3

Situaciones problema con una muestra para promover la investigación en el aula

En este capítulo se presentan variadas situaciones problema relacionadas con el uso de tópicos de estadística no paramétrica, para promover la investigación científica en el salón de clases universitario. En la primera sección se hace referencia al manejo de datos pertenecientes a una variable cualitativa, los cuales están clasificados en dos o más categorías o modalidades y considerados ya sea en una población o en una muestra. En la segunda se explicita la prueba de hipótesis basada en la distribución binomial y en una sola muestra aleatoria. En la tercera se indica la prueba chi-cuadrado con una muestra aleatoria donde la variable cualitativa presenta más de dos categorías. En la cuarta se aborda la prueba K-S de Kolmogorov-Smirnov, para analizar si los datos de una variable continua provienen de una población con una distribución específica. En este marco, se aplica un conjunto de pasos destinados a la ejecución del proceso de inferencia estadística, referido a la prueba de hipótesis.

3.1 Situaciones alusivas al trabajo con datos provenientes de una variable de tipo cualitativo en una población o en una muestra

Para este tipo de situaciones hay que recordar que una variable es cualitativa cuando corresponde a una característica que puede estudiarse en los individuos de una población o de una muestra, pero que sus datos son factibles de clasificarse en categorías o grupos mutuamente excluyentes; es decir, que un individuo no ha de estar en dos o más grupos a la vez. Si los individuos pertenecen a la población objeto de

estudio, será suficiente con realizar estadística descriptiva sobre los datos que conforman las categorías de la variable; en cambio, si los datos de la variable de interés corresponden a una muestra, entonces es posible hacer inferencia estadística y comprobar las hipótesis a que haya lugar.

Por otra parte, los datos de la variable *X* pueden conformar solamente dos categorías o también pueden clasificarse en tres o más grupos. En seguida, se ilustra una situación que puede implicar el uso de inferencia estadística o solamente elementos de estadística descriptiva.

Una empresa registrada en la Cámara de Comercio con el nombre de PAW Productores de artículos W, en el mes de abril de 2022 presentaba un total 50 trabajadores, quienes laboraban en 4 áreas específicas en la ciudad de Cali, codificadas de la siguiente manera:

G: Gestión, P: Producción, C: Comercialización, T: Talento humano
F: Finanzas

G, T, P, P, P, P, P, C, F, C, C, T, G, T, T, P, P, P, P, C, F, F, T, P, G,
G, T, P, P, P, P, P, C, F, C, C, T, G, T, T, P, P, P, P, C, F, F, T, P, G.

Los datos se recolectaron mediante un formulario llamado censo. En este caso, se trata de 50 individuos que pertenecen a una población (Burbano et al., 2021). Por lo tanto, sobre estos datos solamente se harán procesos relativos a la estadística descriptiva. En primera instancia, se identificará la variable y luego se organizarán sus datos en 5 grupos mutuamente excluyentes.

La variable de interés se denota y se define de la siguiente forma:

X: área de trabajo donde se desempeña cada uno de los trabajadores de la empresa PAW Productores de artículos W, localizada en la ciudad de Cali, en el mes de abril de 2022.

Recuérdese que los datos corresponden a cada uno de los valores que admite la variable *X*; la frecuencia absoluta es la cantidad de veces que se repite cierto dato, y la frecuencia relativa corresponde a un porcentaje alusivo a cada frecuencia relativa.

Después de organizar los datos, los 5 grupos o categorías mutuamente excluyentes son los siguientes:

G, G, G, G, G, G, P, P, P, P, P, P, P, P, P, P, P, P, P, P, P, P, P, P, P,
C, C, C, C, C, C, C, C, T, T, T, T, T, T, T, T, T, T, F, F, F, F, F, F

La primera categoría está constituida por los trabajadores que laboran en el área de gestión (G), la segunda categoría la conforman quienes trabajan en el área de producción (P), la tercera corresponde a quienes trabajan en la de comercialización (C), la cuarta es la de talento humano (T) y la quinta categoría es la de finanzas (F).

La frecuencia absoluta para la categoría G es de 6, para P es de 20, para la categoría C es de 8, para T es de 10, y para la categoría F es de 6, para un total de $N = 50$ individuos que constituyen la población objeto de estudio. Ahora, se puede proseguir a organizar los datos en una tabla de frecuencias absolutas, como se presentan en la Tabla 3.1. Además, los valores que se calculen con estos datos se suelen llamar parámetros, porque se están refiriendo a la población de trabajadores de toda la empresa PAW.

Tabla 3.1. Organización de los datos de *X*

X: Área de trabajo	Frecuencia absoluta
G	6
P	20
C	8
T	10
F	6
Total	*N*=50

Fuente: los autores

En la Tabla 3.1 se observa que el área de la empresa con más trabajadores es la de producción (P) con 20 trabajadores; es decir, P es la que presenta la mayor frecuencia absoluta. También, con base en la misma tabla, se determina que las categorías G y F son las que muestran la frecuencia absoluta menor o más baja. En esta población también es posible calcular la medida descriptiva correspondiente al parámetro llamado la moda de *X*, la cual se suele denotar con $M_o(X)$. Entonces, la moda corresponde al dato que presenta la mayor frecuencia absoluta; en esta situación la moda es el dato P (trabajadores en el área de producción). Por lo tanto, se puede denotar y escribir así:

$$M_o(X) = P$$

Una actividad subsiguiente propia de la estadística descriptiva corresponde al proceso de interpretación de la información. En este caso: ¿cómo se interpreta la moda? Recuérdese que la moda es un indicativo de la tendencia de los datos; por lo tanto, aquí indica que en las áreas de la empresa PAW hay una tendencia a que sus trabajadores se ubiquen en el área de producción (P).

Otra actividad inherente a la estadística descriptiva es la representación de la información; para elaborar una representación de los datos pertenecientes a una variable cualitativa es recomendable utilizar un diagrama de pastel o torta, basado en las frecuencias absolutas o en las frecuencias relativas.

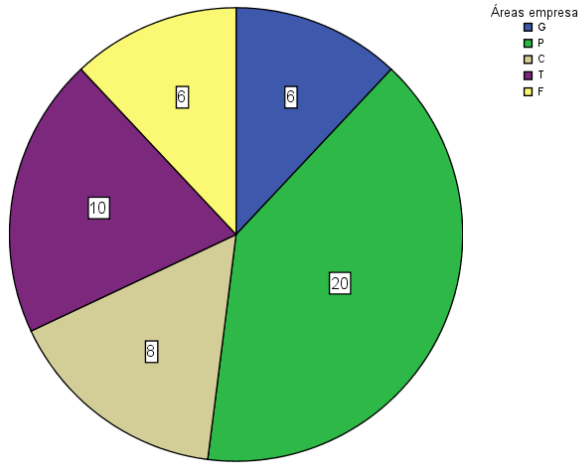

Figura 3.1. Diagrama de torta para la variable X
Fuente: los autores

A continuación, se presenta una tabla de frecuencias relativas asociadas a las 5 categorías en las cuales se han organizado los datos de la variable *X* (Tabla 3.2). De acuerdo con la Tabla 3.2, se puede interpretar que 20 de los 50 trabajadores de la empresa PAW laboran en el área de producción (P); es decir, 20/50 = 0.4 = 0.40 = 40 %; este porcentaje indica que el 44 % de los trabajadores de la mencionada empresa trabajan en el área de producción (P).

Tabla 3.2. Tabla de frecuencias relativas y absolutas

X: Área de trabajo	Frecuencia absoluta	Frecuencia relativa
G	6	0.12 = 12 %
P	20	0.40 = 40 %
C	8	0.16 = 16 %
T	10	0.20 = 20 %
F	6	0.12 = 12 %
Total	N=50	1=100 %

Fuente: los autores

Esta afirmación también equivale a que 40 de cada 100 trabajadores trabajan en el área P, o que 20 de cada 50 trabajan en dicha área. De forma semejante, 10 de los 50 trabajadores trabajan en el área de talento humano (T), lo cual representa al 20 % de los trabajadores quienes trabajan en el área T de la empresa PAW. Así se puede continuar con la interpretación de las frecuencias para las demás categorías.

Algunas veces se acostumbra a representar las frecuencias relativas correspondientes a las categorías de una variable cualitativa mediante un diagrama de barras separadas; en el eje horizontal se ubican las categorías y en el eje vertical se escriben los porcentajes respectivos, como se presentan en la Figura 3.2.

Finalmente, se podría obtener una conclusión como sigue: los trabajadores de la empresa PAW están organizados en 5 áreas de trabajo, de las cuales el área de producción es la que contiene más trabajadores y el área de gestión junto al área de finanzas son las que poseen menos trabajadores. En dicha empresa hay una tendencia a que los trabajadores se ubiquen en el área de producción más que en las otras áreas.

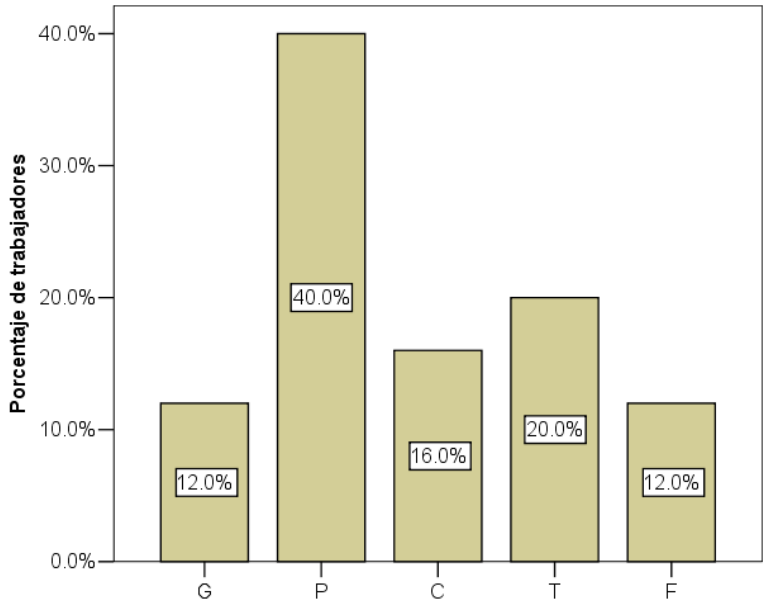

Figura 3.2. Porcentaje de trabajadores por área de trabajo
Fuente: los autores

A continuación, se presenta otra situación hipotética en la cual se puede requerir el uso de estadística descriptiva, inferencia estadística o ambas.

En un estudio sobre control de calidad para el producto A en la empresa T, se toma una muestra aleatoria de 20 unidades, codificadas de la siguiente forma:

B: Buena calidad, D: Defectuoso

B, B, B, D, B, B, B, B, B, B, B, B, B, B, D, B, B, B, B, B.

Los datos fueron colectados a través de un formulario llamado encuesta. En esta situación hipotética se trata de 20 unidades del artículo A (20 individuos), las cuales corresponden a una muestra (Burbano et al., 2021). Por consiguiente, en referencia a estos datos es posible efectuar

procesos sobre la estadística descriptiva y la estadística inferencial. En primer lugar, se identifican las variables y luego se organizan en dos categorías mutuamente excluyentes; estos dos grupos también reciben el nombre de modalidades.

Ahora, la variable de interés, por tener dos categorías, también suele llamarse dicotómica, la cual se mide en una escala nominal y se denota y se define así:

X: estado de calidad de las unidades del producto A seleccionadas aleatoriamente de un proceso productivo en la empresa T.

Al organizar los datos en dos categorías, se obtiene el siguiente arreglo:

B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, D, D.

La primera categoría está conformada por las unidades de buena calidad (B), la segunda categoría la constituyen las unidades defectuosas (D). La frecuencia absoluta para la categoría B es 18 y para la categoría D es 2. Se prosigue a organizar los datos en una tabla de frecuencias absolutas, como se indica en la Tabla 3.3. Adicionalmente, los valores que se calculan con estos datos reciben el nombre de estimadores, porque están vinculados a la muestra aleatoria.

Tabla 3.3. Organización de los datos de *X*

X: Estado de calidad	Frecuencia absoluta
B	18
D	2
Total	n =20

Fuente: los autores

Con base en la Tabla 3.3 se puede determinar que la categoría B es la que presenta la mayor frecuencia absoluta con 18 unidades de buena calidad en la muestra; en contraste, D es la que presenta la menor cantidad de datos, en la muestra de tamaño n = 20. En este caso, la moda para la variable X es B, la cual se denota y escribe de la siguiente manera:

$$m_o(X) = B$$

Para esta situación hipotética, la moda indica que hay una tendencia de que las unidades del artículo A en la empresa T que fueron escogidas aleatoriamente resulten de buena calidad. Una actividad adicional relacionada con la estadística descriptiva consiste en representar la información por medio de un diagrama de torta o pastel, como se puede observar en la Figura 3.3.

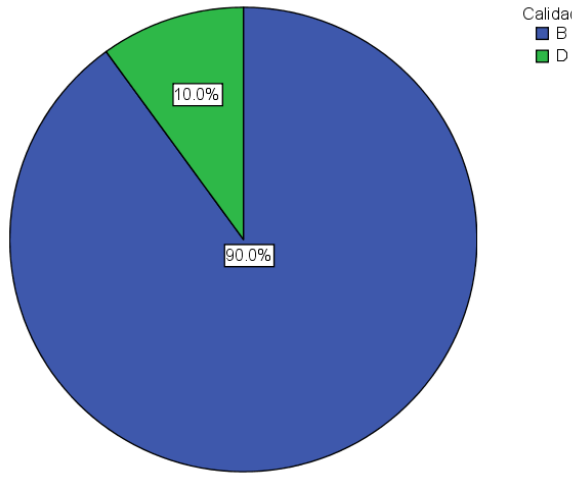

Figura 3.3. Estado de calidad de las unidades del producto A
Fuente: los autores

En seguida, en la Tabla 3.4 se pueden observar las frecuencias relativas a la variable dicotómica X.

En correspondencia con la Tabla 3.4, se interpreta que 18 de los 20 artículos de la muestra aleatoria resultaron de buena calidad, lo cual representa el 90 % de la muestra; en contraste, 2 de los 20 artículos de la mencionada muestra resultaron defectuosos y corresponden al 10 % de esta muestra.

Tabla 3.4. Frecuencias absolutas y relativas para la variable X

X: Estado de calidad	Frecuencia Absoluta	Frecuencia relativa
B	18	0.9 = 90 %
D	2	0.1 = 10 %
Total	n=20	1=100 %

Fuente: los autores

En la Figura 3.4 se pueden observar las respectivas frecuencias relativas.
Finalmente, se puede obtener la siguiente conclusión acerca de los datos de la muestra. Los datos están organizados en dos categorías, de estas, la categoría B es la que más unidades presenta del producto A. En este contexto hay una tendencia a que las unidades de la muestra resulten de buena calidad. Probablemente, al tomar otra muestra aleatoria se obtengan resultados semejantes o quizá diferentes; en consecuencia, a partir de una muestra aleatoria, también será factible realizar procesos de inferencia estadística.

3.2 Prueba de hipótesis basada en la distribución binomial

En diversas circunstancias de investigación es posible que se tenga que trabajar con una población conformada solo por dos categorías; por ejemplo, productos con dos resultados posibles: bueno o defectuoso; también en poblaciones cuyos resultados posibles sean: soltero, casado, masculino, femenino, entre otros. En una población conformada por dos

categorías, saber que la proporción o porcentaje de individuos en la primera clase es *p*, implica que la proporción en la otra categoría es *1 – p = q*.

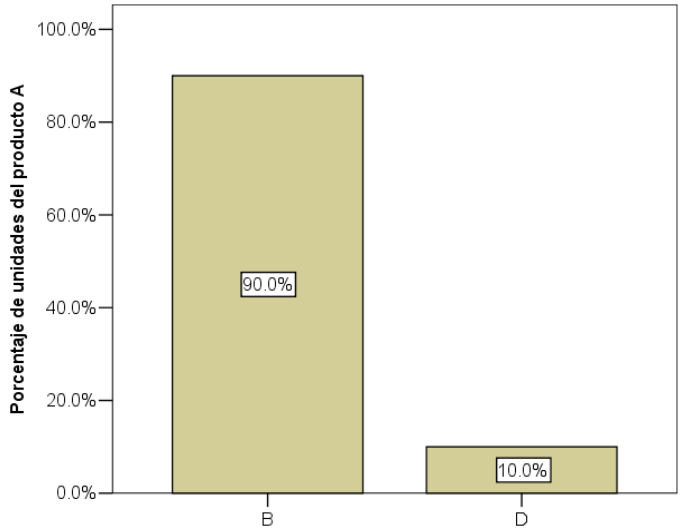

Figura 3.4. Estado de calidad del producto A en la muestra
Fuente: los autores

A pesar de que la proporción *p* puede cambiar de una población a otra, este valor *p* puede considerarse fijo en una población específica; no obstante, aun si se conoce este valor en una determinada población, no se puede esperar que una muestra aleatoria proveniente de tal población contenga exactamente la proporción *p* en la primera categoría y la proporción *q* en la segunda categoría. En este contexto, a los efectos aleatorios en el proceso de muestreo suele atribuírseles el hecho de que la muestra no reproduzca con exactitud los valores de *p* y *q* en esa población.

En estas circunstancias, la distribución muestral correspondiente a la proporción (porcentaje) *p* observada en una muestra aleatoria tomada en

una población de dos categorías es una distribución binomial. Entonces, H_o es la hipótesis nula, la cual informa que p corresponde a un valor específico en la población; esto indica cuán razonables es que la proporción muestral (estimador) provenga de una población de parámetro *p*, o si por el contrario hay diferencias significativas.

En esta situación, para una muestra de tamaño *n*, la probabilidad de que la variable aleatoria *X* sea igual a la cantidad *x* de individuos en la primera categoría y *n*-*x* en la segunda, se puede calcular a través de la distribución binomial (Burbano y Valdivieso, 2015):

$$f(x) = P(X = x) = \begin{cases} \binom{n}{x} p^x q^{n-x} & si\ x = 0,1,2,...,n \\ 0 & \text{en otro caso} \end{cases}$$

Al expresar el combinatorio en términos de números factoriales, la distribución binomial también se puede escribir de la siguiente manera:

$$f(x) = P(X = x) = \begin{cases} \frac{n!}{x!(n-x)!} p^x q^{n-x} & si\ x = 0,1,2,...,n \\ 0 & \text{en otro caso} \end{cases}$$

Donde *n*! = 1*2*3*…*n*; por ejemplo: 4! = 1*2*3*4 = 24; 0! = 1; 1! = 1 y

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Por ejemplo (Burbano et al., 2021):

$$\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{24}{2(2)} = \frac{24}{4} = 6$$

Supóngase que con base en la muestra aleatoria:

B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, D, D

se quiere probar la hipótesis de que no hay diferencias significativas entre la proporción de unidades del producto A que resultan de buena calidad y la proporción de las que resultan defectuosas con un nivel de significancia α del 5 % (α = 0.05).

En concordancia con lo expuesto en Burbano y Valdivieso (2016), un conjunto de pasos para probar esta hipótesis es el siguiente:

1) Planteamiento del sistema de hipótesis

Como se trata de probar que no hay diferencias significativas, la hipótesis nula puede incluir la afirmación de que $p = 0.5$ y la hipótesis alternativa de que p es mayor que 0.5; por lo tanto, el sistema de hipótesis por comprobar se podría escribir así:

H_o: p=0.5
H_1: $p > 0.5$

2) Se fija el nivel de significancia en el valor α = 0.05.

3) La dirección de la prueba irá hacia la derecha; es decir, corresponde a una prueba unilateral derecha.

4) La estadística de prueba se soporta en la distribución binomial.

5) Decisión: si el p-*valor* es menor que $\alpha = 0.05$, entonces se rechaza la hipótesis nula; en caso contrario, no habrá evidencia suficiente para rechazarla y será aceptada.

Aquí, el *p*-valor será igual a sumar las probabilidades para X=18, 19 y 20 calculadas sobre el modelo binomial, es decir:

P-valor = $P(X \geq 18) = P(X = 18) + P(X = 19) + P(X = 20)$

$$P(X = 18) = \binom{20}{18}(0.5)^{18}(0.5)^{20-18} = 190(0.5)^{20} = 0.00018$$

$$P(X = 19) = \binom{20}{19}(0.5)^{19}(0.5)^{20-19} = 20(0.5)^{20} = 0.000019$$

$$P(X = 20) = \binom{20}{20}(0.5)^{20}(0.5)^{20-20} = 1(0.5)^{20} = 0.000009$$

En consecuencia,

p-valor = $P(X \geq 18) = 0.00018 + 0.000019 + 0.000009 = 0.000208$

Como el *p*-valor resultó inferior al nivel de significancia $\alpha = 0.05$, entonces se rechaza la hipótesis nula y se acepta la hipótesis alternativa.

6) Con un nivel de significancia del 5 %, se concluye que sí hay diferencias significativas en la proporción poblacional *p* referida a la cantidad de unidades del producto A con respecto a la proporción de unidades defectuosas.

Otra forma de resolver esta misma situación problema consiste en utilizar una aproximación de la distribución binomial por una distribución

normal estándar cuando el tamaño de la muestra n sea lo suficientemente grande, así:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

Donde $\hat{p}$ corresponde a la proporción muestral y p_0 al valor de la proporción poblacional que aparece en la hipótesis nula, $q_0 = 1 - p_0$; es decir:

$$\hat{p} = \frac{x}{n}$$

Cuando se trabaja con la cantidad x de la variable aleatoria X, una expresión equivalente para la estandarización mencionada es la siguiente:

$$Z = \frac{x - np_0}{\sqrt{np_0 q_0}}$$

Para muestras pequeñas se usa el factor de corrección por continuidad. Esta corrección consiste en disminuir en $0.5/n$ el valor observado en la proporción muestral $\hat{p}$ y el valor esperado en la proporción p_0 poblacional. Cuando $\hat{p} < p_0$ se hace un incremento de $0.5/n$; en cambio, cuando $\hat{p} > p_0$ se hace una disminución de $0.5/n$; en consecuencia,

si $\hat{p} < p_0$ entonces se utiliza $Z = \frac{(\hat{p} + 0.5/n) - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$

si $\hat{p} > p_0$ entonces se emplea $Z = \frac{(\hat{p} - 0.5/n) - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$

Para cuando se trabaje con la cantidad x de la variable aleatoria X, solamente se agrega o disminuye 0.5 a la x; por lo tanto, se utilizarán las siguientes expresiones:

Si $x < np_0$ entonces se utiliza $Z = \dfrac{(x+0.5) - np_0}{\sqrt{np_0 q_0}}$

Si $x > np_0$ entonces se emplea $Z = \dfrac{(x-0.5) - np_0}{\sqrt{np_0 q_0}}$

Para solucionar la situación problema objeto de estudio a partir de la muestra aleatoria:

B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, D, D.

se procedería de la siguiente forma:

1) Planteamiento del sistema de hipótesis

H_o: p=0.5
H_1: $p > 0.5$

2) Se fija el nivel de significancia en el valor $\alpha = 0.05$.

3) La prueba es unilateral derecha.

4) La estadística de prueba se soporta ahora en una distribución normal estándar.

5) Decisión: si el p-valor es menor que $\alpha = 0.05$ entonces se rechaza la hipótesis nula; en caso contrario, se la acepta.

Como

$\hat{p} = \frac{x}{n} = \frac{18}{20} = 0.9$ resultó mayor que $p_0 = 0.5$ entonces se emplea

$$Z = \frac{(\hat{p} - 0.5/n) - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{(0.9 - 0.5/20) - 0.5}{\sqrt{\frac{(0.5)(0.5)}{20}}} = \frac{(0.9 - 0.025 - 0.5}{\sqrt{\frac{0.25}{20}}}$$

$$Z = \frac{0.375}{0.1118} = 3.35$$

En esta situación, el *p*-valor = $P(Z \geq 3.35) = 1 - 0.9996 = 0.0004$

Como el *p*-valor =0.0004 resultó inferior al nivel de significancia α = 0.05, entonces se rechaza la hipótesis nula y se acepta la hipótesis alternativa.

6) Con un nivel de significancia del 5 %, se concluye que *p* es mayor que 0.5; por lo tanto, sí existen diferencias significativas en la proporción poblacional *p* referida a la cantidad de unidades del producto A que resultan de buena calidad con respecto a la proporción de unidades defectuosas. Se ha comprobado que el porcentaje de unidades de buena calidad en la población es superior al 50 %.

En esta misma situación, un inspector de la empresa T, basado en su experiencia de supervisar el proceso de producción, afirma que la proporción (porcentaje) en la población de unidades del producto A es superior al 91 %; usar la muestra obtenida para aceptar o rechazar la afirmación del inspector con un nivel de significancia del 5 %.

Aquí se dispone de la muestra aleatoria de tamaño *n* = 20, dada por:

B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, B, D, D.

Se procedería del siguiente modo:

1) Planteamiento del sistema de hipótesis

H_o: p=0.91
H_1: $p > 0.91$

2) se fija el nivel de significancia en el valor $\alpha = 0.05$.

3) La prueba es unilateral derecha.

4) La estadística de prueba se soporta ahora en una distribución normal estándar.

5) Decisión: si el *p*-valor es menor que $\alpha = 0.05$ entonces se rechaza la hipótesis nula; en caso contrario, se la acepta.

Como

$\hat{p} = \frac{x}{n} = \frac{18}{20} = 0.90$ resultó menor que $p_0 = 0.91$ entonces se emplea

$$Z = \frac{(\hat{p} + 0.5/n) - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{(0.9 + 0.5/20) - 0.91}{\sqrt{\frac{(0.91)(0.09)}{20}}} = \frac{(0.9 + 0.025 - 0.91}{\sqrt{\frac{0.0819}{20}}}$$

$$Z = \frac{0.015}{0.06399} = 0.2344$$

En esta situación, el *p*-valor = $P(Z \geq 0.23) = 1 - 0.591 = 0.409$

Como el P-valor =0.409 resultó mayor que el nivel de significancia α = 0.05, entonces se acepta la hipótesis nula H_o: p=0.91.

6) Con un nivel de significancia del 5 %, se concluye que *p* es igual a 0.91; por lo tanto, el inspector no tiene la razón. En este contexto, la proporción poblacional p relacionada con la cantidad de unidades del producto A que resultan de buena calidad es del 91 %.

La siguiente situación problema ha sido adaptada de Burbano y Valdivieso (2016). Un importador de frutas chilenas afirma que el 96 % de las unidades de un lote grande que le acaba de llegar a su bodega están en buenas condiciones y el resto de la fruta se perderá porque ha arribado en malas condiciones. El vendedor chileno difiere de esa afirmación y le solicita que se tome una muestra aleatoria de 200 unidades de las mencionadas frutas; se encontró que 194 unidades de la muestra sí estaban en buenas condiciones; con la información de esa muestra se desea probar la afirmación del importador con un nivel de significancia del 5 %.

En esta situación, el tamaño de la muestra n = 200 puede considerarse grande. Los datos de la muestra aleatoria también presentan dos categorías, la primera conformada por las frutas que llegan en buenas condiciones a la bodega y la segunda constituida por las frutas arruinadas; en consecuencia, se puede utilizar la aproximación de una distribución binomial por una distribución normal estándar. El proceso de prueba de hipótesis para esta situación puede contener los siguientes pasos:

1) Planteamiento del sistema de hipótesis

H_o: p=0.96
H_1: p diferente de 0.96

2) Se fija el nivel de significancia en el valor $\alpha = 0.05$.

3) Ahora la prueba es bilateral.

4) La estadística de prueba se soporta ahora en una distribución normal estándar.

5) Decisión: si el *p*-valor es menor que $\alpha/2 = 0.025$ entonces se rechaza la hipótesis nula; en caso contrario, dicha hipótesis será aceptada.

Aquí resulta que $1- \alpha = 0.95$; $Z_{\alpha/2} = Z_{0.025} = -1.96$; $Z_{1-\alpha/2} = Z_{0.975} = 1.96$ como se ilustra en la Figura 3.5

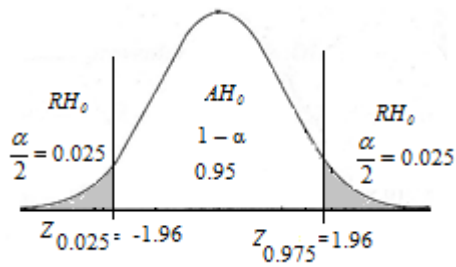

Figura 3.5. Prueba bilateral para la proporción
Fuente: los autores.

Como
$\hat{p} = \frac{x}{n} = \frac{194}{200} = 0.97$, ahora se emplea la siguiente estadística de prueba

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.97 - 0.96}{\sqrt{\frac{(0.96)(0.04)}{200}}} = \frac{0.01}{0.013856} = 0.7217$$

En esta situación, el *p*-valor = $P(Z \geq 0.7217) = 1 - 0.7246 = 0.2758$

Como el *p*-valor =0.2758 resultó mayor que el nivel de significancia α/2 = 0.025, entonces se acepta la hipótesis nula H_o: *p*=0.96.

En este contexto, también se observa que el valor de Z=0.7217 cae en la región AH_0 comprendida entre los valores -1.96 y 1.96 (ver Figura 3.5), por lo tanto, se acepta H_o: *p*=0.96.

6) Con un nivel de significancia del 4 %, se concluye que *p* es igual a 0.96; por lo tanto, el importador tiene la razón. En este contexto, la proporción poblacional *p* asociada a la cantidad de unidades de fruta que llega en buenas condiciones a la bodega es del 96 %.

3.3 Prueba chi-cuadrado para una muestra

Con frecuencia, cuando un investigador emprende sus tareas, su interés puede focalizarse en el número de individuos cuyos datos pueden clasificarse en una categoría de varias que recogen el total de la información. Por ejemplo, la opinión de un grupo de individuos puede clasificarse en una de cinco categorías: muy de acuerdo, de acuerdo, indiferente, en desacuerdo o muy en desacuerdo. En estas circunstancias, la prueba chi-cuadrado es utilizada para probar si existen diferencias significativas entre el número observado de individuos (datos) que se clasifican en cada categoría y el número esperado de datos en cada categoría, con la hipótesis nula.

La hipótesis establece la proporción de individuos que se clasifican en cada una de las categorías de una población de la que se presume provengan los datos; en otros términos, mediante la prueba de hipótesis nula se puede determinar cuáles son las frecuencias esperadas. La estadística de prueba corresponde a una chi-cuadrado, la cual permite

precisar si las frecuencias observadas están próximas a las frecuencias esperadas, en cuyo caso ha de aceptarse la hipótesis nula; de lo contrario, si tales diferencias son muy grandes, se ha de rechazar esta hipótesis.

La estadística de prueba puede expresarse de la siguiente manera:

$$Chi-cudrado = \sum_{i=1}^{k} \frac{\left(O_i - e_i\right)^2}{e_i}$$

Donde O_i corresponde al número observado de individuos clasificados en la categoría *i*, e_i hace referencia al número esperado de individuos en la categoría *i* en concordancia con la hipótesis nula; en este contexto, e_i se obtiene del cociente entre el total de datos *n* (tamaño de la muestra) dividido entre el número de categorías (*k*); el símbolo de la sumatoria indica que se han de sumar todas las *k* categorías. La frecuencia relativa esperada puede estimarse así: $f_i = e_i/n = 1/k$. Si se produce un número alto de concordancias entre las frecuencias observadas y esperadas, es posible que dicha suma sea pequeña, en cuyo caso el valor de chi-cuadrado resultará pequeño y por lo tanto se acepta la hipótesis nula; en caso contrario, si chi-cuadrado resulta lo suficientemente grande, entonces se rechaza esta hipótesis.

Como la estadística (estimador) chi-cuadrado con la hipótesis nula sigue una distribución teórica chi-cuadrado con *k-1* grados de libertad, si *p*-valor es menor que el nivel de significancia α, entonces se rechaza la hipótesis nula; de lo contrario, será aceptada. Es necesario mencionar que cuando *k* sea igual a 2, cada frecuencia esperada debe ser mayor o igual que 5; si *k* es mayor que 5, la chi-cuadrado no se utiliza cuando más del 20 % de las frecuencias esperadas sean menores que 5 (Cochran, 1954).

Algunas veces, las frecuencias esperadas aumentan al combinar categorías adyacentes cuando tal combinación tenga sentido (Burbano et al., 2019).

La siguiente situación problema ha sido planteada de forma semejante a la expuesta por Canavos (1988): el gerente de la empresa WA quiere saber si el número de trabajadores de su empresa que asisten a la oficina de talento humano para recibir consejería laboral se encuentra distribuido de manera equitativa durante los 5 días hábiles de cada semana. De una muestra aleatoria tomada durante 4 semanas completas de labor activa de sus trabajadores, se obtuvieron los siguientes datos sobre X: número de trabajadores que asistieron a la oficina de talento humano para recibir consejería laboral.

Lunes: 50
Martes: 36
Miércoles: 33
Jueves: 40
Viernes: 46

Con un nivel de significancia del 5 %, probar la hipótesis de que el número de trabajadores que asisten a la oficina de talento humano de la empresa WA se distribuyen de manera equitativa durante los 5 días de la semana.

El proceso de prueba de hipótesis para esta situación puede seguir los siguientes pasos:

1) Planteamiento del sistema de hipótesis

H_o: $f_i = 0.2$ para i=1, 2, 3, 4, 5
H_1: f_i difiere de 0.2 para algunos i=1, 2, 3, 4, 5

De forma equivalente, las hipótesis anteriores se pueden plantear de la siguiente forma:

H_o: $e_i = 41$ para i=1, 2, 3, 4, 5
H_1: e_i difiere de 41 para algunos i=1, 2, 3, 4, 5

2) Se fija el nivel de significancia en el valor $\alpha = 0.05$.

3) Se trata de una prueba unilateral derecha sobre una distribución teórica chi-cuadrado con *k-1* = 5-1 = 4 grados de libertad

$$\chi^2_{0.95,4} = 9.4877$$

4) La estadística de prueba se soporta en los datos de la muestra, con n = 205, $k = 5$; e_i = 205/5 = 41.

$$Chi = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \frac{(50-41)^2}{41} + \frac{(36-41)^2}{41} + \frac{(33-41)^2}{41} + \frac{(40-41)^2}{41} + \frac{(46-41)^2}{41}$$

$$Chi = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \frac{(9)^2}{41} + \frac{(-5)^2}{41} + \frac{(-8)^2}{41} + \frac{(-1)^2}{41} + \frac{(5)^2}{41}$$

$$Chi - cuadrado = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \frac{81}{41} + \frac{25}{41} + \frac{64}{41} + \frac{1}{41} + \frac{25}{41} = \frac{196}{41} = 4.78$$

5) Decisión: si el *p*-valor es menor que $\alpha = 0.05$ entonces se rechaza la hipótesis nula; en caso contrario, dicha hipótesis será aceptada.

El *p*-valor = P(Chi >4.78) =0.3106

Como el *p*-valor resultó mayor que 0.05 entonces se acepta la hipótesis nula H_o: $e_i = 41$.

Aquí también resulta que 1- $\alpha = 0.95$; $\chi^2_{0.95,4} = 9.4877$
Como chi-cuadrado = 4.78 resulta menor que 9.4877, entonces se decide aceptar la hipótesis nula. En esta situación problema, por cualesquiera de las dos formas se acepta la hipótesis nula.

6) Con un nivel de significancia del 5 %, se concluye que no existe ninguna razón para considerar que el número de trabajadores que asisten a la oficina de talento humano de la empresa WA no se distribuyen de manera equitativa durante los 5 días de la semana; es decir, que los mencionados trabajadores asisten de manera uniforme con valor esperado de 41 trabajadores por semana.

La siguiente situación problema ha sido adaptada de Siegel (1970): los individuos aficionados a las carreras equinas en torno a una pista circular piensan que hay ventajas para los equinos que ocupan determinadas posiciones en la posta. En una pista para 8 equinos, la posición 1 es aquella que se encuentra más próxima a la baranda interior de esa pista; en cambio, la posición 8 es la que queda en el lado más alejado de esa primera baranda (la más exterior). Los siguientes datos corresponden al número de triunfos logrados por los equinos en cada una de las 8 pistas.

Tabla 3.5. Número de triunfos de los equinos por posición en la pista

Pista	1	2	3	4	5	6	7	8
Número de triunfos	30	20	19	26	18	11	16	12

Fuente: los autores

Con un nivel de significancia del 5 %, se quiere probar si existe algún efecto en correspondencia con la posición en la cual compitieron los equinos durante un periodo determinado de carreras.

El proceso de prueba de hipótesis para esta situación puede contener los siguientes pasos:

1) Planteamiento del sistema de hipótesis

H_o: $f_i = 0.125$ para i=1, 2, 3, …, 8
H_1: f_i difiere de 0.125 para algunos i=1, 2, 3, …, 8

De forma equivalente, las hipótesis anteriores se pueden plantear de la siguiente forma:

H_o: $e_i = 19$ para i=1, 2, 3, …, 8
H_1: e_i difiere de 19 para algunos i=1, 2, 3, …, 8

2) Se fija el nivel de significancia en el valor $\alpha = 0.01$.

3) Se trata de una prueba unilateral derecha sobre una distribución teórica chi-cuadrado con k-1 = 8-1 = 7 grados de libertad

$$\chi^2_{0.95,7} = 14.0671$$

4) La estadística de prueba se soporta en los datos de la muestra, con n = 152, $k = 8$; e_i = 152/8 = 19.

$$Chi = \sum_{i=1}^{8} \frac{(O_i - e_i)^2}{e_i} = \frac{(30-19)^2}{19} + \frac{(20-19)^2}{19} + \frac{(19-19)^2}{19} + \frac{(26-19)^2}{19} + \frac{(18-19)^2}{19} + \frac{(11-19)^2}{19} + \frac{(16-19)^2}{19} \frac{(12-19)^2}{19}$$

$$Chi = \sum_{i=1}^{8} \frac{(O_i - e_i)^2}{e_i} = \frac{(11)^2}{19} + \frac{(1)^2}{19} + \frac{(0)^2}{19} + \frac{(7)^2}{19} + \frac{(-1)^2}{19} + \frac{(-8)^2}{19} + \frac{(3)^2}{19} \frac{(-7)^2}{19}$$

$$Chi - cuadrado = \frac{121}{19} + \frac{1}{19} + \frac{0}{19} + \frac{49}{19} + \frac{1}{19} + \frac{64}{19} + \frac{9}{19} + \frac{49}{19} = \frac{294}{19} = 15.4737$$

5) Decisión: si el *p*-valor es menor que $\alpha = 0.05$, entonces se rechaza la hipótesis nula; en caso contrario, dicha hipótesis será aceptada.

El *p*-valor = P(Chi >15.4737) =0.0304

Como el *p*-valor resultó menor que 0.05, entonces se rechaza la hipótesis nula H_o: $e_i = 19$.

Aquí también resulta que 1- $\alpha = 0.95$; $\chi^2_{0.95,7} = 14.0671$

Como chi-cuadrado = 15.4737 resulta mayor que 14.0671, entonces se decide rechazar la hipótesis nula. En esta situación problema, por cualquiera de las dos maneras se rechaza la hipótesis nula.

6) Con un nivel de significancia del 5 %, se concluye que sí existe un efecto significativo de la posición en la cual compiten los equinos en la pista circular analizada; es decir, que los mencionados equinos se ven afectados en el número de triunfos en correspondencia con la posición de donde inicien su competencia en la posta; así entonces, el valor esperado difiere de 19 triunfos para cada posición.

3.4 Prueba K-S de Kolmogorov-Smirnov para una muestra

Esta prueba se utiliza para analizar si los datos de una variable *X* continua se ajustan a una distribución de probabilidad específica. Por ejemplo, en

diversas ocasiones se requiere analizar si un conjunto de datos cuantitativos con *n* pequeño y en escala de razón o al menos en una escala de intervalo, provienen de una distribución normal con media μ conocida y desviación estándar σ también conocida. Así mismo, con esta prueba se puede determinar si los datos de una muestra aleatoria provienen de una distribución $F_0(x)$ de interés como la distribución exponencial, lognormal, Weibull, gamma, entre otras.

Esta prueba K-S consiste en comparar los valores obtenidos mediante la función empírica $\hat{F}_n(x)$ (función de distribución acumulativa). Las observaciones (datos) se ordenan de menor a mayor de la siguiente manera: $x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(n)}$, la función empírica se denota y define como sigue:

$$\hat{F}_n(x) = \begin{cases} 0 & \text{si } x < x_{(1)} \\ \dfrac{k}{n} & \text{si } x_{(k)} \leq x < x_{(k+1)} \\ 1 & \text{si } x \geq x_{(n)} \end{cases}$$

El proceso de prueba de hipótesis para esta situación puede seguir los siguientes pasos:

1) Planteamiento del sistema de hipótesis

H_o: $F(x) = F_0(x)$
H_1: $F(x)$ difiere de $F_0(x)$

2) Se fija el nivel de significancia en un valor α menor o igual que 0.05.

3) Se trata de una prueba bilateral.

4) La estadística de prueba se soporta en calcular el siguiente valor asociado a la prueba K-S, así (Burbano et al., 2019):

$$D = \underset{1 \le i \le n}{Sup} \left| \hat{F}_n(x_i) - F_0(x_i) \right|$$

En este contexto, el valor D representa la "diferencia absoluta más grande" que se pueda obtener entre la función acumulativa y la probabilidad calculada en el valor x_i sobre la distribución de probabilidad propuesta.

De acuerdo con Burbano *et al.* (2019), en distintos casos es posible utilizar la distribución asintótica para *D*, la cual se puede escribir para n lo suficientemente grande, en los siguientes términos:

$$\lim_{n \to \infty} P\left(D \le \frac{\lambda}{\sqrt{n}} = D_\alpha \right) = 1 - \exp\left(-2\lambda^2\right) = 1 - \alpha$$

En este caso, D_α se ha de seleccionar de tal modo que:

$$P(\text{Re}\,chazar\ \text{Ho/Ho es cierta}) = P(D > D_\alpha / \text{Los datos siguen la distribución propuesta}) = \alpha$$

Por ejemplo, para un nivel de significancia α = 0.05, el valor de λ se puede obtener de la siguiente ecuación:

$1 - exp(-2\lambda^2) = 1 - \alpha/2$

De la solución de esta ecuación resulta un valor λ = 1.36 aprox., en consecuencia, se tiene (Canavos, 1988):

$$D_\alpha = \frac{\lambda}{\sqrt{n}} = \frac{1.36}{\sqrt{n}}$$

5) Decisión: la región crítica o de rechazo de la hipótesis nula ha de satisfacer lo siguiente:

$$P\left(D > \frac{\lambda}{\sqrt{n}} = D_{\alpha} \right) = \alpha$$

En estas circunstancias, si D es mayor que D_{α} entonces se rechaza la hipótesis nula; en caso contrario, se aceptará esta hipótesis.

6) Obtención de una conclusión.

La siguiente situación problema se ha formulado de forma similar a la presentada en Lehmann y D'Abrera (1975): la variable X representa la cantidad de combustible en litros que en una muestra aleatoria de tamaño n =10 los vehículos de la marca A consumen por un lapso de tiempo L; los datos son los siguientes: 13.8, 13.9, 14, 14.2, 12.5, 12.8, 13, 13.4, 13.5, 13.6. Con un nivel de significancia del 5 % se quiere probar la hipótesis de que estos datos provienen de una distribución normal de parámetros $\mu = 13$ y $\sigma = 1$.

El proceso de prueba de hipótesis para esta situación puede incluir los siguientes pasos:

1) Planteamiento del sistema de hipótesis

H_o: $F(x) = F_0(x)$
H_1: $F(x)$ difiere de $F_0(x)$

Donde $F_0(x)$ corresponde a una distribución normal con $\mu = 13$ y $\sigma = 1$.

2) Se fija el nivel de significancia en el valor $\alpha = 0.05$.

3) Se trata de una prueba bilateral.

4) La estadística de prueba se soporta en calcular el siguiente valor vinculado a la prueba K-S, así:

$$D = \underset{1 \le i \le n}{Sup} \left| \hat{F}_n(x_i) - F_0(x_i) \right|$$

En primera instancia se ordenan los datos de menor a mayor: 12.5, 12.8, 13, 13.4, 13.5, 13.6, 13.8, 13.9, 14, 14.2, los cuales corresponden a la serie ordenada (por el rango o posición de cada dato) $x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(10)}$. Para cada dato se utiliza la estandarización

$$Z = \frac{x - \mu}{\sigma}$$

Para cada valor de *Z* obtenido, se determina su probabilidad en una distribución normal estándar; por ejemplo, para *x* = 12.5, su estandarización es

$$Z = \frac{12.5 - 13}{1} = -0.5$$

Su probabilidad a través de la distribución normal estándar es $F_0(x_{(1)}) = F_0(12.5) = 0.3085$.

Para *x*= 12.8, su estandarización es

$$Z = \frac{12.8 - 13}{1} = -0.2$$

Su probabilidad a través de la distribución normal estándar es $F_0(x_{(2)}) = 0.4207$. Para los demás valores de *x* se procederá de similar manera. Por medio de la función acumulativa o distribución empírica se obtienen los valores para $\hat{F}_n(x)$, por ejemplo, $\hat{F}_n(x_{(1)}) = \hat{F}_n(12.5) = \frac{1}{10} = 0.1$,

$\hat{F}_n(x_{(2)}) = \hat{F}_n(12.8) = \frac{2}{10} = 0.2$; así sucesivamente.

Es conveniente mencionar que, en caso de presentarse empates, se divide el número de empates entre el valor de *n*. Con estos resultados se conforma la Tabla 3.6.

Como *D* representa la "diferencia absoluta más grande", entonces para este caso *D* =0.2254.

En esta situación, $D_{\alpha} = 0.410$

5) Decisión: como *D* resultó menor que D_{α} entonces no se rechaza la hipótesis nula, se acepta que los datos de la variable *X* se ajustan a una distribución normal.

Tabla 3.6. Cálculo de la estadística *D*, prueba de Kolmogorov-Smirnov

$X_{(i)}$	$\hat{F}_n(x_{(i)})$	$F_0(x_{(i)})$	D
12.5	0.1	0.3085	0.2085
12.8	0.2	0.4207	0.2207
13	0.3	0.5	0.2
13.4	0.4	0.6554	0.2554
13.5	0.5	0.6915	0.1915
13.6	0.6	0.7257	0.1257
13.8	0.7	0.7881	0.0881
13.9	0.8	0.8159	0.0159
14	0.9	0.8413	0.0587
14.2	1.0	0.8849	0.1151

6) Con un nivel de significancia del 5 % se concluye que los datos de la variable *X* provienen de una distribución normal con $\mu = 13$ y $\sigma = 1$.

La situación problema que se expone a continuación guarda alguna similitud con la indicada por Canavos (1988): la variable *X* representa el número de respuestas correctas obtenidas en una muestra de 16 estudiantes en una prueba interna para el ingreso a la universidad A: 397,

400, 405, 408, 413, 425, 438, 453, 242, 265, 300, 323, 347, 353, 371, 388. Con un nivel de significancia del 5 % se quiere probar la hipótesis de que estos datos provienen de una distribución normal de parámetros μ = 375 y σ = 50.

El proceso de prueba de hipótesis para esta situación puede comprender los siguientes pasos:

1) Planteamiento del sistema de hipótesis

H_o: $F(x) = F_0(x)$
H_1: $F(x)$ difiere de $F_0(x)$

Donde $F_0(x)$ corresponde a una distribución normal con μ = 375 y σ = 50.

2) Se fija el nivel de significancia en el valor α = 0.05.

3) Se trata de una prueba bilateral.

4) La estadística de prueba se soporta en calcular el siguiente valor asociado a la prueba K-S, así:

$$D = \underset{1 \le i \le n}{Sup} \left| \hat{F}_n(x_i) - F_0(x_i) \right|$$

Ahora se ordenan los datos de menor a mayor: 242, 265, 300, 323, 347, 353, 371, 388, 397, 400, 405, 408, 413, 425, 438, 453, los cuales corresponden a la serie ordenada teniendo en cuenta el rango de cada dato, así: $x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(16)}$

Para cada dato se utiliza la estandarización

$$Z = \frac{x - \mu}{\sigma}$$

Para cada valor de Z obtenido se determina su probabilidad en una distribución normal estándar; por ejemplo, para x=242, su estandarización es

$$Z = \frac{242 - 375}{50} = -2.66$$

Su probabilidad a través de la distribución normal estándar es $F_0(x_{(1)}) = F_0(242) = 0.0039$.

Para x= 265, su estandarización es

$$Z = \frac{265 - 375}{50} = -2.2$$

Su probabilidad a través de la distribución normal estándar es $F_0(x_{(2)}) = F_0(265) = 0.0139$. Para los demás valores de x se procederá de similar manera. Por medio de la función acumulativa o distribución empírica se obtienen los valores para $\hat{F}_n(x)$, por ejemplo,

$\hat{F}_n(x_{(1)}) = \hat{F}_n(242) = \frac{1}{16} = 0.0625$,

$\hat{F}_n(x_{(2)}) = \hat{F}_n(265) = \frac{2}{16} = 0.125$; así sucesivamente.

Con estos resultados se conforma la Tabla 3.7.

Como D representa la "diferencia absoluta más grande", entonces para este caso D =0.1207.

En esta situación, D_α = 0.328

5) Decisión: como D resultó menor que D_α entonces no se rechaza la hipótesis nula, se acepta que los datos de la variable X se ajustan a una distribución normal.

6) Con un nivel de significancia del 5 %, se concluye que los datos de la variable X provienen de una distribución normal con con $\mu = 375$ y $\sigma = 50$.

Tabla 3.7. Cálculo de la estadística *D*, prueba de Kolmogorov-Smirnov

$X_{(i)}$	$\hat{F}_n(x_{(i)})$	$F_0(x_{(i)})$	D
242	0.0625	0.0039	0.0586
256	0.125	0.0139	0.1111
300	0.1875	0.0668	0.1207
323	0.25	0.1492	0.1008
347	0.3125	0.2877	0.0248
353	0.375	0.33	0.045
371	0.4375	0.4681	0.0306
388	0.5	0.6026	0.1026
397	0.5625	0.67	0.1075
400	0.625	0.6915	0.0665
405	0.6875	0.7257	0.0382
408	0.75	0.7454	0.0046
413	0.8125	0.7764	0.0361
425	0.875	0.8413	0.0337
438	0.9375	0.8962	0.0413
453	1	0.9406	0.0594

Es conveniente mencionar que este tipo de prueba no paramétrica también se puede realizar de forma directa mediante diversos paquetes estadísticos, entre ellos, el *software* libre R, el *software* con licencia SPSS, Minitab, economapas, SAS, o el lenguaje de programación Python, entre otros.