta como continua, se puede aproximar por una normal bajo ciertas condiciones.

La distribución de probabilidad normal y la curva normal que la representa, tienen las siguientes características:

- La curva normal tiene forma de campana y un solo pico en el centro de la distribución. De esta manera, la media aritmética, la media y la moda de la distribución son iguales y se localizan en el pico. Así, la mitad del área bajo la curva se encuentra a la derecha de este punto central y la otra mitad está a la izquierda de dicho punto.

- La distribución de probabilidad normal es simétrica en torno a su media.

- La curva normal desciende suavemente en ambas direcciones a partir del valor central. Es asintótica, lo cual quiere decir que la curva se acerca cada vez más al eje X pero nunca llega a tocarlo. Es decir, las "colas" de la curva se extienden de manera indefinida en ambas direcciones.



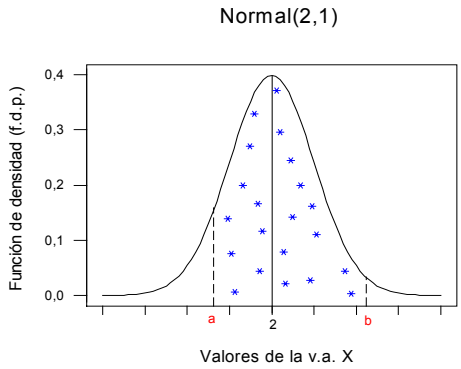
Para indicar que una variable aleatoria (a partir de ahora, v. a.) sigue una distribución normal de media μ y desviación estándar σ usaremos la expresión: $X \sim N(\mu,\sigma)$.

Guía de estudio 3

Definición de función de densidad de población

La probabilidad de que una variable aleatoria (v. a.) *X* tome un valor determinado entre dos números reales *a* y *b* coincide con el área cerrada por la función $f(x) = \frac{\exp^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}}{\sigma\sqrt{2\pi}}$ (**función de densidad de probabilidad**) entre los puntos *a* y *b*, es decir:

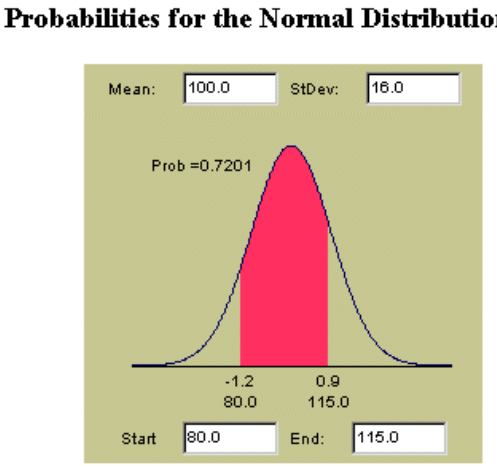
$$P(a \leq X \leq b) = \int_a^b f(x)dx :$$



Como hemos comentado anteriormente, observar que:

- La distribución normal es simétrica respecto de su media μ .
- El área total cerrada por f(x) vale 1, i.e.: $\int_{-\infty}^{+\infty} f(x)dx = 1$.
- Al ser X v.a. continua, P(X=a) = $\int_a^a f(x)dx$ = 0, $\forall \alpha \in P \Rightarrow$ P(X£a) = P(X<a).

Veamos, a través de una sencilla aplicación, este concepto de cómo la distribución normal representa un área bajo la curva. Para eso, podemos consultar el siguiente enlace: http://psych.vermell.edu/~mcclella/java/normal/html donde veremos, cambiando los valores de la media y la desviación estándar, así como los valores entre los cuales queremos calcular la probabilidad, a qué porción de espacio bajo la curva normal corresponde la probabilidad buscada.



2. La distribución normal estándar

Se observó que no hay una sola distribución de probabilidad normal, sino una "familia" de ellas. Como sabemos, cada una de las distribuciones puede tener una media (μ) y una desviación estándar distinta (σ). Por lo tanto, el número de distribuciones normales es ilimitado y sería imposible proporcionar una tabla de probabilidades para cada combinación de μ y σ.

Para resolver este problema, se utiliza un solo "miembro" de la familia de distribuciones normales, aquélla cuya media es 0 y desviación estándar 1 que es la que se conoce como **distribución estándar normal**, de forma que todas las distribuciones normales pueden convertirse en estándar, restando la media de cada observación y dividiendo por la desviación estándar.

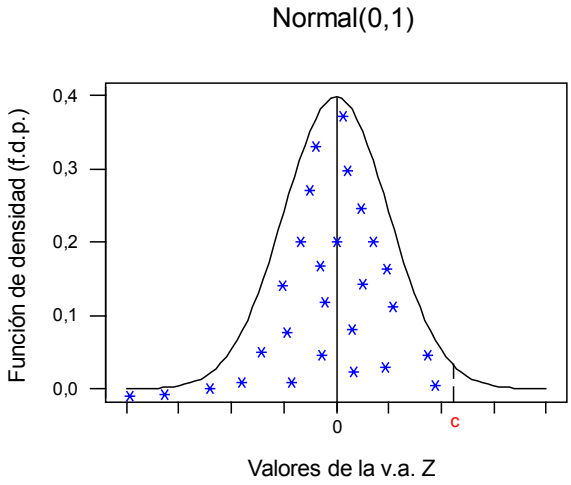
Primero, convertiremos la distribución real en una distribución normal estándar utilizando un valor nombrado Z, o **estadístico Z** que será la distancia entre un valor seleccionado, designado X, y la media μ, dividida por la desviación estándar σ.

Formalmente, si $X \sim N(\mu,\sigma)$, entonces la v. a. $Z = \frac{X-\mu}{\sigma}$ se distribuye según una normal por término medio 0 y desviación estándar 1, i.e.: $Z \sim N(0,1)$, que es la distribución llamada **normal estándar o tipificada**.

De esta manera, un valor Z mide la distancia entre un valor especificado de X y la media aritmética, en las unidades de la desviación estándar. Al determinar el valor Z utilizando la expresión anterior, es posible encontrar el área de probabilidad bajo cualquier curva normal haciendo referencia a la distribución normal estándar en las tablas correspondientes.

Así pues, para averiguar el área anterior utilizaremos la tabla que encontraremos al final de este apartado. Dicha tabla nos proporciona la probabilidad que la v.a. normal estándar *Z* tome un valor situado a la izquierda de un número *c*, i.e.: *P(Z<c)*. En otras palabras, esta tabla nos da el valor del área encerrada por *f(x)* entre -∞ y *c*.

Ejemplo 1:



a) P(Z<1,52) = {ver tabla} = 0,9357

b) P(Z>1,52) = {área total = 1} = 1 - P(Z<1,52) = 0,0643

c) P(0<Z<1,52) = P(Z<1,52) - P(Z<0) = {simetría} = 0,9357 - 0,5000 = 0,4357

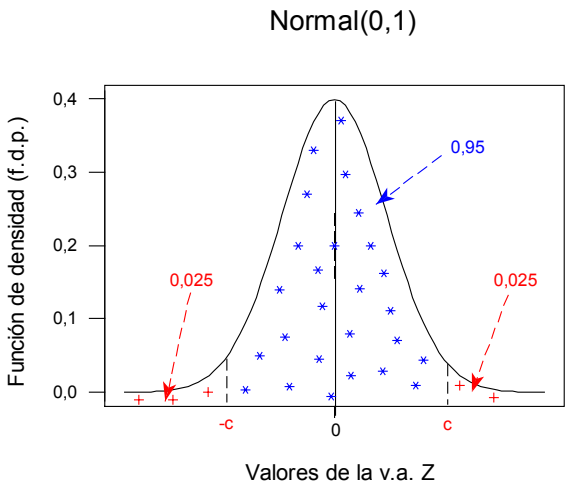
d) P(-2,1<Z<0) = P(Z<0) - P(Z<-2,1) = {ver tabla} = 0,5000 - 0,0179 = 0,4821

Por ejemplo:

a) z(0,25) = nº. que deja un área de 0,25 a su derecha = {tabla} »0,675

ya que P(Z<0,67) = 0,7486 y P(Z<0,68) = 0,7517.

b) Si queremos calcular un nº. real c tal que P(-c<Z<c) = 0,95, nos interesa encontrar z(0,025) {ver gráfico inferior}. Según la tabla, c = z(0,025) = 1,96 ya que P(Z<1,96) = 0,975 y P(Z<-1,96) = 0,025:



Suponemos ahora que X ~ N(100,16).

a) ¿Cuál es la probabilidad de que la variable X tome un valor entre 100 y 115?:

$$P(100<X<115)=P\left(\frac{100-100}{16}<\frac{X-\mu}{\sigma}<\frac{115-100}{16}\right)=P(0<Z<0{,}9375)\approx$$
$$\approx P(Z<0{,}94)-P(Z<0)=0{,}8264-0{,}5000=0{,}3264$$

b) ¿Cuál es la probabilidad de que X tome un valor mayor de 90?:

$$P(X>90)=P\left(\frac{X-\mu}{\sigma}>\frac{90-100}{16}\right)=P(Z>-0{,}63)=1-P(Z<-0{,}63)=$$
$$=1-0{,}2643=0{,}7357$$

3. El Teorema de Chebyshev

Si $X \sim N(\mu,\sigma)$, entonces:

a) $P(\mu-\sigma<X<\mu+\sigma) = 0,68$

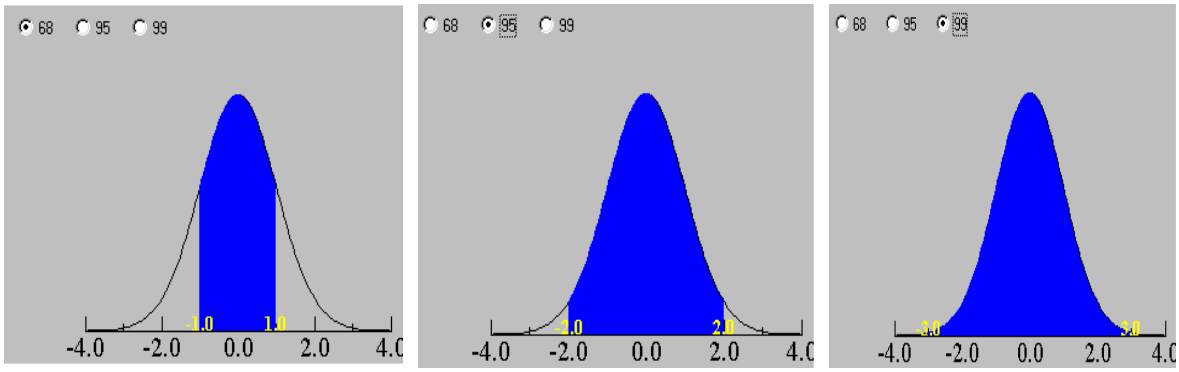
b) $P(\mu-2\sigma<X<\mu+2\sigma) = 0,95$

c) $P(\mu-3\sigma<X<\mu+3\sigma) = 0,97$

i.e., el 68% (aproximadamente) de los valores que tome la v.a. X estarán situados a una distancia de la media inferior a una desviación estándar. Análogamente, el 95% de los valores estarán situados a menos de 2 veces la desviación estándar, y un 99,7% de dichos valores se encontrarán dentro de un radio de 3 sigma.

Por lo tanto, para una distribución normal, la mayor parte de todos los valores caen a tres desviaciones estándar de la media.

Los *applets* que aparecen a continuación permiten identificar los respectivos porcentajes del área bajo la curva:



Guía de estudio 3

Área bajo la curva normal estándar: P(Z < z) donde Z sigue una distrib. N(0,1)

	segundo decimal de z									
z	**0,00**	**0,01**	**0,02**	**0,03**	**0,04**	**0,05**	**0,06**	**0,07**	**0,08**	**0,09**
-3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
-3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
-3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
-3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
-3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
-2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
-2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
-2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
-2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
-2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
-2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
-2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
-2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
-2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
-2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
-1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
-1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
-1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
-1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
-1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
-1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
-1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
-1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
-1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
-1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
-0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
-0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
-0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
-0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
-0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
-0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
-0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
-0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
-0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Bibliografía, materiales complementarios y enlaces de interés

- Como bibliografía complementaria se puede consultar la que figura en el Plan Docente.

ESTADÍSTICA

Módulos 7, 8, 9, 10 y 11

Guía de estudio – GES_4

MUESTREO Y DISTRIBUCIONES MUESTRALES

- Presentación de la Guía de Estudio (GES)
- Objetivos
- Contenidos
- Bibliografía

Presentación

Esta cuarta Guía de Estudio (**GES_4**) pretende orientar el estudio de los contenidos de los Módulos 7-11. Esta **GES** contiene el siguiente material:

1. Definición de muestra aleatoria simple y error en el muestreo
2. Definición del Teorema de las Distribuciones muestrales
3. Definición del Teorema Central del Límite

Materiales: para trabajar esta **GES4** necesitáis los materiales básicos de la asignatura (módulos 7-10).

Calendario: la temporización de la **GES4** será la prevista en el Plan Docente.

Objetivos

Con el estudio de esta **GES** se pretende que el estudiante alcance los siguientes objetivos:

1. Entender la necesidad de porqué en numerosas ocasiones una muestra es la única forma factible de conocer una población.
2. Explicar los métodos utilizados para seleccionar una muestra
3. Entender cómo se diseña una distribución muestral para la media de la muestra
4. Entender la importancia del Teorema Central del Límite, así como su aplicación

Contenidos

1. Una breve introducción a la distribución muestral

A menudo necesitamos estudiar las propiedades de una determinada población, pero nos encontramos con el inconveniente de que ésta es demasiado numerosa como para analizar a todos los individuos que la componen. Por tal motivo, recurrimos a extraer una muestra de la misma y a utilizar la información obtenida para hacer inferencias sobre toda la población. Estas estimaciones serán válidas sólo si la muestra tomada es "representativa" de la población.

Así, el muestreo es una técnica que utilizaremos para inferir algo respecto de una población mediante la selección de una muestra de esa población. Posteriormente, construiremos una distribución muestral de medias de la muestra, para entender la forma en que ésta tiende a agruparse en torno a la media de la población.

En muchos casos, el muestreo es la única manera de poder obtener alguna conclusión de una población, entre otras causas, por el coste económico y el tiempo empleado que supondría estudiar a todos los miembros de una población.

Definición de muestra aleatoria simple

En principio, podríamos distinguir dos tipos de muestra: la ***probabilística*** y la ***no probabilística***, en el sentido en que una **muestra probabilística** es una muestra seleccionada de tal forma que cada elemento de la población tiene la misma probabilidad de formar parte de la muestra.

De esta manera, si se utilizan métodos no probabilísticos, no todos los elementos de la población tienen la misma posibilidad de ser incluidos. En este caso, diríamos que los resultados están **sesgados**, lo cual quiere decir que tal vez los resultados de la muestra no sean representativos de la población.

Una forma de asegurarnos de que el subconjunto escogido es representativo de toda la población consiste en tomar una **muestra aleatoria simple**, la cual se caracteriza por:

1. Cada miembro de la población tiene la misma probabilidad de ser elegido, y

2. Las observaciones son elegidas siguiendo una secuencia aleatoria.

Error en el muestreo

Tras entender la importancia de escoger una muestra representativa de la población, veamos que para lograr esto, podemos seleccionar, por ejemplo, una muestra aleatoria simple de la población, pero es muy improbable que la media de la muestra sea idéntica a la media de la población.

De la misma manera, tal vez la desviación estándar u otra medición que se calcule con base a la muestra no sea igual al valor correspondiente de la población

Por tanto, es posible que existan ciertas diferencias entre los estadísticos de la muestra (como la media o la desviación estándar), y los parámetros de población correspondientes. A dicha diferencia se la conoce como **error de muestreo**.

2. Distribución muestral de la media de las muestras

Consistiría en una distribución de probabilidad de todas las medias posibles de las muestras de un tamaño de muestra dado.

Así pues, dada una población (a la cual representaremos por la v.a. X), podemos extraer de la misma k muestras, cada una de ellas de tamaño n. Para cada una de las k muestras podemos calcular un estadístico, p.e., la media de las n observaciones que la componen.

Así tendremos un total de k nuevos valores $\bar{x}_i, i = 1,...,k$. Podemos asociar estos valores a una nueva v.a. $\overline{X}$, cuya distribución llamaremos **distribución muestral**.

Una de las propiedades más importantes es la siguiente:

Teorema: Distribución de las medias muestrales

Sea *X* una v.a. **cualquiera** de media μ y desviación típica σ, entonces:

- Si consideramos **todas** las muestras aleatorias posibles, cada una de ellas de tamaño *n*, se cumplirá que $\mu_{\bar{x}} = \mu$ y $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.
- Además, si *X* sigue una distribución normal, $\overline{X}$ también será normal.

3. El Teorema Central del Límite

Sea *X* una v.a. **cualquiera** de media μ y desviación típica σ, entonces:

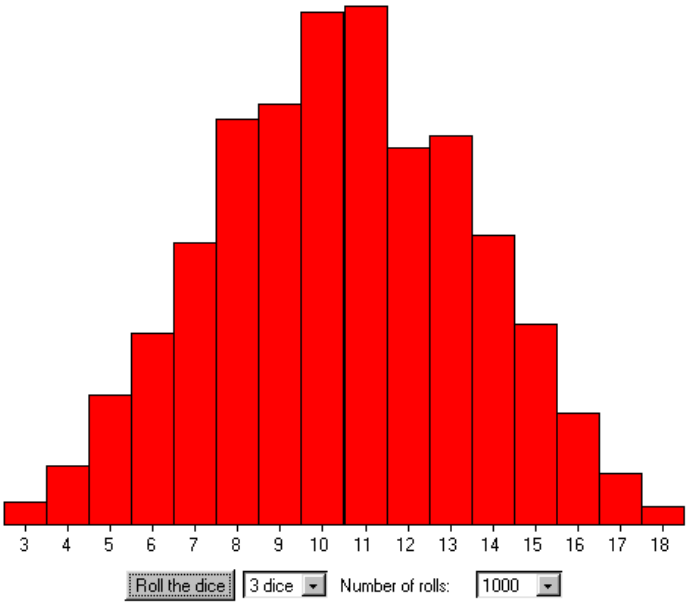
Si el tamaño muestral *n* es "suficientemente grande" (en la práctica suele valer *n>30*), la distribución de las medias muestrales se aproxima a la de una normal, i.e.:

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

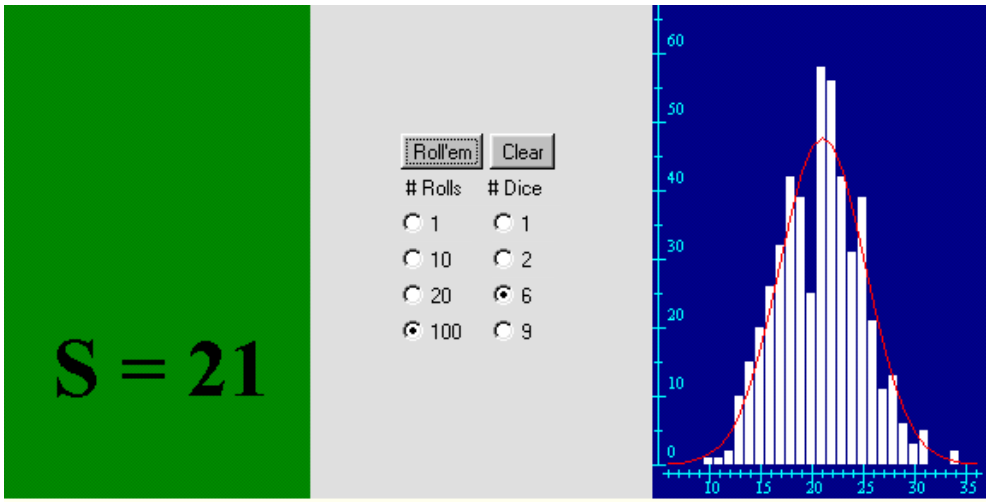
La importancia del TCL radica en que **sea cuál sea** la distribución de la población original (v.a. *X*), conforme el tamaño de las muestras (*n*) aumenta, la distribución de las medias se va aproximando a la de una normal (de la cual conocemos muchas propiedades).

Así, si la población tiene una distribución de probabilidad normal, entonces, para cualquier tamaño de muestra la distribución del muestreo de la media también tendrá una distribución normal. Si la distribución de la población es simétrica (pero no normal), se verá que surge la forma normal como lo establece el Teorema Central del Límite aún con muestras de al menos 30 para observar las características de normalidad.

Un ejemplo gráfico que muestra el Teorema Central del Límite, lo podemos encontrar en el siguiente enlace: http://www.unalmed.edu.co/~estadist/C.L.T/T_C_L.htm, de forma que cambiando el tamaño de la muestra veremos cómo va variando dicho gráfico, obtendremos representaciones similares a la siguiente:



Otro ejemplo similar al anterior, lo podemos encontrar en: http://www.ideamas.cl/cursoProb/javaEstat/central_limit_theorem/clt.html. Nuevamente, cambiando los datos veremos cómo la distribución resultante se va aproximando a una distribución normal:



Ejemplo 1:

El Presidente de una multinacional de telecomunicaciones, está preocupado por el número de teléfonos móviles defectuosos, producidos por su empresa. En promedio, 110 teléfonos al día son devueltos por este problema, con una desviación estándar de 64. ¿Cuál es la probabilidad de que en los próximos 48 días, el número medio de teléfonos devueltos diariamente sea menor de 120?

Tenemos a la variable X como la distribución de las devoluciones diarias de teléfonos móviles defectuosos. Se toma una muestra aleatoria de tamaño muestral 48 y como n es mayor que 30, podemos usar el teorema central del límite para poder afirmar que la distribución muestral se aproxima a la normal.

Para que se ordene la reorganización del proceso productivo, la probabilidad de que la media de teléfonos devueltos al día durante los próximos 48 días ha de ser menor que

0,8, por tanto, debemos calcular la probabilidad de que la media no sea mayor que 120,

$$P(\overline{X} \le 120) = P(\overline{X} \le 120) = P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{120-110}{64/\sqrt{48}}\right) = P(Z \le 1{,}08) \xrightarrow{TABLAS} 0{,}8599 = 0{,}8599$$

Observamos, pues, que existe una probabilidad de 0,86 de que no se devuelva en promedio más de 120 teléfonos al día durante los próximos 48 días.

Ejemplo 2:

Un supermercado financiero de Internet sabe que el número de operaciones diarias que realizan sus clientes sigue una distribución normal de media 120 y de desviación estándar de 34. Si se toma la muestra de días correspondiente al último mes de septiembre, ¿Cuál es la probabilidad de que la media diaria de operaciones que han realizado los clientes de este supermercado financiero on-line a lo largo del mes de septiembre, esté entre 110 y 140?

Tenemos a la variable X como la distribución del número diario de operaciones que llevan a cabo los clientes de este supermercado financiero en Internet. Se toma una muestra aleatoria de tamaño muestral 30, los días del último mes de septiembre, y como n es, en este caso igual a 30, podemos usar el TCL para poder afirmar que la distribución muestral obtenida se aproxima a una distr. normal.

$$P(110 < \overline{X} < 140) = P\left(\frac{110-120}{34/\sqrt{30}} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \le \frac{140-120}{34/\sqrt{30}}\right) = P(-1{,}61 < Z < 3{,}22) = P(Z < 3{,}22) - P(Z < -1{,}61) =$$

$$= P(Z < 3{,}22) - P(Z < -1{,}61) = 0{,}9994 - 0{,}0537 = 0{,}9457$$

En conclusión, podemos decir que existe una probabilidad del 94,57% de que el número medio de operaciones que hayan realizado los clientes de este supermercado financiero on-line durante el mes de septiembre esté entre 110 y 140.

Bibliografía, materiales complementarios y enlaces de interés

- Como bibliografía complementaria se puede consultar la que figura en el Plan Docente.

ESTADÍSTICA

Capítulos: 12, 13, 14 y 15

Guía de estudio – GES5

INTERVALOS DE CONFIANZA

- Presentación de la Guía de Estudio (GES)
- Objetivos
- Contenidos
- Bibliografía
- Fe de erratas

Presentación

Esta Guía de Estudio (**GES_5**) pretende orientar el estudio de los contenidos de los Capítulos: 12, 13, 14 y 15. Esta **GES_5** incorpora el siguiente material:

1. Un breve recordatorio del teorema central del límite (TCL).
2. La estimación de un parámetro estadístico.
3. La estimación de la media aritmética cuando conocemos la varianza poblacional.
4. La estimación de la media aritmética con una varianza poblacional desconocida.
5. La utilización de la "t" de Student como alternativa a la utilización de la distribución normal estándar. Se incorpora un ejemplo y la tabla de la t-Student.

Materiales: para trabajar esta **GES_5** se necesitan los materiales básicos de la asignatura (Capítulos: 12, 13, 14, y 15).

Calendario: la temporización de la **GES_5** será la prevista en el Plan Docente.

Objetivos

Con el estudio de la **GES_5** se pretende que el estudiante consiga los siguientes objetivos:

1. Avanzar en el conocimiento de la estimación de los parámetros estadísticos básicos: la media aritmética y la desviación estándar, especialmente utilizando los procedimientos adecuados en cada caso.
2. Conocer la utilidad y el uso de los **intervalos de confianza**, así como interpretar sus resultados.
3. Poder utilizar el estadístico "t-Student" como alternativa al estadístico "z" a la hora de hacer los cálculos del intervalos de confianza.

Contenidos

1. Recordatorio del TCL (Teorema Central del Límite)

RECORDATORIO DEL TCL

- **Teorema (Distribución de las Medias Muestrales):** Sea *X* una v.a. **cualquiera** de media μ y desviación típica σ, entonces:
  1. Si consideramos **todas** las muestras aleatorias posibles, cada uno de ellas de medida *n*, se verificará que $\mu_{\bar{x}} = \mu$ y $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.
  2. Además, si *X* sigue una distribución normal, $\overline{X}$ también será normal.
- **Teorema Central del Límite:** Sea *X* una v.a. **cualquiera** de media μ y desviación típica σ, entonces:

  Si la medida muestral *n* es "suficientemente grande" (en la práctica suele valer *n>30*), la distribución de las medias muestrales se aproxima a la de una normal, i.e.:

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

2. Estimación de un parámetro estadístico

ESTIMACIÓN DE UN PARÁMETRO ESTADÍSTICO

- En ocasiones, estaremos interesados en estimar el valor de algún parámetro poblacional (como la media μ o la desviación estándar σ), el cual desconocemos. Con el fin de realizar tal estimación, cogeremos una muestra de la población y calcularemos el parámetro muestral asociado ($\bar{x}$ para la media, *s* para la desviación estándar, etc.). El valor de este parámetro muestral será **la estimación puntual** del parámetro poblacional.
- Hay dos propiedades que son francamente deseables en cualquier estimador muestral: que sea sesgado y que tenga poca variabilidad:
  1. Un estimador es **sesgado** cuando la media de su distribución muestral asociada coincide con la media de la población. Eso sucede, por ejemplo, con el estimador $\bar{x}$, ya que $\mu_{\bar{x}} = \mu$.
  2. La **variabilidad** de un estimador viene determinada por el cuadrado de su desviación estándar. En el caso del estimador $\bar{x}$, su desviación estándar es $\sigma_{\bar{x}} = \sigma / \sqrt{n}$, también llamado **error estándar** de μ . Observar que cuanto mayor sea la medida de la muestra *n,* menor será la variabilidad del estimador $\bar{x}$ (y por lo tanto, mejor será nuestra estimación).
  - Supongamos una población *X* que sigue una distribución cualquiera con media μ y desviación estándar σ

1. Sabemos (por el TCL) que para valores grandes de n, la media muestral $\bar{x}$ sigue una distribución aproximadamente normal con media $\mu_{\bar{x}} = \mu$ y desviación estándar $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.
2. Por otra parte, el Teorema de Chebyshev nos decía que, en una distribución normal, aproximadamente un 95% de los datos estaban situados a una distancia inferior a dos desviaciones estándar de la media.

   De lo mencionado antes deducimos que: $P(\mu - 2\sigma_{\bar{x}} < \bar{x} < \mu + 2\sigma_{\bar{x}}) = 0{,}95$ i.e.,

   $$0{,}95 = P(\bar{x} < \mu + 2\sigma_{\bar{x}}) - P(\bar{x} < \mu - 2\sigma_{\bar{x}}) = P(\mu > \bar{x} - 2\sigma_{\bar{x}}) - P(\mu > \bar{x} + 2\sigma_{\bar{x}})$$

   i.e., $P(\bar{x} - 2\sigma_{\bar{x}} < \mu < \bar{x} + 2\sigma_{\bar{x}}) = 0{,}95$

   El resultado anterior nos da un método para calcular a partir de $\bar{x}$ y de σ un intervalo real, tal que la probabilidad de que la media de la población μ esté contenida en él es de 0,95.
3. Estos tipos de intervalos se llaman **intervalos de confianza** de un parámetro poblacional. El **nivel de confianza (1 - α)** del intervalo es la probabilidad de que éste contenga al parámetro poblacional. En el ejemplo anterior, el nivel de confianza era del 95% (i.e.,α = 0,05).

3. Estimación de la media aritmética conociendo la varianza poblacional

ESTIMACIÓN DE μ CON σ CONOCIDA

- Supongamos una población *X* (que sigue una distribución cualquiera), con media μ (desconocida) y desviación estándar σ conocida, se trata de encontrar un intervalo de confianza a nivel (1 - α) para μ
- **Supuesto:** $\bar{X}$ se distribuye según una normal.
- Recordatorio TCL: Si *X* se distribuye normalmente → $\bar{X}$ también lo hará. En caso contrario, necesitaremos coger una medida muestral *n* "grande" (generalmente, *n > 30* es suficiente).
- Bajo el supuesto anterior, el intervalo de confianza a nivel *(1 - α)* para la media poblacional μ viene dado por:

$$\bar{x} \pm z\left(\alpha/2\right) * \frac{\sigma}{\sqrt{n}}$$

  donde *zα(/2)* es el valor que, en una normal estándar, deja a su derecha un área de *α/2*. El **error máximo de estimación** es la mitad de la longitud del intervalo, i.e.: $E = z\alpha^{(/2)} * /n1/2.\sigma$
- Hay una relación inversa entre el error máximo de *estimación E* y la medida muestral *n* (si cogemos muestras mayores, el error máximo disminuye). Por otro lado, cuanto mayor sea el nivel de confianza *(1- α)*, mayor será la amplitud del intervalo (y por lo tanto, mayor será el error máximo de estimación). Así, si queremos aumentar el nivel de confianza sin incrementar la amplitud del intervalo (cosa deseable), tendremos que coger muestras de mayor medida.

4. Estimación de la media aritmética cuando no se conoce la varianza poblacional

ESTIMACIÓN DE μ CON σ DESCONOCIDA

1. Reparto una población *X* (que sigue una distribución cualquiera), con media μ y desviación estándar σ desconocidas, se trata de encontrar un intervalo de confianza a nivel *(1 - α)* para μ
2. En este caso no podemos utilizar el método anterior, sino que tendremos que aproximar el valor de σ por s (desviación muestral estándar), y utilizar la distribución t-Student con n-1 grados de libertad (siendo n la medida de la muestra escogida).
3. Las **t-Student** son una familia de distribuciones (una para cada valor del parámetro "Grado de Libertad") con las siguientes propiedades (para GL > 2):
   1. Las t-Student son simétricas por término medio 0 y varianza mayor que 1.
   2. Así como el GL aumenta, la distribución se va aproximando a una normal estándar.
4. **Supuesto:** $\overline{X}$ se distribuye según una normal.

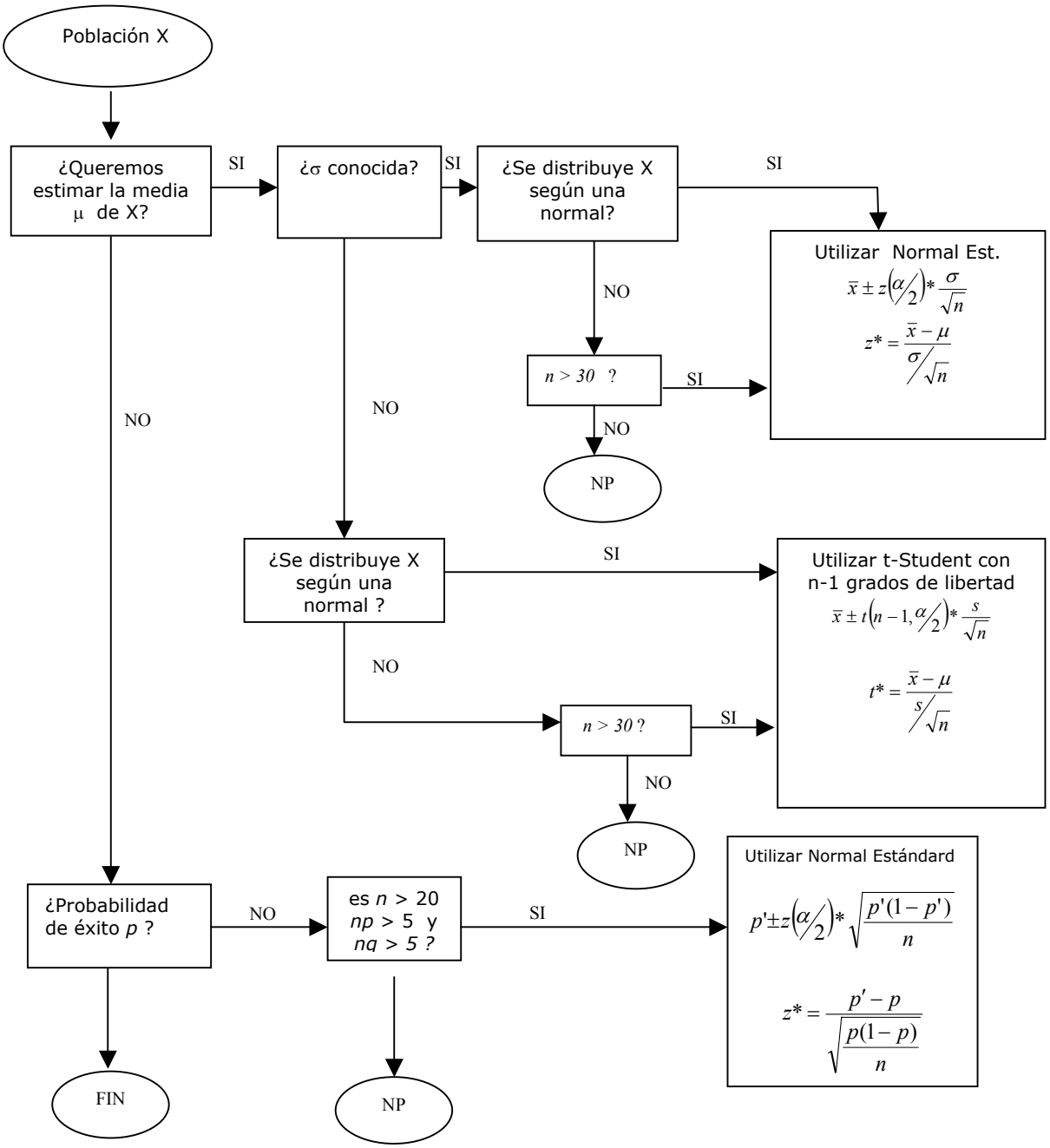
   Recordatorio TCL: Si *X* se distribuye normalmente → $\overline{X}$ también lo hará. En caso contrario, necesitaremos coger una medida muestral *n* "grande" (generalmente, *n > 30* es suficiente).
5. Bajo el supuesto anterior, el intervalo de confianza a nivel *(1- α)* para la media poblacional μ viene dado por:

$$\overline{x} \pm t\left(n-1, \alpha/2\right) * \frac{s}{\sqrt{n}}$$

   donde *t(n-1,α/2)* es el valor que, en una t-Student con *n-1* grados de libertad, deja a su derecha uno área de *α/2.* El **error máximo de estimación** es la mitad de la longitud del intervalo, i.e.:

   *E = t(n-1,/2) * s α/n1*/2.

INT. DE CONFIANZA Y CONTR. DE HIPÓTESIS (1 POBLACIÓN)



NP significa que tenemos que utilizar métodos No Paramétricos (fuera del contenido del curso).

5. La utilización de la "t" de Student

APUNTES DE REFUERZO A LA UNIDAD 14. (PÁGINA 92 DEL MÓDULO DIDÁCTICO)

En la práctica estadística casi nunca conocemos la desviación estándar (o la varianza) de la población con la que estamos trabajando sino que hace falta que lo estimemos a través de la desviación estándar (o la varianza) de una muestra representativa de la población.

Ya habéis visto cómo es posible estimar la desviación estándar de la población σ mediante la desviación estándar de la muestra s, la cual es la raíz cuadrada de la varianza de la muestra s^2, que se calcula según la fórmula:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \quad \text{ó} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Pero todos los cálculos que antes hacíamos con una distribución normal a través de las puntuaciones estandarizadas *z* cuando conocíamos los parámetros de la población estudiada, hará falta que ahora las hagamos con una distribución *t*, que tiene una forma similar a la normal pero es más dispersa que la curva normal. De hecho, la distribución de la *t* DEPENDE DEL TAMAÑO de la muestra que estemos utilizando, siendo más disperso cuanto más pequeña sea la muestra. En concreto, la distribución *t* se define en términos de n-1, es decir, el tamaño de la muestra menos 1. Son los llamados"grados de libertad" (mirad la expresión de la distribución *t de Student*).

EJEMPLO

Se está llevando a cabo un estudio entre una muestra de profesionales que se dedican a corregir artículos de diario para que pueden ser considerados como un grupo normativo. Se les pasa una prueba y se obtienen las siguientes puntuaciones (que se refieren al número de aciertos):

Individuo	Puntuación
1	19
2	32
3	25
4	18
5	31
6	27
7	23
8	20
9	30
10	26
11	21

12	27
13	31
14	25
15	29
16	19
17	30
18	26
19	17
20	24
21	31
22	23
23	18
24	30
25	29
26	22

Queremos calcular la media de esta muestra, y por eso aplicamos la fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{o} \quad \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

El número de individuos de la muestra, *n*, son 26. Es decir, X_i va desde X_1 hasta X_{26}, siendo X_1 igual en 19 y X_{26} igual en 22. En definitiva, $\sum_{i=1}^{n} x_i$ será la suma de todas las X_i desde la primera a la última 653.

La media de la muestra, $\bar{x}$, será, en consecuencia, 653 / 26 = 25'12.

Aplicando la fórmula de la varianza muestral obtenemos que s^2 = 22'67 y, en consecuencia, la desviación estándar de la muestra será s = 4'76.

En definitiva, la puntuación muestral de los correctores y correctoras es de 25'12 y la desviación estándar muestral de 4'76.

Y así podemos encontrar todos y cada uno de los indicadores que tenemos a la Unidad 14 del manual.

Así podemos calcular el error estándar de la media:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{4'76}{\sqrt{26}} = 0'93$$

Y el margen de error como $t_{\alpha/2,}$ n-1 * $s_{\bar{X}}$. Siendo α el error y n-1 los grados de libertad.

Si α = 0'10, α/2=0'05 y t0'05,25 = -1,708.

Si α = 0'05, $\alpha/2$=0'025 y t0'025,25 = -2 '060.

Si α = 0'01, $\alpha/2$=0'005 y t0'005,25 = -2 '787.

(en la Unidad 14 del manual leemos ...)

$t_{\alpha/2,\ n-1}$ es el valor de la distribución t (con n-1 grados de libertad) tal que el 100*(1-α)% del área está incluido entre $\pm$ $t_{\alpha/2,\ n-1}$, tendremos tres posibilidades, en consecuencia:

Si decidimos que α = 0'10, 100*(1-0'10)% = 90% del área buscada caerá entre $\pm$ 1,708.

Si decidimos que α = 0'05, 100*(1-0'05)% = 95% del área buscada caerá entre $\pm$ 2'060.

Si decidimos que α = 0'01, 100*(1-0'01)% = 99% del área buscada caerá entre $\pm$ 2'787.

En definitiva, el margen de error sería:

Con un nivel de confianza del 90%, el margen de error será de 0'93*1,708 = 1,58.

Con un nivel de confianza del 95%, el margen de error será de 0'93*2'060 = 1'92.

Con un nivel de confianza del 99%, el margen de error será de 0'93*2,787 = 2,59.

Y ahora ya podemos encontrar el intervalo de confianza para los diferentes márgenes de error (apartado 5 en la página 95 del manual), es decir,

Con un nivel de confianza del 90%, el intervalo de confianza será de 25'12 $\pm$ 1,58.

Con un nivel de confianza del 95%, el intervalo de confianza será de 25'12 $\pm$ 1'92.

Con un nivel de confianza del 99%, el intervalo de confianza será de 25'12 $\pm$ 2,59.

Vemos algunos ejemplos más:

1) Queremos conocer el valor de la distribución t que corta el 2 '5% de cada cola (es decir $\alpha/2$=0'025). En otras palabras, el valor de t que corta 0'025 para la cola derecha y 0'025 para la cola izquierda, dejando un 0'95 en el área central (es decir, un nivel de confianza del 95%).

Si el tamaño muestral es 10, los grados de libertad serán 9.

Miramos la tabla de la distribución t de Student: una probabilidad acumulada de 0'025 (que corresponde a la cola izquierda) con 9 grados de libertad resulta una t de -2 '262. Como la tabla da el punto hasta el cual el área especificada está incluida, obtenemos este número negativo.

Como la curva es simétrica, el otro 0'025 estará en una posición simétrica pero a la derecha, es decir, la otra t = +2'262.

Una manera alternativa de encontrar esta última t con las tablas sería sumar el 0'025 de la cola izquierda y los 0'95 del área central, es decir, podríamos buscar el área que corresponde a 0'025+0'95 = 0'975. Si miramos a la tabla qué t corresponde a una área de 0'975 con 9 grados de libertad encontraremos el mismo resultado, es decir, 2'262.

Remarcar que la tabla de la distribución t de Student que tenéis no es muy precisa, y su precisión es muy menor que la que se consigue con un programa estadístico.

2) Calculad el área en la cola de la distribución t con 24 grados de libertad a la derecha del valor 2'492.

Si buscamos en la tablas un valor de t = 2'492 (con 24 grados de libertad) encontramos que corresponde a uno área de 0'99. Pero esta área es hacia la izquierda, no cabe a la derecha; para encontrar el área hacia la derecha hará falta que calculemos su complementaria, es decir,

1-0'99 = 0'01.

3) Para una distribución de 55 grados de libertad, cuál es el área bajo la curva entre los valores -2 '004 y 2' 004.

A la izquierda de un valor T de -2 '004 (con 55 grados de libertad) hay una probabilidad acumulada de 0'025 (mirad tabla distribución t). Como tratamos con una distribución simétrica, a la derecha de 2'004 también tendremos una probabilidad de 0'025 (siempre con 55 grados de libertad). En definitiva, el área bajo la curva entre los valores T de -2 '004 y 2'004 (con 55 grados de libertad) será 1-0' 025-0 '025 = 1-0' 050 = 0'95.

4) Suponemos que el área entre dos puntos -T y +T es igual en 0'90. Cuál es el valor de T para una distribución t con (a) 9 grados de libertad?; (b) 100 grados de libertad?; (c) 1000 grados de libertad?

Si entre -T y + T hay un área de 0'90 querrá decir que el error es de 0'10, es decir, que en cada cola hay un error de 0'05.

Entonces:

9 grados de libertad, el valor de T será de ± 1'833.

100 grados de libertad, el valor de T será de ± 1'660.

1000 grados de libertad, el valor de T será de ± 1'646.

Bibliografía, materiales complementarios y enlaces de interés

- Como bibliografía complementaria podéis consultar la que figura en el Plan Docente.

Fe de erratas

- Página 85: A la actividad 12.3 dice: "En conexión con la pregunta 13.3" y tendría que decir "En conexión con la pregunta 12.2".
- Capítulo 16, Pág. 108, habla de los ejercicios 17.1, 17.2 y la figura 17.3. En todos los casos tendría que ser 16.1, 16.2 y figura 16.3.
- A la resolución de la actividad 13.2 tanto a los "sí", como a los no calcula los intervalos de confianza con el mismo error estándar, cuándo son diferentes. La solución correcta tendría que ser:

0'315 +/- 0'0120 = [0'303, 0'327]

Distribución *t-Student*

Distribución t-Student

Grados libertad v	PROBABILIDAD ACUMULADA						
	0,001	0,005	0,01	0,025	0,05	0,1	0,2
1	-318,289	-63,656	-31,821	-12,706	-6,314	-3,078	-1,376
2	-22,328	-9,925	-6,965	-4,303	-2,920	-1,886	-1,061
3	-10,214	-5,841	-4,541	-3,182	-2,353	-1,638	-0,978
4	-7,173	-4,604	-3,747	-2,776	-2,132	-1,533	-0,941
5	-5,894	-4,032	-3,365	-2,571	-2,015	-1,476	-0,920
6	-5,208	-3,707	-3,143	-2,447	-1,943	-1,440	-0,906
7	-4,785	-3,499	-2,998	-2,365	-1,895	-1,415	-0,896
8	-4,501	-3,355	-2,896	-2,306	-1,860	-1,397	-0,889
9	-4,297	-3,250	-2,821	-2,262	-1,833	-1,383	-0,883
10	-4,144	-3,169	-2,764	-2,228	-1,812	-1,372	-0,879
11	-4,025	-3,106	-2,718	-2,201	-1,796	-1,363	-0,876
12	-3,930	-3,055	-2,681	-2,179	-1,782	-1,356	-0,873
13	-3,852	-3,012	-2,650	-2,160	-1,771	-1,350	-0,870
14	-3,787	-2,977	-2,624	-2,145	-1,761	-1,345	-0,868
15	-3,733	-2,947	-2,602	-2,131	-1,753	-1,341	-0,866
16	-3,686	-2,921	-2,583	-2,120	-1,746	-1,337	-0,865
17	-3,646	-2,898	-2,567	-2,110	-1,740	-1,333	-0,863
18	-3,610	-2,878	-2,552	-2,101	-1,734	-1,330	-0,862
19	-3,579	-2,861	-2,539	-2,093	-1,729	-1,328	-0,861
20	-3,552	-2,845	-2,528	-2,086	-1,725	-1,325	-0,860
21	-3,527	-2,831	-2,518	-2,080	-1,721	-1,323	-0,859
22	-3,505	-2,819	-2,508	-2,074	-1,717	-1,321	-0,858
23	-3,485	-2,807	-2,500	-2,069	-1,714	-1,319	-0,858
24	-3,467	-2,797	-2,492	-2,064	-1,711	-1,318	-0,857
25	-3,450	-2,787	-2,485	-2,060	-1,708	-1,316	-0,856
26	-3,435	-2,779	-2,479	-2,056	-1,706	-1,315	-0,856
27	-3,421	-2,771	-2,473	-2,052	-1,703	-1,314	-0,855
28	-3,408	-2,763	-2,467	-2,048	-1,701	-1,313	-0,855
29	-3,396	-2,756	-2,462	-2,045	-1,699	-1,311	-0,854
30	-3,385	-2,750	-2,457	-2,042	-1,697	-1,310	-0,854
35	-3,340	-2,724	-2,438	-2,030	-1,690	-1,306	-0,852
40	-3,307	-2,704	-2,423	-2,021	-1,684	-1,303	-0,851
45	-3,281	-2,690	-2,412	-2,014	-1,679	-1,301	-0,850
50	-3,261	-2,678	-2,403	-2,009	-1,676	-1,299	-0,849
55	-3,245	-2,668	-2,396	-2,004	-1,673	-1,297	-0,848
60	-3,232	-2,660	-2,390	-2,000	-1,671	-1,296	-0,848
70	-3,211	-2,648	-2,381	-1,994	-1,667	-1,294	-0,847
80	-3,195	-2,639	-2,374	-1,990	-1,664	-1,292	-0,846
90	-3,183	-2,632	-2,368	-1,987	-1,662	-1,291	-0,846
100	-3,174	-2,626	-2,364	-1,984	-1,660	-1,290	-0,845
120	-3,160	-2,617	-2,358	-1,980	-1,658	-1,289	-0,845
1000	-3,098	-2,581	-2,330	-1,962	-1,646	-1,282	-0,842

Distribución t-Student

Grados libertad v	PROBABILIDAD ACUMULADA						
	0,8	0,9	0,95	0,975	0,99	0,995	0,999
1	1,376	3,078	6,314	12,706	31,821	63,656	318,289
2	1,061	1,886	2,920	4,303	6,965	9,925	22,328
3	0,978	1,638	2,353	3,182	4,541	5,841	10,214
4	0,941	1,533	2,132	2,776	3,747	4,604	7,173
5	0,920	1,476	2,015	2,571	3,365	4,032	5,894
6	0,906	1,440	1,943	2,447	3,143	3,707	5,208
7	0,896	1,415	1,895	2,365	2,998	3,499	4,785
8	0,889	1,397	1,860	2,306	2,896	3,355	4,501
9	0,883	1,383	1,833	2,262	2,821	3,250	4,297
10	0,879	1,372	1,812	2,228	2,764	3,169	4,144
11	0,876	1,363	1,796	2,201	2,718	3,106	4,025
12	0,873	1,356	1,782	2,179	2,681	3,055	3,930
13	0,870	1,350	1,771	2,160	2,650	3,012	3,852
14	0,868	1,345	1,761	2,145	2,624	2,977	3,787
15	0,866	1,341	1,753	2,131	2,602	2,947	3,733
16	0,865	1,337	1,746	2,120	2,583	2,921	3,686
17	0,863	1,333	1,740	2,110	2,567	2,898	3,646
18	0,862	1,330	1,734	2,101	2,552	2,878	3,610
19	0,861	1,328	1,729	2,093	2,539	2,861	3,579
20	0,860	1,325	1,725	2,086	2,528	2,845	3,552
21	0,859	1,323	1,721	2,080	2,518	2,831	3,527
22	0,858	1,321	1,717	2,074	2,508	2,819	3,505
23	0,858	1,319	1,714	2,069	2,500	2,807	3,485
24	0,857	1,318	1,711	2,064	2,492	2,797	3,467
25	0,856	1,316	1,708	2,060	2,485	2,787	3,450
26	0,856	1,315	1,706	2,056	2,479	2,779	3,435
27	0,855	1,314	1,703	2,052	2,473	2,771	3,421
28	0,855	1,313	1,701	2,048	2,467	2,763	3,408
29	0,854	1,311	1,699	2,045	2,462	2,756	3,396
30	0,854	1,310	1,697	2,042	2,457	2,750	3,385
35	0,852	1,306	1,690	2,030	2,438	2,724	3,340
40	0,851	1,303	1,684	2,021	2,423	2,704	3,307
45	0,850	1,301	1,679	2,014	2,412	2,690	3,281
50	0,849	1,299	1,676	2,009	2,403	2,678	3,261
55	0,848	1,297	1,673	2,004	2,396	2,668	3,245
60	0,848	1,296	1,671	2,000	2,390	2,660	3,232
70	0,847	1,294	1,667	1,994	2,381	2,648	3,211
80	0,846	1,292	1,664	1,990	2,374	2,639	3,195
90	0,846	1,291	1,662	1,987	2,368	2,632	3,183
100	0,845	1,290	1,660	1,984	2,364	2,626	3,174
120	0,845	1,289	1,658	1,980	2,358	2,617	3,160
1000	0,842	1,282	1,646	1,962	2,330	2,581	3,098

Capítulos: 16, 17 y 18

Guía de estudio - GES_6

CONTRASTE DE HIPÓTESIS

- Presentación de la Guía de Estudio (GES)
- Objetivos
- Contenidos
- Bibliografía
- Fe de erratas

Presentación

Esta Guía de Estudio (**GES_5**) pretende orientar el estudio de los contenidos de los Capítulos: 16, 17, y 18 relacionados con el contraste de hipótesis.

Esta **GES_5** incorpora el siguiente material:

1. Contraste de hipótesis para una población.
2. Contraste de hipótesis para dos poblaciones.
3. El contraste de hipótesis paso a paso.
4. Algunos ejemplos de contraste de hipótesis.

Materiales: para trabajar esta **GES_5** se necesitan los materiales básicos de la asignatura (Capítulos: 16, 17, y 18).

Calendario: la temporización de la **GES_5** será la prevista en el Plan Docente.

Objetivos

Con el estudio de la **GES_5** se pretende que el estudiante consiga los siguientes objetivos:

1. Introducir al estudiante en el conocimiento del contraste de hipótesis, tanto con respecto a una sola muestra como a la comparación entre muestras diferentes.
2. Conocer y entender los nuevos conceptos como: hipótesis nula e hipótesis alternativa, nivel de significación, significación estadística ...
3. Conocer el concepto de error de tipo I y error tipo II, y entender tanto su significado como su interrelación.

Contenidos

1. Contraste de hipótesis para una población

INTRODUCCIÓN AL CONTRASTE DE HIPÓTESIS

- Un **contraste de hipótesis** es un proceso estadístico que permite escoger una hipótesis de trabajo de entre dos posibles y antagónicas. El contraste empieza con la formulación de dos hipótesis sobre el valor de algún parámetro poblacional, siendo ambas incompatibles (si una es cierta, la otra necesariamente tiene que ser falsa). Supondremos cierta una de ellas, en la que nombraremos **hipótesis nula H_0**, y trataremos de determinar hasta qué grado las observaciones registradas son coherentes con H_0. Sólo en caso de que haya fuertes indicios de incompatibilidad entre el supuesto de que H_0 sea cierto y los datos obtenidos empíricamente, descartaremos H_0 como hipótesis de trabajo y, en su lugar, cogeremos como cierta la **hipótesis alternativa H_1**. Dos ejemplos de contrastes de hipótesis serían:

$$(i)\ \begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases} \qquad (ii)\ \begin{cases} H_0 : \sigma = 2{,}5 \quad (\leq) \\ H_1 : \sigma > 2{,}5 \end{cases}$$

Contraste Bilateral ($\neq$) Contraste Unilateral ($>$)

- En el siguiente esquema se representan las cuatro combinaciones posibles (en función de la decisión de que cogemos y de la certeza o no de la hipótesis nula) de todo contraste de hipótesis:

DECISIÓN ELEGIDA	**Hipótesis Nula H_0**	
	Verdadera	**Falsa**
No descartar H_0	Decisión correcta de tipo A **Probabilidad 1-α**	Error de tipo II **Probabilidad β**
Descartar H_0	Error de tipo I Probabilidadα	Decisión correcta de tipo B **Probabilidad 1-β**

Tendremos una **decisión correcta de tipo A** cuando hayamos optado por no descartar la hipótesis nula y resulte que ésta es cierta. Por su parte, una **decisión correcta de tipo B ocu**rrirá cuando hayamos decidido descartar la hipótesis nula y resulte que esta era falsa. Hablaremos de **error de tipo I** cuando hayamos descartado la hipótesis nula siendo ésta cierta (error que se considera como muy grave). Finalmente, ocurrirá un **error de tipo II** cuando hayamos optado por no descartar la hipótesis nula y resulte que ésta es falsa.

Dado que descartaremos o no la hipótesis nula a partir de muestras obtenidas (es decir, no dispondremos de información completa sobre la población), no será posible garantizar que la decisión tomada sea la correcta. Lo que sí podremos hacer es controlar la probabilidad de cometer un error. Denotaremos para α el **nivel de significación** o probabilidad de cometer un error de tipo I, y para β la probabilidad de cometer un error de tipo II. Con el fin de controlar ambos errores, los asignaremos probabilidades "pequeñas" (usualmente de 0,01 o 0,05). Llamaremos **potencia del contraste** en 1-β, ya que este número es la probabilidad de rechazar la hipótesis nula siendo ésta falsa. Es fundamental hacer notar en este punto que α, β y la medida muestral *n* están interrelacionados, de manera que si hacemos disminuir cualquiera de ellos alguno de los dos restantes tendrán que aumentar. Así, p.e., si queremos coger a un α menor deberemos aceptar que aumente β o bien incrementar la medida de la muestra n.

Finalmente, llamaremos **estadístico de contraste** en una v.a. calculada a partir de las observaciones muestrales, la cual se utiliza conjuntamente con un criterio de decisión (establecido a priori) para determinar si tenemos que descartar o no la hipótesis nula.

CONTRASTES SOBRE μ CON σ CONOCIDA

- Reparto una población *X* (que sigue una distribución cualquiera), con mediaμ (desconocida) y desviación estándarσ conocida, se trata de contrastar alguno de los tres tests siguientes:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} \quad \text{o bien} \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \quad \text{o bien} \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

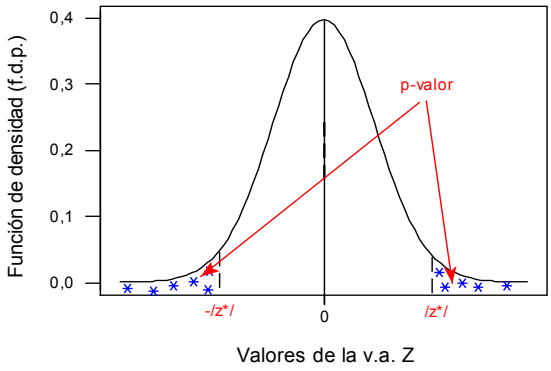
- **Supuesto:** $\overline{X}$ se distribuye según una normal.

  Recordatorio TCL: Si *X* se distribuye normalmente → $\overline{X}$ también lo hará. En caso contrario, necesitaremos coger una medida muestral *n* "grande" (generalmente, *n > 30* es suficiente).

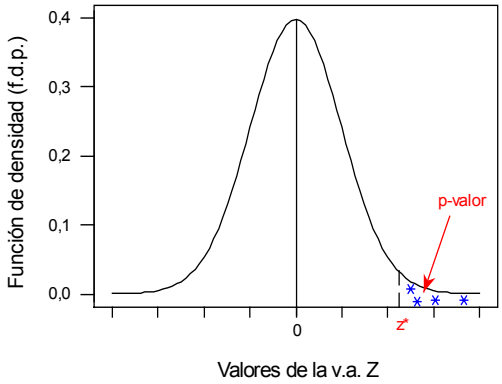
- **Estadístico de contraste:** $z^* = \dfrac{\overline{x} - \mu}{\sigma / \sqrt{n}} \approx N(0,1)$.

- Definimos el **p-valor** como la probabilidad de que, suponiendo cierta H_0, el estadístico de contraste coja un valor al menos tan extremo como lo que se obtiene a partir de las observaciones muestrales, i.e., el p-valor es el área de la cola de la distribución (o colas si el test es bilateral) definida a partir del estadístico de contraste:
  - Si H_1 contiene ">"$\Rightarrow$ *p-valor = P(Z>z*).*
  - Si H_1 contiene "<"$\Rightarrow$ *p-valor = P(Z<z*).*
  - Si H_1 contiene $\neq$"$\Rightarrow$ *p-valor = P(Z< -/z* ó Z>z/*)/ = 2 P(Z< /z/*) .*

P-valor cuando el test es bilateral



P-valor cuando H1 contiene ">"



El p-valor nos proporciona el grado de credibilidad de la hipótesis nula: si el valor de *p* es "muy pequeño" (inferior a 0,001), significaría que la hipótesis nula es del todo increíble (en base a las observaciones obtenidas), y por lo tanto la descartaríamos; si el valor de *p* está entre 0,05 y 0,001 significaría que hay fuertes evidencias en contra de la hipótesis nula, por lo que la rechazaríamos o no en función del valor que hubiéramos asignado (a priori) a α. Finalmente, si el valor de *p* es "grande" (superior a 0,05), no se tendrían motivos suficientes como para descartar la hipótesis nula, por lo que la cogeríamos como cierta.

- **Criterio de decisión:** Descartaremos H_0 si **p-valor**$\leq\alpha$ (normalmente α = 0,05).

CONTRASTES SOBRE μ CON σ DESCONOCIDA

- Reparto una población *X* (que sigue una distribución cualquiera), con media μ y desviación estándarσ desconocidas, se trata de contrastar alguno de los tres tests siguientes:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases} \quad \text{o bien} \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases} \quad \text{o bien} \begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

- **Supuesto:** $\overline{X}$ se distribuye según una normal.

  Recordatorio TCL: Si *X* se distribuye normalmente → $\overline{X}$ también lo hará. En caso contrario, necesitaremos coger una medida muestral *n* "grande" (generalmente, *n > 30* es suficiente).

- **Estadístico de contraste:** $t^* = \dfrac{\overline{x} - \mu}{s/\sqrt{n}} \approx t - Student(n-1)$.

- **Criterio de decisión:** Descartaremos H_0 si **p-valor** $\leq \alpha$ (normalmente α = 0,05).

- **Ejemplo de cómo calcular el p-valor usando la tabla de la t-Student:**

  Suponemos que estamos en un contraste unilateral donde H_1 contiene "<" con n =7, y que el estadístico de contraste obtenido es t* =-2,80. Vamos a la tabla, en la fila de 6 grados de libertad y, teniendo en cuenta que el área a la izquierda de -2,8 es la misma que el área a la derecha de 2,8, buscamos el valor 2,8 (o lo que más se aproxime a ello) en la mencionada fila. Aunque 2,8 no aparece, sabemos que es entre 2,4469 y 3,1427. El área la derecha de estos dos valores es 0,025 y 0,01 respectivamente. Por lo tanto, ya sabemos que el área a la derecha de 2,8 es entre 0,025 y 0,01. Se sigue pues que el p-valor estará entre 0,025 y 0,01.

¡Atención!: si el contraste es bilateral (aparece un $\neq$ en H_1), el p-valor asociado a t* =-2,8 será dos veces el área a la derecha de 2,8 (o dos veces el área a la izquierda de -2,8).

CONTRASTES SOBRE LA PROB. DE ÉXITO p

- Suponemos que una población *X* con probabilidad de éxito *p* desconocida. Con el fin de estimar este parámetro, cogemos una muestra de medida *n* y definimos la **probabilidad muestral de éxito** como: *p' = número de éxitos observados / n.* Se tratará de contrastar alguno de los tres tests siguientes:

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases} \quad \text{o bien} \begin{cases} H_0 : p = p_0 \\ H_1 : p < p_0 \end{cases} \quad \text{o bien} \begin{cases} H_0 : p = p_0 \\ H_1 : p > p_0 \end{cases}$$

- **Supuesto 1:** La distribución de *X* es aproximadamente normal.

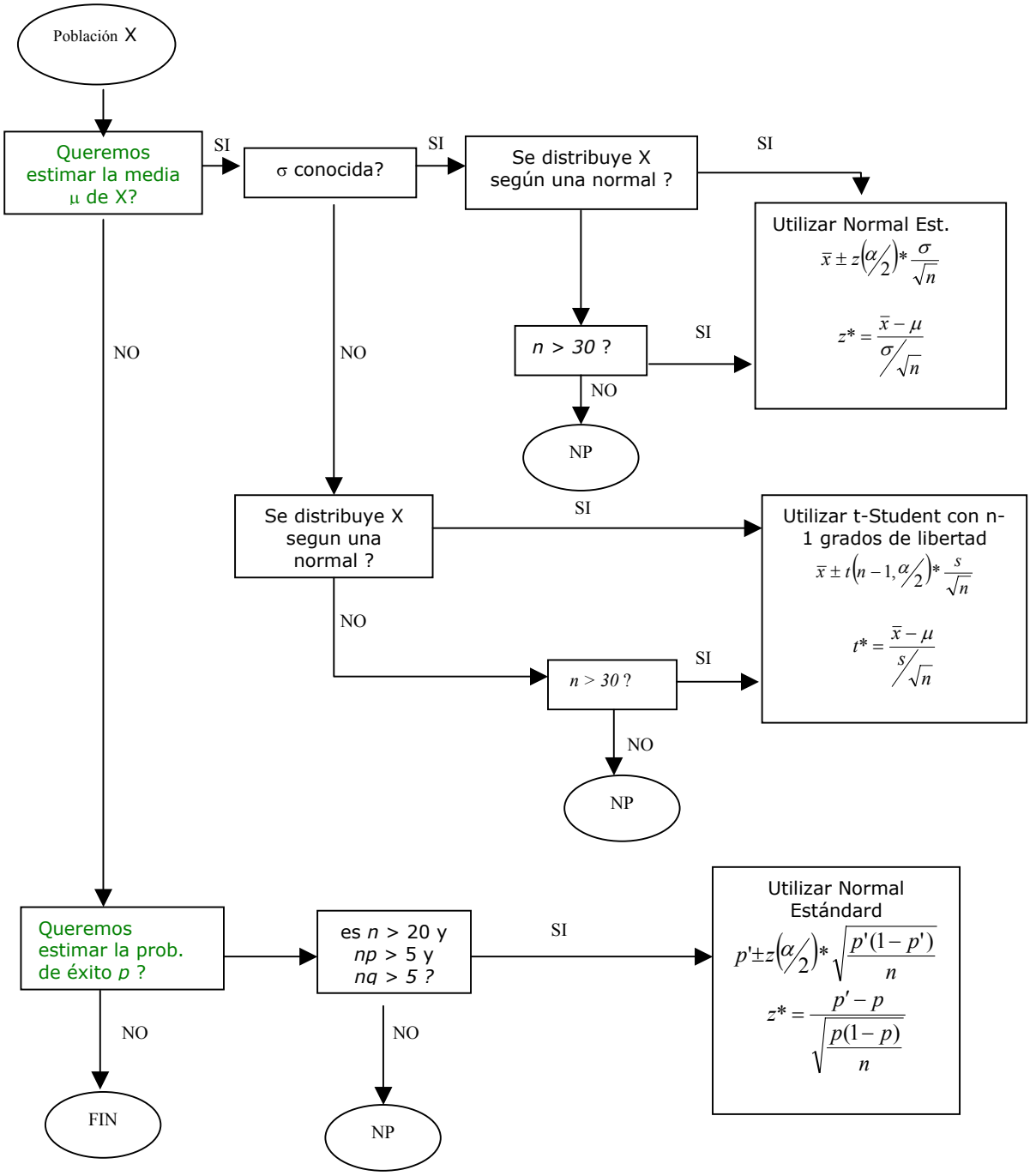
  Recordamos que si *n* $\geq$ 20, *n*p* $\geq$ 5, y *n*(1-p)* $\geq$ 5, entonces $X \approx N\left(np, \sqrt{np(1-p)}\right)$.

- **Supuesto 2:** Las *n* observaciones que constituyen la muestra han sido seleccionadas de manera aleatoria e independiente de una población que no ha cambiado durante el muestreo.

- **Estadístico de contraste:** $z^* = \dfrac{p' - p}{\sqrt{\dfrac{p(1-p)}{n}}} \approx N(0,1)$

- **Criterio de decisión:** Descartaremos H_0 si **p-valor** $\leq \alpha$ (normalmente α = 0,05).

INT. DE CONFIANZA Y CONTR. DE HIPÓTESIS (1 POBLACIÓN)



NP significa que tendremos que utilizar métodos No Paramétricos (fuera del contenido del curso)

2. Contraste de hipótesis para dos poblaciones

MUESTRAS DEPENDIENTES E INDEPENDIENTES

- El que dos muestras sean dependientes o no viene determinado por las fuentes (personas u objetos) que nos aportan las observaciones. Si en la obtención de ambas muestras se han utilizado las mismas fuentes o fuentes asociadas, tenemos dos **muestras dependientes**. Por el contrario, si se han utilizado fuentes completamente diferentes hablaremos de muestras independientes.
- Suponemos que, al iniciar el semestre, seleccionamos al azar a 30 alumnos matriculados en Estadística I y les pasamos un test de conocimientos previos. Al final del semestre, seleccionamos a otros 30 alumnos al azar y les pasamos un test de conocimientos adquiridos durante el curso. En tal caso, consideraríamos ambas muestras como independientes. Por el contrario, si el test de conocimientos adquiridos se realizará en los mismos 30 alumnos que hicieron el test inicial, entonces hablaríamos de muestras dependientes.

INFERENCIAS SOBRE μA - μB EN DOS MUESTRAS DEPENDIENTES

- Dadas dos muestras dependientes, X_A y $X_{B,}$ cada una de ellas de medida *n,* consideraremos la v.a. que resulte de calcular su diferencia: *de* = S_{HA} - $X_{B.}$ Denotaremos para μ_{de} μ= *A*-μ_B y σ_{de} en su media y desviación estándar respectivamente. Por lo tanto, hacer inferencias sobre la diferencia de las dos medias muestrales dependientes será equivalente a hacerlas μ sobre.
- **Supuesto:** S_{HA} y $_{XB}$ siguen una distribución normal.
- **Observación:** Si S_{HA}~ *N*(μ_A $\sigma_{,A}$) y $_{XB}$~ *N*(μ_B $\sigma_{,B}$) → *de = XA-X,Bσ N(*μ_A*-B ,*σ_A*-μB).*
- El **intervalo de confianza**, a nivel 1-α, para μ_{de} μ= *A*-μ_B viene dado por la expresión:

$$\bar{d} \pm t\left(n-1, \alpha/2\right)\frac{s_d}{\sqrt{n}}$$

donde *t(n-*α*1,/2)* es el valor que, en una t-Student con *n-1* grados de libertad, deja a su derecha uno área de α*/2,* y $_{sd}$ es la desviación estándar muestral de la v.a. *d* .

- El **estadístico de contraste** para el test $\begin{cases} H_0 : \mu_d = \mu_0 \\ H_1 : \mu_d \neq \mu_0 \end{cases}$ (o bien $<$ ó $>$) es:

$$t^* = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}} \approx t - Sudent(n-1)$$

INFERENCIAS SOBRE μA - μB EN DOS MUESTRAS INDEPENDIENTES

- Trabajaremos ahora con dos muestras independientes, S_{HA} y $X_{B,}$ de medidas n_A y n_B respectivamente.
- **Supuesto:**S_{HA} y $_{XB}$ siguen una distribución normal.
- **Observación:**Si S_{HA}~ *N*(μ_A $\sigma_{,A}$) y $_{XB}$~ *N*(μ_B $\sigma_{,B}$)→

$$\rightarrow \bar{X}_A - \bar{X}_B \approx N(\mu_A - \mu_B, \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}) .$$

- El **intervalo de confianza**, a nivel 1-α, para μA-μ_B viene dado por la expresión:

$$(\bar{x}_A - \bar{x}_B) \pm t\left(\min\{n_A - 1, n_B - 1\}, \alpha/2\right)\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

donde $t(min\{\ldots\}, \alpha/2)$ es el valor que, en una t-Student con los grados de libertad indicados, deja a su derecha uno área de $\alpha/2$, y s_A, s_B son las desviaciones estándar de las muestras.

- El **estadístico de contraste** para el test $\begin{cases} H_0 : \mu_A - \mu_B = \mu_0 \\ H_1 : \mu_A - \mu_B \neq \mu_0 \quad (\text{o bien} < \text{ ó } >) \end{cases}$ es:

$$t^* = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \approx t - Sudent\left(\min\{n_A - 1, n_B - 1\}\right)$$

INFERENCIAS SOBRE pA - pB EN DOS MUESTRAS INDEPENDIENTES

- En este caso, tendremos dos v.a. independientes S_{HA} y XB. A partir de ellas definimos las **probabilidades muestrals de éxito** como:

$$p_A \text{ '}= xA / n_A \text{ y } p_{B'} = x_B / n_B$$

- **Supuesto 1:** Las distribuciones de S_{HA} y $_{XB}$ son aproximadamente normales.

Recordamos que si $n\ 20, \geq n*p\ 5, \geq$ y $n*(1-p)\ 5, \geq$ entonces $X \approx N\left(np, \sqrt{np(1-p)}\right)$.

- **Supuesto 2:** Las observaciones de cada muestra han sido seleccionadas de forma aleatoria de dos poblaciones independientes que no cambian durante el proceso de muestreo.
- **Observación:** Para muestras suficientemente grandes (i.e., se cumple el supuesto 1), si

$$(p'_A - p'_B) \approx N\left(p_A - p_B, \sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}\right)$$

- El **intervalo de confianza**, a nivel 1-α, para p_A-p_B viene dado por la expresión:

$$(p'_A - p'_B) \pm z\left(\alpha/2\right)\sqrt{\frac{p'_A(1-p'_A)}{n_A} + \frac{p'_B(1-p'_B)}{n_B}}$$

donde $z\alpha(/2)$ es el valor que, en una normal estándar, deja a su derecha uno área de $\alpha/2$.

- El **estadístico de contraste** para el test $\begin{cases} H_0 : p_A - p_B = 0 \\ H_1 : p_A \neq p_B \quad (\text{o bien} < \text{ ó } >) \end{cases}$ será:

$$z^* = \frac{(p'_A - p'_B)}{\sqrt{p(1-p)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$ si p_A y p_B son conocidos, o bien

$$z^* = \frac{(p'_A - p'_B)}{\sqrt{p'_P(1-p'_P)\left(\frac{1}{n_A}+\frac{1}{n_B}\right)}}$$ si p_A y p_B son desconocidos, siendo $p'_P = \frac{x_A + x_B}{n_A + n_B}$

3. El contraste de hipótesis paso a paso

CONTRASTE DE HIPÓTESIS.

Para poder saber si un determinado resultado encontrado en un experimento o en una determinada observación se debe o no al azar realizamos un **test de hipótesis**. Hacer un test hipótesis es comparar dos hipótesis complementarias y decidir con cuál de las dos nos quedamos.

Un test de hipótesis también se llama *contraste de hipótesis o prueba de significación*.

Las dos hipótesis complementarias se llaman:

A) Hipótesis nula, H_0

B) Hipótesis alternativa, H_1

Se llama hipótesis nula porque aquello que habitualmente se plantea el investigador es negarla, de manera que se pueda rechazar la hipótesis nula y, en consecuencia, aceptar la hipótesis alternativa.

En el mundo de la estadística nunca podemos estar completamente seguros de la decisión tomada, es decir, la decisión que tomemos en la elección de una u otra de las hipótesis (H_0 o H_1) contendrá un cierto riesgo de error, una cierta probabilidad de equivocarnos en nuestra elección. Este riesgo de error es resume en el siguiente cuadro:

	Afirmamos que H_0 es cierta	**Rechazamos el H_0 = H_0 es falsa**
En realidad, H_0 es cierta	Correcto	Error de tipo I o α
En realidad, el H_0 es falsa	Error de tipo II o β	Correcto

El error de tipo I o α viene dado por el mismo valor de α que nos aparecía cuando hacíamos muestreo. Es el error más que hay que controlar, y lo hacemos mediante el **nivel de significación**.

El error de tipo II o β es desconocido y varía en relación inversa a cómo α varía.

En una investigación, habitualmente, lo que se busca es negar la hipótesis nula y, en consecuencia, aceptar la alternativa. De hecho, la hipótesis alternativa es, normalmente, aquello que estamos buscando. Por eso, buscamos minimizar el error de tipo I. A la posibilidad de rechazar el H_0 (tomando H_1) y equivocarnos es lo que denominamos **error de tipo I, α o nivel de significación**, y será fijado por los investigadores.

Los valores más habituales para este **nivel de significación** (α) son 0'05 y 0'01. Son valores *totalmente arbitrarios*. Ahora bien, desde el punto de vista de la objetividad científica, este valor se tendría que *establecer a priori*, antes de la obtención de los resultados. De manera complementaria, tenemos el **nivel de confianza**, que es 1-α, y que habitualmente toma los valores de 0'95 y 0'99.

Así, con un nivel de significación previamente fijado, cuando en un test rechazamos la hipótesis nula, podemos afirmar con un 1-α de confianza que la hipótesis alternativa se cumpla.

Mientras que α se determina a priori, no podemos hacer lo mismo con, β el error de tipo II. Sin embargo, cometer un error de tipo *II no es tan grave como* cometer un error de tipo I. De hecho, cuando aceptamos la hipótesis nula estamos diciendo que no tenemos bastante evidencia estadística para rechazarla.

PROCEDIMIENTO

¿Qué es contrastar? Nuestra hipótesis se planteará para ver si podemos afirmar que la media encontrada en una muestra ($\bar{x}$) realmente se corresponde con la media de una población (μ). O bien si una proporción encontrada en una muestra (p) realmente se corresponde con la proporción de una población (π).

Para comprobarlo, calcularemos uno estadístico de contraste que *contrastaremos* con los valores de las tablas. Dependiendo de si conocemos o no la desviación estándar de la población, el estadístico de contraste se comparó con la distribución *t* o con la distribución *z*. Con la *t* será cuando no se conozca la desviación estándar, con la *z* cuando SÍ se conozca la desviación estándar.

El estadístico de contraste se calcula de la siguiente manera:

Media	$\frac{\bar{x}-\mu}{error\_estàndard}$
Proporción	$\frac{p-\pi}{error\_estàndard}$

Donde el error estándar está:

Media	$\frac{s}{\sqrt{n}}$
Proporción	$\sqrt{\frac{p(1-p)}{n}}$

- **<u>Ejemplo de la página 116 del módulo didáctico: CONTRASTAR UNA MEDIA</u>**

  1º paso: Planteamiento de hipótesis

  H_0: μ = 35 milímetros

  H_1: $\mu \neq$ 35 milímetros

  2º paso: Calcular el estadístico de contraste (Éste es un paso que es diferente en cada ocasión y se trata sólo de aplicar la fórmula concreta en cada caso.)

  s^2 = 0,015

  $s = \sqrt{0{,}015}$

  error estándar = $\frac{\sqrt{0{,}015}}{\sqrt{84}}$ = 0,0153

estadístico de contraste = $\frac{\bar{x} - \mu}{error\_estàndard} = \frac{35{,}02 - 35}{0{,}0153}$ = 1,306

3º paso: Conclusión

El estadístico de contraste habrá que compararlo con una $t_{\alpha/2, n-1}$ = $t_{0{,}025,63}$

No encontramos en las tablas la t para 63 grados de libertad, así pues tendremos que utilizar una t para 60 grados de libertad.

$t_{0{,}025,60} = \pm 2{,}000$

Como 1,306 es menor que la t, entonces nada se opone a aceptar la H_0, es decir, la media es todavía de 35 milímetros.

- **Ejemplo de la página 117 del módulo didáctico: CONTRASTAR UNA PROPORCIÓN**

1º paso: Planteamiento de hipótesis

$H_0: \pi \leq 0{,}5$

$H_1: \pi > 0{,}5$

Se trata, en consecuencia, de un contraste unilateral por la derecha

2º paso: Cálculo estadístico de contraste

Conocemos la variancia de la población, ya que si π = 0,5, la varianza será igual en 0,5 multiplicado por 1- 0,5. En consecuencia:

error estándar = $\sqrt{\frac{0{,}5(1-0{,}5)}{1500}}$ = 0,0129

estadístico de contraste = $\frac{p - \pi}{error\_estàndard} = \frac{0{,}52 - 0{,}5}{0{,}0129}$ = 1,550

3º paso: Conclusión

Utilizaremos la distribución *z*, ya que conocemos la varianza de la población. Se trata de un contraste unilateral, así que habrá que utilizar una z_α = $z_{0{,}05}$ = 1,645

Como 1,550 es menor que 1,645, entonces nada se opone a aceptar la H_0, es decir, la proporción encontrada no es significativamente superior a un 50%, no hay mayoría.

- **Los cuatro pasos en un test de hipótesis.**

1º paso: Formulación de las hipótesis de trabajo y de las hipótesis estadísticas

Hacer un test de hipótesis es plantearse dos cuestiones complementarias y sacar una conclusión. Siempre ponemos en la hipótesis alternativa que aquello que vemos en las dos muestras refleja realmente el comportamiento de las dos poblaciones que comparamos. En general, el planteamiento es el siguiente:

A) H_0 = Igualdad (es decir, el signo "igual" siempre va a la hipótesis nula, ya sea solo o con uno "mayor o igual" o uno "menor o igual")

B) H_1 = No igualdad (es decir, los signos "mayor", "menor" siempre van a la hipótesis alternativa)

Un ejemplo: queremos conocer si el sueldo de los documentalistas de Barcelona es significativamente igual o significativamente desigual a los de los documentalistas de Madrid (se trata de una comparación entre dos medias). Entonces tendremos que:

$H_0: \mu_{Madrid} = \mu_{Barcelona}$

$H_1: \mu_{Madrid} \neq \mu_{Barcelona}$

Otro ejemplo: en un cuestionario preguntamos si se conoce un tema A o si se conoce un tema B para ver el conocimiento diferenciado de estos dos temas, obtendremos una proporción π_a de los que conocen A y una proporción π_b de los que conocen B (se trata de la comparación entre dos proporciones). La formulación estadística de estas hipótesis sería la siguiente:

$H_0: \pi_a = \pi_b$

$H_1: \pi_a \neq \pi_b$

Eso sería un **test bilateral**, pero también podemos trabajar con un test unilateral, ya sea por la derecha o por la izquierda. Si seguimos en el ejemplo de las proporciones...

Un test **unilateral por la derecha** sería:

$H_0: \pi_a \leq \pi_b$

$H_1: \pi_a > \pi_b$

O también:

$H_0: \pi_a = \pi_b$; es decir, $\pi_a - \pi_b = 0$

$H_1: \pi_a > \pi_b$; es decir, $\pi_a - \pi_b > 0$

Un test **unilateral por la izquierda** sería:

$H_0: \pi_a \geq \pi_b$

$H_1: \pi_a < \pi_b$

O también:

$H_0: \pi_a = \pi_b$; es decir, $\pi_a - \pi_b = 0$

$H_1: \pi_a < \pi_b$; es decir, $\pi_a - \pi_b < 0$

2º paso: Calculamos el estadístico de contraste (Éste es un paso que es diferente en cada ocasión y se trata sólo de aplicar la fórmula concreta en cada caso.)

Primero hay que encontrar el error estándar de la diferencia entre muestras

Media	$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$
Proporción	$s_{p_1-p_2} = \sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

Después calcularemos el estadístico de contraste

Media	$\frac{\bar{x}_1 - \bar{x}_2}{error\_estàndard}$
Proporción	$\frac{p_1 - p_2}{error\_estàndard}$

3º paso: Escoger una zona de aceptación y de rechazo adecuadas

Eso se hace comparándolo con las tablas *de la normal* o con las tablas de la *t de Student*.

Establecemos un nivel de significación (α) y habrá que calcular su z_α correspondiente (o la t_α).

Por ejemplo, si α = 0'05, en caso de tratarse de una prueba bilateral, entonces, $\alpha/2$ = 0'025, en consecuencia, $z_{\alpha/2} = \pm 1'96$.

Si se trata de una prueba unilateral, entonces α = 0'05 estará toda en un lado, y z_α = 1'645. Será negativa en el caso de una prueba unilateral izquierda y positiva en el caso de una prueba unilateral derecha.

4º paso: Tomar una decisión estadística

Si la Z observada entra dentro de la **zona de aceptación**, entonces aceptamos la H_0

Si la Z observada entra dentro de la **zona crítica**, entonces rechazamos la H_0 y, en consecuencia, aceptamos la H_1

En el caso de α = 0'05, si la prueba es bilateral:

La zona de aceptación será [-1 '96, 1'96] y la de rechazo desde $-\infty$ (es decir el máximo a la izquierda de la distribución) hasta -1 '96 y de +1'96 hasta $+\infty$ (es decir, el máximo a la derecha de la distribución)

Si la prueba es unilateral izquierda:

La zona de aceptación será [-1 '645, $+\infty$] y la de rechazo será [$-\infty$, -1'645]

Si la prueba es unilateral derecha:

La zona de aceptación será [$-\infty$,1'645] y la de rechazo será [1'645, $+\infty$]

Aquí tenéis unas tablas con los valores de "z" para los niveles de significación más usuales.

Nivel de significación	'10	0'05	0'01	0'005	0'001
Valores críticos de "z" para un contraste de una sola cola por la izquierda	-1'28	-1'645	-2'33	-2'58	-2'88
Valores críticos de "z" para un contraste de una sola cola por la derecha	1'28	1'645	2'33	2'58	2'88
Contraste bilateral	-1'645 i 1'645	-1'96 i 1'96	-2'58 i 2'58	-2'81 i 2'81	-3'08 i 3'08

- **Modificaciones al ejemplo página 122-123 del módulo didáctico: comparación de dos medias**

  Tenemos dos consumos de marcas de coches, la marca R y la marca S. La hipótesis que queremos corroborar es que los dos coches tienen consumos diferentes.

  Una vez hemos llegado al paso 2 (calcular el estadístico de contraste), con una t en qué el $\alpha/2$ es igual en 0'005 (ya que α es igual en 0'01) y unos grados de libertad igual en 130. Nosotros no tenemos esta t y, en consecuencia, por aproximación, utilizaremos una t con 120 grados de libertad. Este valor es igual a $\pm$ 2'617.

  La zona de aceptación estará pues de -2 '617 hasta +2'617. El estadístico de contraste queda fuera de la zona de aceptación, rechazamos en consecuencia la hipótesis nula y aceptamos la alternativa: las dos marcas de coche tienen consumos de combustible diferentes.

- **Modificaciones al ejemplo página 124-125 del módulo didáctico: comparación de dos proporciones**

  Tenemos dos encuestas sobre aceptación a la legalización de marihuana y queremos contrastar que ha habido un cambio significativo en la opinión en este aspecto.

  Consideramos que conocemos la varianza de la población y utilizamos, pues, una distribución z. El nivel de significación es 0'05, motivo por el cual $\alpha/2$ será igual en 0'025. La z correspondiente será 1'96. La zona de aceptación será de -1 '96 hasta 1'96, y el estadístico de contraste cae claramente fuera de esta zona, en consecuencia, rechazamos la hipótesis nula: se ha dado un cambio de actitud en la población norteamericana.

3. Algunos ejemplos de contraste de hipótesis:

Ejemplo 1: Página 117 del manual: CONTRASTAR UNA PROPORCIÓN MUESTRAL

Datos que nos proporcionan:

Proporción de la muestra = 52% = 0'52

n= 1500 norteamericanos

Nivel de significación = = 5%α = 0'05

1º paso: planteamiento de las hipótesis

Hipótesis nula: no es una mayoría significativa: la proporción es menor o igual al 50%

H_0 : $\mu \leq$ 0'50.

Hipótesis alternativa: es una mayoría significativa: la proporción es mayor al 50%

H_1 : μ > 0'50.

2º paso: cálculo de estadístico de contraste

Hará falta primero calcular *la desviación estándar de la proporción* $= \sqrt{\frac{\pi(1-\pi)}{n}} = \sqrt{\frac{0'50(1-0'50)}{1500}} =$ 0'0129

$$EC = \frac{p - \pi}{\sigma_p} = \frac{0'52 - 0'50}{0'0129} = 1'550$$

3º paso: construir zona de aceptación y rechazo

Se trata de un contraste de hipótesis unilateral por la derecha, donde $\alpha = 0'05$.

Éste 0.05 está por la derecha, es decir, en nuestras tablas habrá que buscar el área complementaria = 0'95 (ya que las tablas que tenemos nos muestran siempre el área a la izquierda del punto que tenemos).

Como estamos contrastando los parámetros de una muestra utilizaremos la distribución t de Student.

Los grados de libertad serán n-1 = 1550.

Pero nuestras tablas no nos permiten unos grados de libertad por encima de 1000, así, buscaremos el valor que corresponde a unos grados de libertad de 1000.

$t_{0'95,1000} = 1'646$

En consecuencia, la zona de aceptación será (∞: 1'646).

Y la zona de rechazo será (1'646: ∞).

4º paso: conclusión

El EC cae DENTRO de la zona de aceptación, NADA SE OPONE A ACEPTAR LA H_0.

No es una mayoría significativa: la proporción es menor o igual al 50%.

No hay una mayoría a favor de la despenalización de la posesión de marihuana.

Ejemplo 2: CONTRASTAR LA DIFERENCIA DE MEDIAS DE UNA MUESTRA DEPENDIENTE

Muestra dependiente = Dos experimentos con la misma muestra.

Nos interesa estudiar la diferencia entre uno y otro tratamiento.

H_0 = no hay cambio (la diferencia es igual en 0).

H_1 = hay cambio (la diferencia no es igual en 0).

- **Vemos un ejemplo:**

  Un club de esquí organiza un curso de buen estado físico de dos semanas para ejecutivos. Hacen pesar a cinco de los participantes seleccionados al azar antes del curso y después del curso. Los resultados son los siguientes:

Número	Peso anterior	Peso posterior
Josep Maria	81	77
Xavier	77	76
Pere	75	73
Albert	88	83
Ricard	76	74

Contrastad si ha habido una reducción de peso significativa (contrastadlo a nivel del 5% y suponed una distribución normal para los datos.

Datos que nos proporcionan:

En primer lugar, calculamos la diferencia de peso antes y después del curso, para cada uno de los 5 sujetos de la muestra. Eso nos permite elaborar la media de estas diferencias

Media de las diferencias = $\bar{x}_{diferències} = \frac{4+1+2+5+2}{5}$ = 2'8

También podemos calcular la varianza = 2'7.

n= 5 individuos.

Nivel de significación = α = 5% = 0'05.

1º paso: planteamiento de las hipótesis

Hipótesis nula: no hay cambio: la diferencia de medias es 0.

$H_0 : \mu = 0$.

Hipótesis alternativa: hay cambio: la diferencia no es 0.

$H_1 : \mu \neq 0$. **(** esta hipótesis se podría formular de diferente manera si lo que nos interesa verificar es si los participantes perdieron peso, como parece que era el objetivo).**

2º paso: cálculo de estadístico de contraste

Hará falta primer calcular *el error estándar de la media* = $\frac{desviació\_estàndard\_mostra}{\sqrt{n}}$

$s_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{\sqrt{2'7}}{\sqrt{5}}$ = 0'73

Estadístico de contraste = $\frac{mitjana\_mostral - mitjana\_població}{error\_estàndard\_de\_la\_mitjana}$

EC = $\frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{2'8}{0'73}$ = 3'81

3º paso: construir zona de aceptación y rechazo

Se trata de un contraste de hipótesis bilateral, así pues habrá que utilizar la mitad de α en cada lado.

α = 0'05; $\alpha/2$ = 0'025.

Como estamos contrastando los parámetros de una muestra utilizaremos la distribución t de Student.

Los grados de libertad serán n-1 = 4.

La t que habrá que buscar en las tablas será $t_{0'025,4 =}$ -2 '776.

En consecuencia, las zonas de rechazo serán (-∞, -2'776) i (2'776 , ∞).

Y la zona de aceptación será (-2 '776: 2'776).

4º paso: conclusión

El EC cae FUERA de la zona de aceptación, dentro de la zona de rechazo derecha.

Rechazamos el H_0, la diferencia de medias no es 0.

Se ha dado una pérdida de peso antes y después del tratamiento.

Ejemplo 3: Página 123 del manual: CONTRASTAR DOS MEDIAS

Datos que nos proporcionan:

$\bar{x}$ = 12'1 km/l coches R y 13'9 km/l coches S

n = 82 coches R y 50 coches S

s = 2 '8 coches R y 4' 0 coches S

Nivel de significación = α = 0'01

1º paso: planteamiento de las hipótesis

Hipótesis nula: las dos medias provienen de la misma población.

$H_0 : \mu_R = \mu_S$.

Hipótesis alternativa: las dos medias no provienen de la misma población.

$H_1 : \mu_R \neq \mu_S$.

2º paso: cálculo de estadístico de contraste

Primero habrá que calcular *la desviación estándar común de las medias*

$$s = \sqrt{\frac{(82-1)\cdot 2'8^2 + (50-1)\cdot 4'0^2}{(82+50-2)}} = 3'304$$

Error estándar de la diferencia de medias:

$$s_{\bar{x}_R - \bar{x}_S} = 3'304\sqrt{\frac{1}{82}+\frac{1}{50}} = 0'5928$$

Finalmente, el estadístico de contraste

$$EC = \frac{(13'9 - 12'1)}{0'5928} = 3'036$$

3º paso: construir zona de aceptación y rechazo

Se trata de un contraste de hipótesis bilateral, donde α = 0'01, es decir, $\alpha/2$ = 0'005

Como estamos contrastando los parámetros de una muestra utilizaremos la distribución t de Student.

Los grados de libertad serán n1 + n2 - 2 = 82 + 50 - 2 = 130

Pero nuestras tablas no nos permiten 130 grados de libertad, así, buscaremos el valor que más se aproxima, es decir 120.

$t_{0'005,120}$ = -2'617

En consecuencia, las zonas de **rechazo** serán (∞: -2'617) i (2'617: ∞)
Y la zona **de aceptación** será (-2'617:2'617).

4º paso: conclusión

El EC cae FUERA de la zona de aceptación, SE NIEGA LA $H_{0;}$ SE ACEPTA LA H_1.
Las dos marcas de coches tienen consumo de combustibles diferentes.

Bibliografía, materiales complementarios y enlaces de interés

- Como bibliografía complementaria podéis consultar la que figura en el Plan Docente.

Fe de erratas

Capítulo 17, página 113: Al tercer párrafo, cuarta línea dice: "estas muestras tienen la media 0 y la estándar 1/raíz cuadrada de 100"; y tendría que decir "el error estándar".
Capítulo 16, página 108, cuando dice ejercicios 17.1, 17.2 y figura 17.3. En todos los casos tendría que ser 16 en lugar de 17.

Capítulos: 19, 20, 21 y 22

Guía de estudio - GES_7

CORRELACIÓN Y REGRESIÓN LINEAL

Presentación

Esta Guía de Estudio (**GES_7**) pretende orientar el estudio de los contenidos de los Capítulos: 19, 20, 21 y 22 relacionados con la correlación y regresión lineal.

Esta **GES_7** incorpora el siguiente material:

1. Correlación lineal
2. Regresión lineal.
3. Inferencia sobre la recta de regresión

Materiales: para trabajar esta **GES_7** se necesitan los materiales básicos de la asignatura (Capítulos: 19, 20, 21 y 22).

Calendario: la temporización de la **GES_7** será la prevista en el PlanDocente.

Objetivos

Con el estudio de la **GES_7** se pretende que el estudiante consiga los siguientes objetivos:

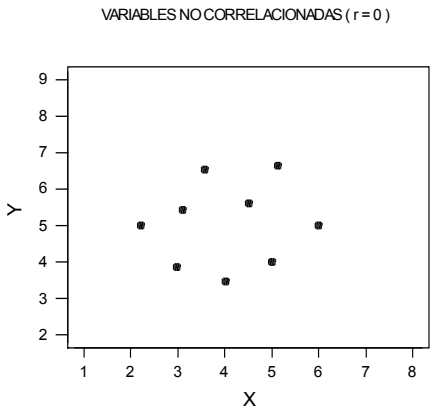
1. Introducir al estudiante en el conocimiento de la correlación lineal.
2. Entender el concepto de regresión lineal y saber calcular una recta de regresión.

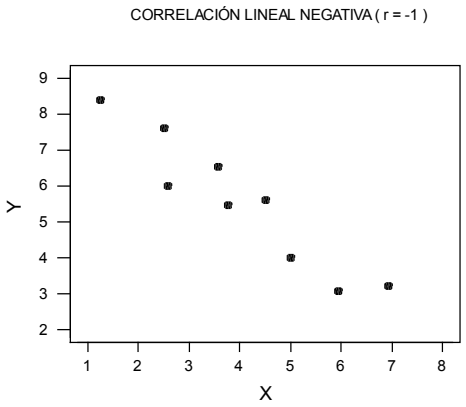
Contenidos

1. Correlación lineal

- En ocasiones nos puede interesar estudiar si existe o no algún tipo de relación entre dos variables aleatorias. Así, por ejemplo, podemos preguntarnos si hay alguna relación entre las notas de la asignatura Estadística I y las de Matemáticas I. Una primera aproximación al problema consistiría en dibujar en el plano R^2 un punto por cada alumno: la primera coordenada de cada punto sería su nota en estadística, mientras que la segunda sería su nota en matemáticas. Así, obtendríamos una nube de puntos que podría indicarnos visualmente la existencia o no de algún tipo de relación (lineal, parabólica, exponencial, etc.) entre ambas notas.
- En particular, nos interesa cuantificar la intensidad de la relación **lineal** entre dos variables. El parámetro que nos da tal cuantificación es el **coeficiente de correlación lineal de Pearson r**, cuyo valor oscila entre −1 y +1 :

$$-1 \le r = \frac{Cov(X,Y)}{s_X s_Y} \le +1$$





- Como se observa en los diagramas anteriores, el valor de *r* se aproxima a +1 cuando la correlación tiende a ser lineal directa (mayores valores de *X* significan mayores valores de *Y*), y se aproxima a –1 cuando la correlación tiende a ser lineal inversa.

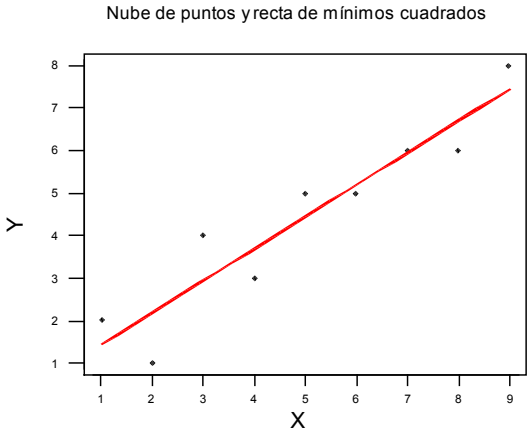
- Es importante notar que la existencia de correlación entre variables no implica causalidad.

- ¡Atención!: si no hay correlación de ningún tipo entre dos v.a., entonces tampoco habrá correlación lineal, por lo que r = 0. Sin embargo, el que ocurra r = 0 sólo nos dice que no hay correlación lineal, pero puede que la haya de otro tipo.

- Denominamos **coeficiente de determinación R^2** al cuadrado de *r* . Podemos interpretar R^2 como el porcentaje de la variación en *Y* que viene "explicado" por el modelo lineal obtenido: a mayor porcentaje mejor es nuestro modelo para "predecir" el comportamiento de la v.a. *Y*.

2. Regresión lineal

- En aquellos casos en que el coeficiente de regresión lineal sea "cercano" a +1 o a –1, tiene sentido considerar la ecuación de la recta que "mejor se ajuste" a la nube de puntos (recta de mínimos cuadrados). Uno de los principales usos de dicha recta será el de predecir o estimar los valores de Y que obtendríamos para distintos valores de X.



- La ecuación de la **recta de mínimos cuadrados** (en forma punto-pendiente) es la siguiente:

$$y - \bar{y} = \frac{Cov(X,Y)}{s_x^{\,2}}(x - \bar{x})$$

SUPUESTOS DEL MODELO DE REGRESIÓN LINEAL

- En el caso en que nuestras observaciones sean una muestra aleatoria proveniente de una población, estaremos interesados en realizar inferencias sobre la misma. A fin de que estas inferencias sean "estadísticamente razonables", se han de cumplir las siguientes condiciones:

  1. En la población, la relación entre las variables *X* e *Y* debe ser aproximadamente lineal, i.e.: $y = \beta_0 + \beta_1 x + \varepsilon$, siendo ε la v.a. que representa los **residuos** (diferencias entre el valor estimado por el modelo y el verdadero valor de *Y*).

  2. Los residuos se distribuyen según una Normal de media 0, i.e.: $\varepsilon \approx N(0,\sigma^2)$.

  3. Los residuos son independientes unos de otros.

  4. Los residuos tienen varianza σ^2 constante.

- Afortunadamente, el modelo de regresión lineal es bastante "robusto", lo que significa que no es necesario que las condiciones anteriores se cumplan con exactitud (en particular las tres últimas).

3. Inferencias sobre la pendiente de la recta (β1)

- Supongamos que hemos extraído una muestra de *n* pares de valores *(x_i , y_i)*, *i = 1,…,n*, procedentes de una población *(X,Y)*, y que hemos calculado el coeficiente de regresión lineal asociado *r* , el cual, por ser próximo a +1 o a −1, parece indicar la existencia de una correlación lineal entre los valores de la muestra. ¿Es suficientemente "significativo" el valor de *r* como para asegurar la existencia de una correlación lineal entre las dos v.a. que conforman la población?. En otras palabras,

¿podemos afirmar que el **coeficiente de correlación lineal poblacional** ρ es significativamente distinto de cero?

- Una forma alternativa de plantearse la cuestión anterior sería: a partir de la muestra podemos calcular la ecuación de la recta de mínimos cuadrados asociada, la cual podemos escribir como $\hat{y} = b_0 + b_1 x$, donde b_0 y b_1 son estimaciones de los valores "verdaderos" β_0 y β_1. La pregunta ahora sería: dado un valor cualquiera de la v.a. *X*, ¿es buena la estimación que obtenemos de *Y* dada por la recta de mínimos cuadrados obtenida? En otras palabras, ¿es posible afirmar que la **pendiente de la recta de regresión poblacional** β_1 es significativamente distinta de cero? De ser así, tendríamos que, en efecto, existe una correlación lineal entre ambas variables poblacionales.

- ¡Observación importante!: $r = 0 \leftrightarrow b_1 = 0$ (ya que el numerador de ambos parámetros es el mismo). Por tal motivo, los dos contrastes siguientes son equivalentes:

$$(\text{i}) \quad \begin{Bmatrix} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{Bmatrix} \qquad \text{y} \qquad (\text{ii}) \quad \begin{Bmatrix} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{Bmatrix}$$

- El estadístico (t-Student) que se utiliza para realizar el test (ii) es el siguiente:

$$t = \frac{b_1 - \beta_1}{s_{b1}} \approx t\left(n-2, \alpha/2\right), \text{ donde } s_{b1} = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{(n-2)\left[\sum x^2 - \frac{\left(\sum x\right)^2}{n}\right]}}$$

donde $t(n-2,\alpha/2)$ es el valor asociado a una t-Student con $n-2$ grados de libertad que deja a su derecha un área de $\alpha/2$ (o, equivalentemente, deje a su izquierda un área de $1 - \alpha/2$).

- Nota: si en vez de realizar el contraste bilateral (ii) deseamos hacer un contraste unilateral (en el cual la hipótesis alternativa sería $H_1 : B_1 > 0$ ó $H_1 : B_1 < 0$), deberemos sustituir en la fórmula anterior $\alpha/2$ por α (ya que ahora trabajaremos con una única cola de la distribución).

- Finalmente, también podemos obtener el intervalo de confianza para β_1 a nivel de confianza $(1-\alpha)$ utilizando la expresión:

$$b_1 \pm t\left(n-2, \alpha/2\right) * s_{b1}$$

Bibliografía, materiales complementarios y enlaces de interés

- Como bibliografía complementaria podéis consultar la que figura en el Plan Docente.