3. Análisis univariable y bivariable

3.1. Análisis univariable

Como se ha visto, los métodos de análisis univariable se utilizan para estudiar el comportamiento de las variables de forma individual.

Podéis consultar el subapartado 2.2 de este módulo didáctico.

3.1.1. Distribución de frecuencias

Las distribuciones de frecuencias permiten obtener una primera aproximación de la tendencia de los resultados, ya que indican el número de individuos que tanto en valores absolutos como en valores porcentuales han mencionado cada uno de los códigos posibles (respuestas) que puede tomar una variable determinada.

La ventaja principal de las distribuciones de frecuencias es que pueden llevarse a cabo sea cual sea la escala en que están medidas las variables que se deben analizar.

Los resultados de los ejemplos presentados a continuación, y también los ejemplos expuestos en el resto de los métodos de análisis que se explican en este módulo, provienen del tratamiento de los datos con el paquete estadístico SPSS.

Ejemplo de distribución de frecuencias

En la tabla siguiente exponemos la distribución de frecuencia correspondiente a la variable V6 "Número de litros de leche consumidos en el hogar a la semana", extraída de un estudio sobre los hábitos de consumo de productos lácteos en el que se entrevistó a 836 principales responsables de la compra en el hogar. El universo objeto de estudio lo constituyeron hogares de la ciudad de Barcelona consumidores de un litro de leche a la semana, como mínimo.

En el paquete estadístico SPSS, la información que proporciona la distribución de frecuencias de la variable que se analiza se presenta en las seis columnas siguientes:

V6 "Número de litros de leche que consumen a la semana"

	Value Label	Value	Frequency	Percent	Valid percent	Cum percent
1	litro/semana	1	15	1.8	1.9	1.9
2	litros/semana	2	100	12.0	12.5	14.3
3	litros/semana	3	118	14.1	14.7	29.1
4	litros/semana	4	67	8.0	8.4	37.4
5	litros/semana	5	75	9.0	9.4	46.8
6	litros/semana	6	70	8.4	8.7	55.5
7	litros/semana	7	89	10.6	11.1	66.6

V6 "Número de litros de leche que consumen a la semana"

Value Label	Value	Frequency	Percent	Valid percent	Cum percent
8 litros/semana	8	52	6.2	6.5	73.1
9 litros/semana	9	19	2.3	2.4	75.4
10 litros/semana	10	62	7.4	7.7	83.2
11 litros/semana	11	7	.8	.9	84.0
12 litros/semana	12	55	6.6	6.9	90.9
13 litros/semana	13	3	.4	.4	91.3
14 litros/semana	14	37	4.4	4.6	95.9
15 litros/semana	15	7	.8	.9	96.8
16 litros/semana	16	7	.8	.9	97.6
17 litros/semana	17	1	.1	.1	97.8
18 litros/semana	18	4	.5	.5	98.3
20 litros/semana	20	5	.6	.6	98.9
21 litros/semana	21	5	.6	.6	99.5
22 litros/semana	22	1	.1	.1	99.6
24 litros/semana	24	3	.4	.4	100.0
	.	34	4.1	Missing	
	Total	836	100.0	100.0	

Valid cases 802 Missing cases 34

Distribución de frecuencias.

Value Label: lista las etiquetas asignadas por el investigador a cada código posible de la variable. Si la variable se ha medido en una escala de tipo métrico, como es el caso que nos ocupa, no es necesario etiquetar los códigos, ya que el valor del código es suficiente para saber que corresponde a un consumo determinado de litros de leche a la semana.

Value: indica los distintos valores que toma la variable.

Ejemplo

En nuestro ejemplo, hay hogares que consumen desde 1 litro hasta 24 litros de leche a la semana.

Frequency: indica el número de individuos que, en valores absolutos, han mencionado cada uno de los valores posibles de la variable.

Ejemplo

En nuestro ejemplo, 15 hogares consumen 1 litro de leche a la semana, 100 consumen 2 litros..., y 34 hogares no han contestado el número de litros de leche que consumen. Este dato en SPSS está representado por un punto (*missing value*).

Percent: indica el porcentaje de individuos que, sobre el total de los entrevistados, han mencionado cada uno de los valores que toma la variable.

Ejemplo

En nuestro ejemplo, el 1,8% de los hogares consume 1 litro de leche a la semana, el 12% consume 2 litros..., y el 4,1% de los hogares entrevistados no ha contestado a esta pregunta.

Valid percent: indica el porcentaje de individuos que han mencionado cada uno de los posibles valores de la variable, tomando como base de cálculo no

la totalidad de los entrevistados, como en el caso de *percent*, sino la totalidad de los entrevistados que han respondido a la pregunta.

Ejemplo

En nuestro ejemplo, la base de cálculo serían los 802 hogares que han respondido el número de litros de leche que consumen (836 menos los 34 que no han respondido).

Cum percent: expresa el porcentaje acumulado, es decir, el porcentaje de individuos que han mencionado un valor determinado o alguno de los valores anteriores a éste. Se calcula, igual que el *valid percent*, sobre el número de individuos que han contestado a la pregunta y no sobre la totalidad de los entrevistados.

Ejemplo

En nuestro ejemplo, en el 66,6% de los hogares que han respondido se consumen de 1 a 7 litros de leche a la semana.

La información proporcionada para una distribución de frecuencias se puede sintetizar mediante el cálculo de los tipos estadísticos descriptivos que veremos a continuación: los estadísticos descriptivos que permiten medir la **tendencia central** y los que permiten medir la **dispersión**.

3.1.2. Medidas de tendencia central

Los tipos estadísticos que miden la tendencia central permiten apreciar cuál es el comportamiento medio de cada variable. Los tres más utilizados son la **moda**, la **mediana** y la **media**.

En el cuadro siguiente se presenta el resultado de estos tres indicadores de tendencia central para la variable "Número de litros de leche que consumen a la semana".

```
Mean 6.685          Median 6.000          Mode 3.000
```

Medidas de tendencia central (en litros).

Media (*mean*): indica cuál es el valor medio de la variable. Es el cociente entre la suma ponderada de cada valor de la variable por el número de individuos que la han mencionado, y el número total de individuos:

$$\overline{X} = \frac{\sum_{c=1}^{C} f_c \cdot x_c}{n}$$

donde:

C = número de categorías de la variable, $c = 1, ..., C$;

x_c = valor tomado por la categoría c de la variable X;

f_c = número de individuos que han mencionado la categoría c de la variable X;

n = número total de individuos.

Ejemplo

En nuestro ejemplo, la media de litros de leche consumidos por hogar y semana es:

$$\overline{X} = \frac{15 \cdot 1 + 100 \cdot 2 + 118 \cdot 3 + ... + 3 \cdot 24}{802} = 6{,}685 .$$

Mediana (*median*): indica el valor de la distribución que divide la muestra en dos partes iguales o aproximadamente iguales.

Ejemplo

Siguiendo el ejemplo, la mediana es 6, lo que significa que el 55,5% de los hogares consume 6 litros o menos de leche a la semana y que el 44,5% consume más de 6 litros de leche a la semana.

Moda (*mode*): indica el valor de la respuesta más mencionada.

Ejemplo

En el ejemplo anterior, la moda es 3 litros de leche a la semana, ya que el 14,1% de los hogares dice que consume 3 litros a la semana y éste es el valor con un porcentaje superior de citaciones.

La tendencia central de una variable se mide con uno de estos tres tipos estadísticos descriptivos, según la escala utilizada:

1) Si las variables están medidas en **escalas ordinales**, el indicador apropiado será la **mediana**.

2) Si las variables están medidas en **escalas nominales**, el indicador apropiado será la **moda**.

3) Si las variables están medidas en **escalas cuantitativas**, la medida de tendencia central adecuada será la **media**.

En este último caso, también hay que tener en cuenta que la media se calcula a partir de todos los valores de la distribución y, por lo tanto, es altamente sensible a los valores extremos, ya sean bajos o altos, los cuales suelen denominarse ***outliers***. Si hay *outliers*, la media no es una medida adecuada de la tendencia central y hay que recurrir a la mediana o a la moda.

3.1.3. Medidas de dispersión

Las medidas de dispersión permiten analizar la variabilidad de la distribución, es decir, determinar si las respuestas que han dado las personas entrevistadas se han concentrado sólo en unos cuantos valores o si, por el contrario, han sido muy variadas. La dispersión se mide respecto del comportamiento medio de la variable, por lo que la elección de la medida de dispersión que hay que utilizar también depende de la escala en que esté medida la variable que se analiza.

Si la escala de medida es cualitativa (nominal u ordinal), la única medida de dispersión que puede utilizarse para medir el grado de concentración de las respuestas es la frecuencia relativa de la moda, es decir, el porcentaje de individuos que ha mencionado el valor modal.

Ejemplo

En la tabla siguiente, podemos apreciar que en la variable "Situación laboral del principal responsable de las compras en el hogar" las respuestas están concentradas mayoritariamente en el valor modal; el 60,9% trabaja por cuenta ajena.

V152 "Situación laboral actual del responsable del hogar"

Value Label	Value	Frequency	Percent	Valid percent	Cum percent
Trabaja por cuenta propia	1	216	25.8	25.8	25.8
Trabaja por cuenta ajena	2	509	60.9	60.9	86.7
Inactivo	3	111	13.3	13.3	100.0
	Total	836	100.0	100.0	
Valid cases 836		Missing cases	0		

Distribución de la situación laboral del principal responsable del hogar.

La medida que permite evaluar la dispersión de las respuestas respecto de la media cuando la escala de medida es cuantitativa es la **varianza** *(variance)*. Otras medidas de dispersión que permiten completar la información suministrada por la varianza son las siguientes:

- El **coeficiente de simetría** *(skewness)*: indica el grado de simetría o asimetría de la distribución.

- El **coeficiente de apuntamiento** (*curtosis*): valora si las respuestas están concentradas en pocos valores o están repartidas.

A continuación, se expone en qué consiste cada uno.

Ejemplo

En el cuadro se presenta el resultado de estos indicadores de dispersión para la variable "Número de litros de leche que consumen a la semana".

Variance	18.016	Std dev	4.245
Kurtosis	1.290	kewness	1.118

Medidas de dispersión del número de litros de leche que consumen a la semana.

1) La **varianza** es la suma de las diferencias entre la media de la distribución y un valor cualquiera de la distribución. Así pues, cuando los datos están concentrados en torno a la media, la varianza es pequeña, y cuando están repartidos, la varianza es elevada. El cálculo de la varianza utiliza la media al cuadrado de las desviaciones de todos los valores observados respecto de la media. Nunca puede ser negativa. En el caso de datos agrupados en categorías, la fórmula de la varianza es:

$$S^2 = \frac{\sum_{c=1}^{C} f_c (x_c - \bar{x})^2}{n-1}.$$

Ejemplo

Si aplicamos esta fórmula a nuestro ejemplo, tenemos que el resultado de la variable "Número de litros de leche que consumen a la semana" es 18,016:

$$S^2 = \frac{15(1-6{,}685)^2 + 100(2-6{,}685)^2 + \ldots + 3(24-6{,}685)^2}{802-1} = 18{,}016\,.$$

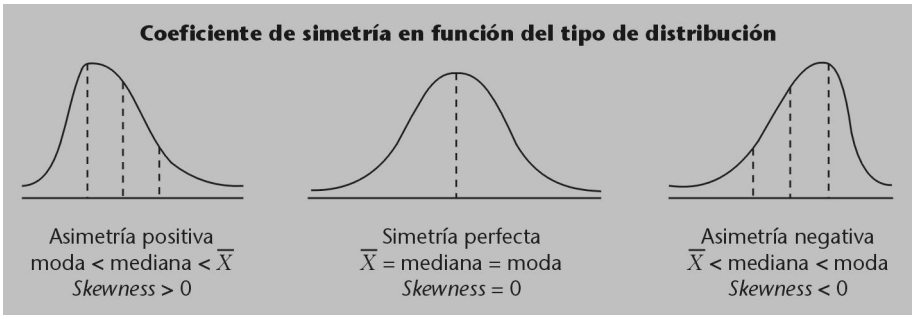
El valor de la varianza está en unidades al cuadrado y, por lo tanto, este resultado no es directamente comparable con el resto de la información. En cualquier caso, lo que suele hacerse es efectuar la raíz cuadrada de este valor y obtener así la desviación típica de la variable, que ya estará expresada en la misma unidad que los datos, y no en unidades al cuadrado. En nuestro ejemplo la desviación típica (como muestra el cuadro anterior) es de 4,245 litros.

2) El **coeficiente de simetría** (*skewness*) indica el grado de simetría de la distribución y permite ver rápidamente si es simétrica o asimétrica.

Una distribución es simétrica cuando el número de observaciones que hay a cada uno de los lados del centro de la distribución son iguales y las desviaciones positivas y las correspondientes desviaciones negativas respecto de la media también son iguales; en consecuencia, la media, la moda y la mediana coinciden en el mismo valor. En este caso, podemos afirmar que la distribución es normal, por lo que el coeficiente de simetría es igual a cero.

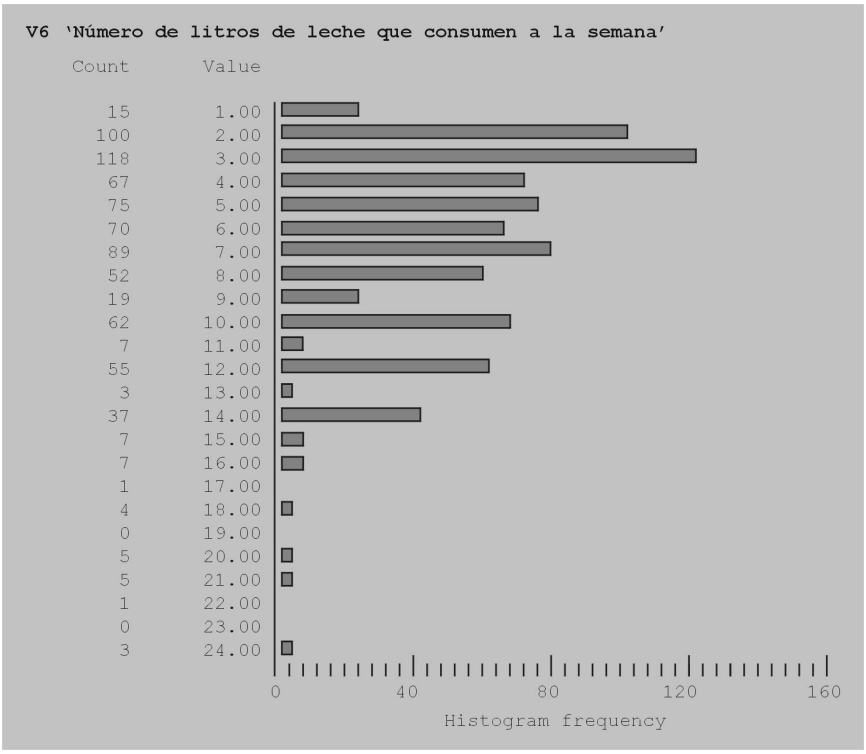
Una distribución es asimétrica cuando las desviaciones positivas y negativas respecto de la media no son iguales. Cuanto más alejado de cero sea el coeficiente de simetría, más asimétricas son las respuestas a la izquierda

(los individuos están más concentrados en valores o códigos bajos), y cuanto más alejado de cero y negativo sea el coeficiente de simetría, más asimétricas son las respuestas a la derecha (los individuos están más concentrados en valores o códigos altos).



Ejemplo

En nuestro ejemplo, el coeficiente de simetría es positivo, 1,118; eso significa que los individuos están concentrados en valores bajos de la distribución. Efectivamente, al representar la distribución de la V6 gráficamente, se aprecia este resultado:

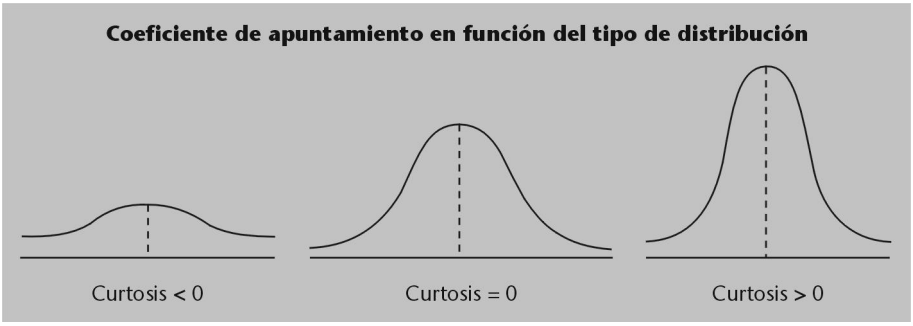


Histograma del número de litros de leche que consumen a la semana.

3) El **coeficiente de apuntamiento** indica el grado de concentración de las respuestas. Un coeficiente de apuntamiento igual a cero significa que la distribución de las respuestas se aproxima a la de una distribución normal en cuanto a su altura.

Cuanto más alejado de cero y positivo sea este coeficiente, más concentradas están las respuestas en unos cuantos valores de la distribución. Cuanto más

alejado de cero y negativo sea este coeficiente, más repartidas están las respuestas a lo largo de un gran número de valores de la distribución, tal como muestran los gráficos siguientes:



Ejemplo

En nuestro ejemplo, el coeficiente de apuntamiento es positivo, 1,290, lo que indica que los individuos están concentrados en pocos valores de la distribución. Efectivamente, el 73,5% de los hogares consume entre 2 y 9 litros de leche a la semana, y las respuestas van desde 1 litro hasta 24 litros de leche a la semana.

3.1.4. Inferencia estadística

En investigación comercial, una vez obtenidos los resultados es importante validarlos, es decir, ver si hay diferencias entre los resultados obtenidos en la investigación y unos valores determinados conocidos *a priori* o teóricos; en caso de que las haya, hay que comprobar si estas diferencias son estadísticamente significativas o si, por el contrario, se deben al azar.

Ejemplo

En el estudio sobre el mercado de productos lácteos se podría tener el propósito de verificar si es posible afirmar que el número medio de litros de leche consumidos por hogar en la ciudad de Barcelona es de uno al día, es decir, siete a la semana, en lugar de los 6,685 litros a la semana que daba el resultado de la media.

El proceso que hay que seguir para validar los resultados se denomina **test de inferencia estadística**.

Las etapas que deben seguirse para llevar a cabo este proceso son las siguientes:

1) Establecer la hipótesis nula H_0 y su alternativa H_1.

2) Elegir un nivel de significación α.

3) Elegir el estadístico adecuado para contrastar H_0 y calcularlo bajo la hipótesis nula H_0.

4) Determinar el valor crítico a partir del cual rechazamos H_0 (zona de rechazo).

5) Comparar el valor del estadístico con el valor teórico para determinar si es necesario o no rechazar H_0 con el nivel de significación especificado.

Etapa 1: establecer la hipótesis nula H_0 y su alternativa H_1

Para contrastar un resultado determinado, es necesario plantear *a priori* dos hipótesis:

1) Hipótesis nula (H_0): la diferencia entre *X* e *Y* es estadísticamente nula y, por lo tanto, puede afirmarse que se debe a las oscilaciones del azar.

2) Hipótesis alternativa (H_1): la diferencia entre *X* e *Y* es estadísticamente significativa y, por lo tanto, puede afirmarse que no se debe a las oscilaciones del azar.

El test de inferencia estadística consiste en contrastar estas dos hipótesis con el fin de verificar cuál de las dos es cierta. Según un principio general de este tipo de test, todas las diferencias se deben al azar mientras no se demuestre lo contrario, por lo cual lo que siempre se somete a comprobación es la hipótesis nula H_0. Rechazar la hipótesis nula H_0 supone aceptar automáticamente la hipótesis alternativa H_1 y, por el contrario, aceptar la hipótesis nula H_0 supone rechazar automáticamente la hipótesis alternativa H_1.

Etapa 2: elegir un nivel de significación a

Tal como se muestra en el cuadro que hay a continuación, la decisión a la cual se llega después de haber finalizado el test siempre lleva asociados dos tipos de error:

1) El **error de tipo I** se comete cuando se rechaza la hipótesis nula y ésta, en realidad, es verdadera. La probabilidad de cometer un error de tipo I está representada por α y se denomina **nivel de significación**. El nivel de significación lo fija *a priori* el investigador y es el riesgo de error que se está dispuesto a asumir en caso de que rechace la hipótesis nula y ésta sea verdadera. De forma convencional, suelen elegirse niveles de significación 0,05 y 0,01. Es decir, que se está dispuesto a asumir un error del 5% o del 1% en el momento de rechazar la hipótesis nula.

2) El **error de tipo II** se comete cuando se acepta la hipótesis nula y ésta, en realidad, es falsa. La probabilidad de cometer un error del tipo II se denomina

riesgo β. Este riesgo siempre es desconocido, ya que generalmente no se conocen los parámetros de la población. Por lo tanto, es imposible saber si acertamos al aceptar la hipótesis nula.

Cuadro de decisión			
		Hipótesis nula H_0	
		Verdadera	**Falsa**
Decisión	Se acepta	Correcta	Error tipo II (riesgo β)
	Se rechaza	Error tipo I (riesgo α)	Correcta

Etapa 3: elegir el estadístico adecuado para contrastar H_0 y calcularlo bajo la hipótesis nula H_0

El estadístico que hay que utilizar para testar H_0 dependerá, una vez más, de la escala en que esté medida la variable que se analiza. Si la variable es cualitativa, debe calcularse el estadístico de la khi-cuadrado, y si es cuantitativa, debe calcularse el estadístico *Z* o el estadístico *t*, según el tamaño de la muestra. Si ésta es superior a treinta casos, hay que aplicar el estadístico Z, y si es inferior a treinta, hay que aplicar el estadístico *t*. La fórmula de cálculo de cada uno es la siguiente:

1) Estadístico de la **khi-cuadrado**:

$$\chi^2 = \sum_{i=1}^{K} \frac{(O_i - T_i)^2}{T_i},$$

donde:

O_i = frecuencia observada de la categoría *i*;

T_i = frecuencia teórica de la categoría *i*;

k = número de categorías.

Según esta fórmula, cuanto mayor sea la distancia entre O_i y T_i, mayor será el valor de χ^2 y más elevadas las posibilidades de rechazar H_0.

2) Estadísticos ***Z*** y ***t***:

a) En caso de que el valor que hay que analizar sea una media:

$$Z = \frac{|\overline{X} - \mu|}{\frac{S}{\sqrt{n}}}; \qquad t = \frac{|\overline{X} - \mu|}{\frac{S}{\sqrt{n}}},$$

donde X es la media observada en la muestra, μ la media observada en la población o norma, s la desviación típica de la muestra y n el tamaño de la muestra.

b) En caso de que el valor que hay que analizar sea una proporción:

$$Z = \frac{p - \theta}{\sqrt{\frac{p(1-p)}{n}}}; \qquad t = \frac{p - \theta}{\sqrt{\frac{p(1-p)}{n}}}$$

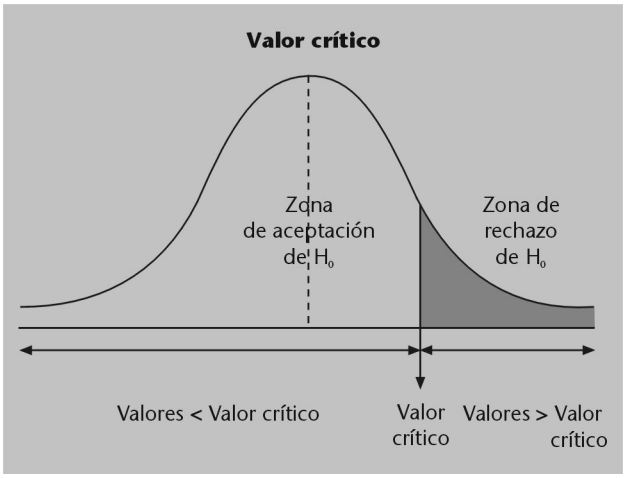
donde p es el porcentaje observado en la muestra y θ el porcentaje observado en la población o norma.

Los estadísticos Z y t se calculan a partir de la misma fórmula; la única diferencia entre sí es la obtención del valor crítico con el que deben compararse.

Podéis consultar la etapa 4 para la determinación del valor crítico.

Etapa 4: determinar el valor crítico a partir del cual rechazamos H_0 (zona de rechazo)

Para contrastar H_0 también es preciso definir lo que se denomina el **valor crítico** a partir del cual se determina la zona de rechazo de H_0, es decir, la zona de la distribución del estadístico en la que corresponde rechazar la hipótesis nula en el caso de que el valor del estadístico pertenezca a esta zona de la distribución, tal como aclara este gráfico.



El valor crítico se obtiene a partir de la distribución del estadístico que se utiliza: la distribución de la khi-cuadrado en el caso del cálculo del estadístico de la khi-cuadrado, la distribución normal en el caso del cálculo del estadístico Z y la distribución t de Student en el caso del cálculo del estadístico t.

Podéis consultar el anexo 3 al final de este módulo didáctico.

Etapa 5: comparar el valor del estadístico con el valor crítico para determinar si tenemos que rechazar H_0 o no con el nivel de significación especificado

Si el valor del estadístico es superior al valor crítico, es decir, si se sitúa en la zona de rechazo de H_0, tenemos que rechazar H_0. Al contrario, si el valor del estadístico es inferior al valor crítico, es decir, si se sitúa en la zona de aceptación de H_0, no podemos rechazar H_0 con el nivel de significación especificado.

A continuación veremos dos ejemplos de aplicación de un test de inferencia estadística.

Primer ejemplo

Supongamos que tenemos que contrastar la hipótesis de que el nivel de estudios de los individuos entrevistados en el estudio sobre el mercado de productos lácteos es igual al

nivel de estudios de la población. El nivel de estudios obtenido en la muestra y el nivel de estudios de la población son los expuestos en el cuadro siguiente:

V149 "Grado de instrucción del responsable del hogar"

Value Label	Value	Porcentaje muestra	Porcentaje población.
hasta primarios	1	20.1	56.0
secundarios	2	39.7	35.0
superiores	3	40.2	9.0
	Total	100.0	100.0

Grado de instrucción de la muestra y de la población.

Parece que el examen de las frecuencias indica que el nivel de estudios de los individuos entrevistados difiere del nivel de estudios de la población. El resultado de la prueba estadística nos indicará si esta afirmación es correcta. Con esta finalidad, seguimos cada una de las etapas planteadas más arriba.

1. La hipótesis nula H_0 es que no hay diferencias entre la frecuencia observada y la frecuencia teórica o de la población. Sólo intervienen las diferencias debidas al azar. La hipótesis alternativa H_1 es que hay diferencias significativas entre la frecuencia observada y la frecuencia teórica o de la población.

2. El nivel de significación elegido (que corresponde al riesgo en que se incurriría si se rechazara H_0 por error) es de 0,05 ($\alpha = 5\%$).

3. La prueba estadística adecuada para una variable ordinal es la prueba de la khi-cuadrado. El cálculo del estadístico χ^2 aplicando la fórmula es el siguiente:

$$\chi^2 = \frac{(20 \cdot 1 - 56)^2}{56} + \frac{(39 \cdot 7 - 35)^2}{35} + \frac{(40 \cdot 2 - 9)^2}{9} = 131{,}81$$

4. El valor crítico χ_c^2 se obtiene a partir de la distribución de la khi-cuadrado. La lectura de la distribución se efectúa para un nivel de significación α y para unos grados de libertad determinados, en este caso $k - 1$ grados de libertad, donde k es el número de categorías de la variable analizada. En el ejemplo χ_c^2 (2 grados de libertad, $\alpha = 5\%$) = 5,99 .

5. El valor observado de la khi-cuadrado ($\chi^2 = 131{,}81$) es superior al valor crítico ($\chi_c^2(2{,}5\%) = 5{,}99$), rechazamos H_0. La distribución del grado de instrucción de los individuos de la muestra no se ajusta a la distribución del grado de instrucción en la población.

Segundo ejemplo

En este segundo ejemplo queremos averiguar si puede afirmarse que el número medio de litros de leche consumidos por hogar en la ciudad de Barcelona es de uno al día, es decir, de siete a la semana.

Podéis consultar el anexo 1 al final de este módulo didáctico.

1. La hipótesis nula H_0 es que el número medio de litros de leche consumidos a la semana y por hogar ($\overline{X} = 6{,}685$)no es diferente de una norma (μ) de 7 litros de leche consumidos a la semana y por hogar. La hipótesis alternativa H_1 es $\overline{X} < \mu$.

$$H_0 \ \overline{X} = 7$$

$$H_1 \ \overline{X} < 7$$

2. El nivel de significación es $\alpha = 5\%$.

3. El estadístico adecuado es el estadístico *Z*, ya que la variable es métrica y el tamaño de la muestra es superior a 30.

Con los datos de la distribución obtenemos:

$$Z = \frac{|6{,}685 - 7|}{\frac{4{,}245}{\sqrt{802}}} = 2{,}101 .$$

4. El valor crítico se obtiene a partir de la distribución normal. En nuestro caso, teniendo en cuenta un nivel de significación del 5%, el valor crítico es igual a 1,645.

Podéis consultar el anexo 2 al final de este módulo didáctico.

5. El valor del estadístico *Z* es superior al valor crítico, por tanto rechazamos H_0. La conclusión de la prueba es que hay el 95% de posibilidades de que el consumo de leche medio por hogar y a la semana observado en la muestra sea inferior a 7 litros.

3.2. Análisis bivariable

Los métodos de análisis bivariable se utilizan para estudiar las relaciones que hay entre variables tomadas de dos en dos.

Podéis consultar el subapartado 2.2 de este módulo didáctico.

3.2.1. Análisis bivariable entre dos variables cualitativas: tablas de contingencia

Las tablas de contingencia analizan la distribución de frecuencia conjunta de dos variables de tipo cualitativo. Las categorías de una variable se cruzan con las categorías de la otra, de modo que la distribución de una variable se subdivide de acuerdo con las categorías de la otra variable.

Las tablas de contingencia constituyen uno de los instrumentos más utilizados en investigación comercial porque los resultados son fácilmente interpretables y comprensibles para directivos con pocos conocimientos estadísticos, lo que permite utilizar de una manera rápida los resultados de la investigación en acciones empresariales.

Ejemplo

Presentamos a continuación un ejemplo de tabla de contingencia entre dos variables extraídas del estudio sobre los hábitos de consumo de productos lácteos. Las variables analizadas son las siguientes:

a. "Situación laboral del principal responsable de las compras en el hogar" (V152), codificada en tres grupos:

1. Trabaja por cuenta propia.

2. Trabaja por cuenta ajena.

3. Inactivo.

b. "Su grado de instrucción" (V149), codificado en dos grupos:

1. Sin estudios universitarios.

2. Con estudios universitarios.