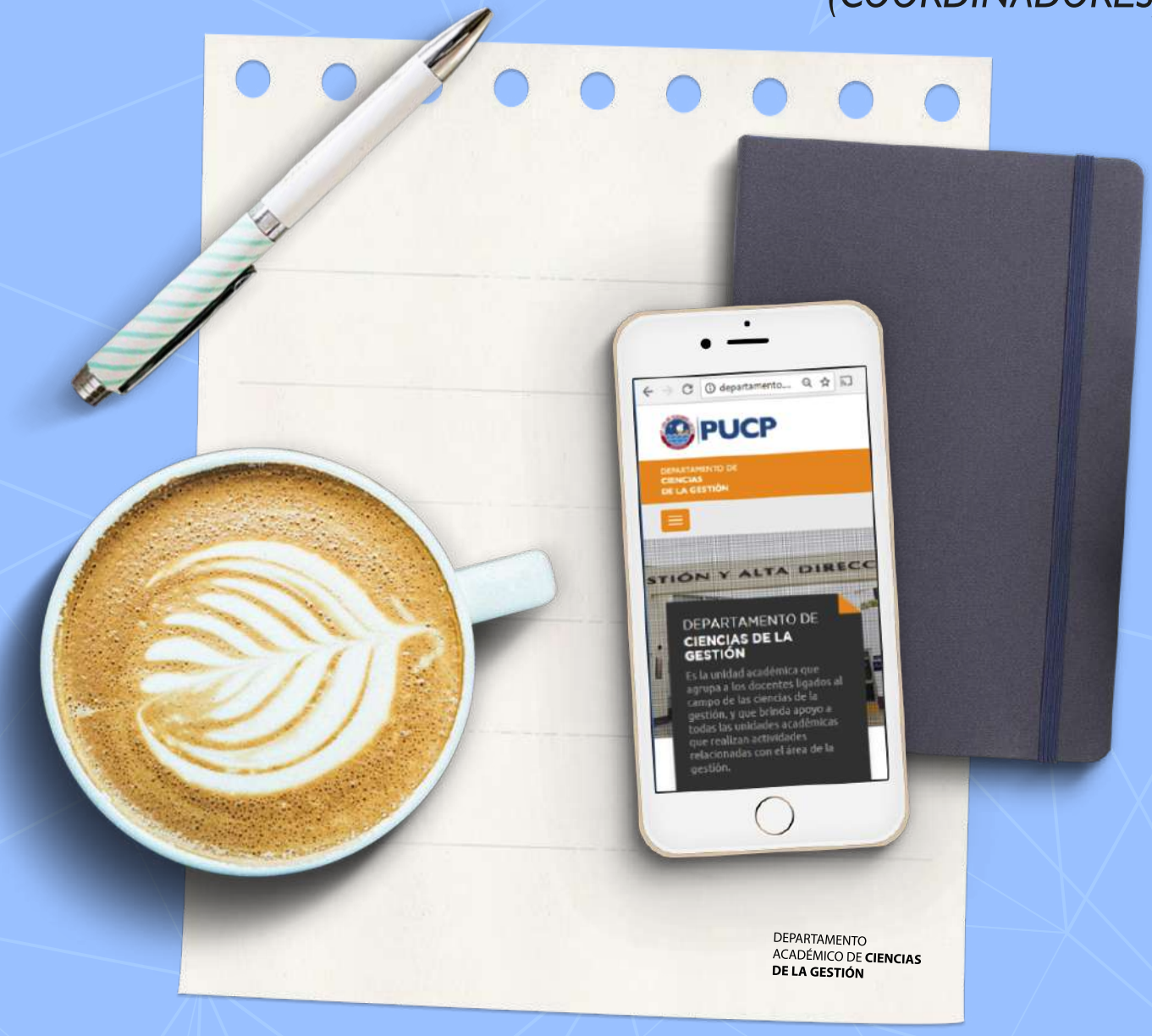
APUNTES DE CLASE #3
MÉTODOS DE INVESTIGACIÓN CUANTITATIVA

MILOS LAU, BERLAN RODRÍGUEZ, FÁTIMA PONCE, ANDRÉS YÁÑEZ Y MILAGROS MENDOZA (COORDINADORES)



DEPARTAMENTO
ACADÉMICO DE **CIENCIAS**
DE LA GESTIÓN

PUCP

APUNTES DE CLASE #3

MÉTODOS DE INVESTIGACIÓN CUANTITATIVA

Milos Lau
Berlan Rodríguez
Fátima Ponce
Andrés Yáñez Cárdenas
Milagros Mendoza Babilón
(Coordinadores)

Juan Azula Pastor
Hans Burkli Burkli
Alberto Díaz Milla
Jorge Hernández Garavito
Denzel Glandel Tafur
Jessica Leguía Llanos
Diego Mendoza García
Marvin Padilla Trujillo
Javier Vidal Cuba

DEPARTAMENTO
ACADÉMICO DE **CIENCIAS**
DE LA **GESTIÓN**

PUCP

MÉTODOS DE INVESTIGACIÓN CUANTITATIVA

Publicación: Apuntes de Clase #3

Dirección: Av. Universitaria 1801, San Miguel Lima

Teléfono: (51-1) 626-2000

www.pucp.edu.pe

Índice

Índice de tablas

Índice de figuras

Presentación

Este documento contiene ejercicios y casos tipo realizados en el curso de nivel pregrado Métodos de Investigación Cuantitativa, el cual es dictado en la Facultad de Gestión y Alta Dirección de la Pontificia Universidad Católica del Perú (PUCP). Es importante mencionar que este curso capacita al alumno para que desarrolle métodos de recojo de datos que se adecuen a la problemática a investigar, construya mediciones, aplique herramientas alineadas con los objetivos de la investigación, estructure la información y presente resultados a fin de participar en estudios cuantitativos relevantes que permitan mejorar la gestión social, pública o privada de las organizaciones en las que se desenvuelva a nivel profesional.

Asimismo, la introducción al uso de *software* como Excel y SPSS ayuda a que el alumno fortalezca su capacidad de análisis inferencial mediante el manejo de bases de datos; ello con la finalidad de contribuir a su formación como tomador de decisiones y creador de valor.

En este sentido, el objetivo de estos apuntes es mejorar la comprensión y aplicación de las herramientas empleadas en el curso mediante el desarrollo de ejercicios y casos aplicados. La comprensión de las mismas supone que el estudiante haya aprendido previamente los conceptos básicos de los cursos de estadística y matemáticas proporcionados en Estudios Generales Letras. Este documento está organizado de acuerdo con el orden de los temas consignados en el sílabo del curso.

Fátima Ponce
Agosto de 2019

1. Métodos de investigación, usos e importancia

En este primer capítulo se busca que el alumno logre reconocer las diferencias entre los enfoques de investigación cuantitativa y cualitativa mediante el uso de ejercicios aplicados a la gestión. Para esta primera parte, se tiene el objetivo de reforzar el planteamiento de un problema de investigación, sus objetivos y definición de variables (tipos y operacionalización). Asimismo, en este capítulo se presenta el uso de los programas Excel y SPPS como herramientas de apoyo para el análisis de la data proporcionada en cada ejercicio.

Palabras clave

- Enfoques de investigación
- Matriz de variables
- Operacionalización de variables
- Comandos básicos de SPSS
- Comandos básicos de Excel

Ejercicio I: Construcción de matriz de variables

Usted ha sido contratado por la Facultad de Gestión y Alta Dirección para realizar una investigación que pruebe la siguiente hipótesis general: "El rendimiento de los alumnos en las clases prácticas (dirigidas y calificadas) durante el ciclo tiene relación positiva con la nota final del curso".

Considerando esta hipótesis, elabore una matriz de variables con 2 variables intermedias y sus respectivos indicadores. Además, señale el tipo de variable y su definición instrumental.

Tabla 1-1: Matriz de variables

Variable general	Variable intermedia	Variable específica o indicador	Tipo de variable	Definición instrumental
(Característica general que queremos conocer)	*(Aspecto o elemento en el que se descompone la variable general)*	*(Unidad de medida a ser observada)*	*(Por su nivel de medición)*	*(Fuentes de las que se obtendrá la información)*
Nota final del curso				

Ejercicio II: Análisis y procesamiento de datos en SPSS y Excel

Para realizar este ejercicio, emplee el archivo en Excel "BD_1".

Realice un análisis cuantitativo a partir de la data proporcionada en la BD_1. La información que encontrará está relacionada con ingresos monetarios, estado civil, grado de instrucción y género. Esta base corresponde a la Encuesta de Vida y Pobreza-ENAHO / Año: 2015 / Periodo: Trimestre 4 / Módulo 5: Empleo e Ingresos publicada por el Instituto Nacional de Estadística e Informática (INEI).

Para este análisis, utilice el archivo en Excel proporcionado para esta unidad que podrá encontrar en el *link* con el *drive* compartido en la bibliografía que aparece al final de este documento.

A continuación, se muestran relaciones entre variables que debe tener en cuenta en su análisis. Recuerde que para este análisis puede apoyarse en el uso de Excel o SPSS.

- Ingresos y género
- Grado de instrucción y estado civil
- Género y grado de instrucción
- Estado civil e ingresos

A partir del análisis realizado con las variables listadas arriba, valide las siguientes afirmaciones con verdadero o falso, según corresponda:

a) El porcentaje de personas casadas con estudios superiores es mayor que el de solteros con estudios superiores. ( )

b) El ingreso promedio mensual de las mujeres es igual al de los hombres. ( )

c) Hay un mayor número de mujeres con estudios superiores que con primaria, secundaria y sin nivel de estudios. ( )

d) El ingreso promedio mensual de los divorciados es menor que el de los viudos. ( )

e) No se evidencian estadísticas de personas divorciadas sin nivel de estudios. ( )

Caso aplicado: Diagnóstico para la promoción de proyectos de desarrollo

Una entidad de cooperación internacional desea promover proyectos de desarrollo bajo el enfoque de capacidades en Latinoamérica. Con esa consigna, ha contratado a un grupo de investigación para que, antes de emprender cualquier acción, realice un diagnóstico rápido de la situación general en estos países. En el Perú, el grupo de investigación hizo una revisión rápida de la literatura y data disponible; asimismo, planteó varias líneas de investigación.

a) Tipos de investigación

De acuerdo con los contenidos listados en la tabla "Enfoque de investigación" que se encuentra a continuación, señale el tipo de enfoque de investigación (cuantitativo, cualitativo o mixto) al que se encuentran asociados los siguientes enunciados:

Tabla 1-2: Enfoque de investigación

Enunciado	Enfoque de investigación
"Se plantea recoger información sobre las horas de trabajo semanal de la/el jefe del hogar, el monto total de sus ingresos y cómo lo distribuye en sus gastos mensuales".	
"Para evaluar el impacto de la educación en las condiciones de vida de las familias, se desea conocer las historias de vida de los hijos, haciendo énfasis en su desarrollo académico (fortalezas y limitaciones) y la toma de decisiones de los padres respecto a la educación de sus hijos".	
"El objetivo principal es determinar, a través de entrevistas a profundidad, cuáles son los bienes más valorados por las familias y, así, tener una mejor aproximación a las condiciones de vida de las familias y posibles alternativas de mejora".	
"Una vez determinados los ingresos familiares y comparados con la media de la localidad, se busca saber cómo es percibida esa diferencia en las familias".	
"Ya que se apunta a la creación de un perfil socioeconómico basado en las condiciones de vida, se aplicarán encuestas que identifiquen qué bienes se encuentran en el hogar, cuántas habitaciones hay y el número de personas residentes en la casa".	
"Esta parte del estudio incluirá un acompañamiento a las madres de familia en varios momentos del día, de manera que se observe cuál es la importancia de las labores domésticas para el hogar y su influencia en las condiciones de vida de la familia".	
"A través de una serie de indicadores, se pretende llevar a cabo una medición del aporte económico que supone el trabajo doméstico y su influencia en las condiciones de vida de la familia".	

b) Definición de población y muestra

Considerando el primer enfoque de investigación de la tabla anterior, se plantea recoger información sobre las horas de trabajo semanal de la/el jefe de hogar, el monto total de sus ingresos y cómo lo distribuye en sus gastos mensuales. Teniendo en cuenta que la comunidad donde se planea recoger la información cuenta con 10 000 familias, señale lo siguiente:

b.1. ¿Quiénes son los elementos de estudio de la investigación (población)?

b.2. ¿Quiénes van a proveer la información que necesitamos? ¿Es viable recoger información en toda la población? ¿Qué recursos nos demandaría hacer esto?

c) Tipos de variable y operacionalización[1]

Los investigadores encontraron que una forma de medir las condiciones de vida era evaluando el acceso a servicios básicos; mientras que las capacidades por educación podían ser medidas con el nivel educativo; y la solvencia económica, con los ingresos. Además, todos estos datos se encontraban en la Encuesta Nacional de Hogares (ENAHO) que el INEI realiza cada año. Algunas de las variables más importantes resultaron ser "Condiciones de vida", "Solvencia económica" y "Educación".

A continuación, se detallan algunas de las variables o indicadores que se usarán en el estudio. Pero, antes de iniciar, los investigadores deben definir qué tipo de variable están usando.

- Años de escolaridad de la/el jefe de hogar
- Gasto promedio mensual en artículos de lujo
- Rango de gasto promedio mensual en educación y salud
- Último grado de instrucción alcanzado por el padre
- Acceso a los servicios básicos
- Nivel socioeconómico
- Gasto promedio mensual en alimentos
- Número de habitaciones en el hogar
- Ingreso promedio mensual
- Número de horas de trabajo semanal
- Número de miembros en el hogar

[1] La operacionalización de una variable consiste en definir las variables específicas o indicadores que la componen, así como su tipo y clasificación.

Para tener un planteamiento más ordenado y claro, los investigadores iniciaron el trabajo de operacionalizar las variables en la siguiente tabla según las variables generales e intermedias ya establecidas.

Tabla 1-3: Operacionalización de variables

Variable general	Variable intermedia	Variable específica o indicador	Tipo y clasificación de variable
Condiciones de vida	Características de la vivienda		
Educación	Nivel educativo		
Solvencia económica	Empleo e ingresos		
	Capacidad de gasto		

Solucionario

Ejercicio I: Construcción de matriz de variables

Tabla 1-4: Solución matriz de variables

Variable general	Variables intermedia	Variable específica o indicador	Tipo de variable	Definición instrumental
Nota final del curso	Aprovechamiento de PD	Asistencia a las PD	Cuantitativa Cualitativa	Registro de asistencia
		Participación en clase	Cualitativa	Evaluación docente
	Rendimiento PC	Promedio de notas PC	Cuantitativa	Registro de notas

Ejercicio II: Análisis y procesamiento de datos en SPSS y Excel

Para este ejercicio, emplee el archivo en Excel "SOL_1".

Antes de empezar con el análisis de las variables, es importante que entienda cómo se debe trabajar con una base de datos. Para esto, utilice el archivo en Excel proporcionado para esta unidad ("BD_1") y apóyese en el uso de SPSS o en el mismo Excel, en su defecto.

¿Cómo ingreso al *software*?

En SPSS:

- **Abrir el *software*:**
  Busque en la barra de inicio de Windows el *software* "IBM SPSS Statistics".
- **Abrir un archivo/base de datos:**
  Antes de seguir este proceso, asegúrese de que la base tenga en la primera fila el nombre de las variables y, a partir de la segunda fila, los datos de la base.

  Archivo > Abrir > Datos > Archivos de tipo: Excel o SPSS Statistics (según el tipo de la BD) > Señalar el disco/ruta donde fue descargada o guardada la BD > Abrir > Aceptar > Corrobore que en "Vista de variables" aparezcan los siguientes datos correctos: nombre de la variable, etiqueta, valores y medida según corresponda.

- **Copiar una base de datos:**
  Cuando se tiene problemas para importar la base a SPSS, copiar la base a la vista de

datos es una opción.

Seleccione la base (solo los datos, no la fila con el nombre de las variables) > Copie la base (CTRL+C) y pegue la base en "Vista de datos" (CTRL+V) > Vaya a "Vista de variables" y complete los siguientes datos: nombre de la variable, etiqueta, valores y medida según corresponda.

En Excel:

- **Abrir el *software*:**
  Busque en la barra de inicio de Windows el *software* "Excel".

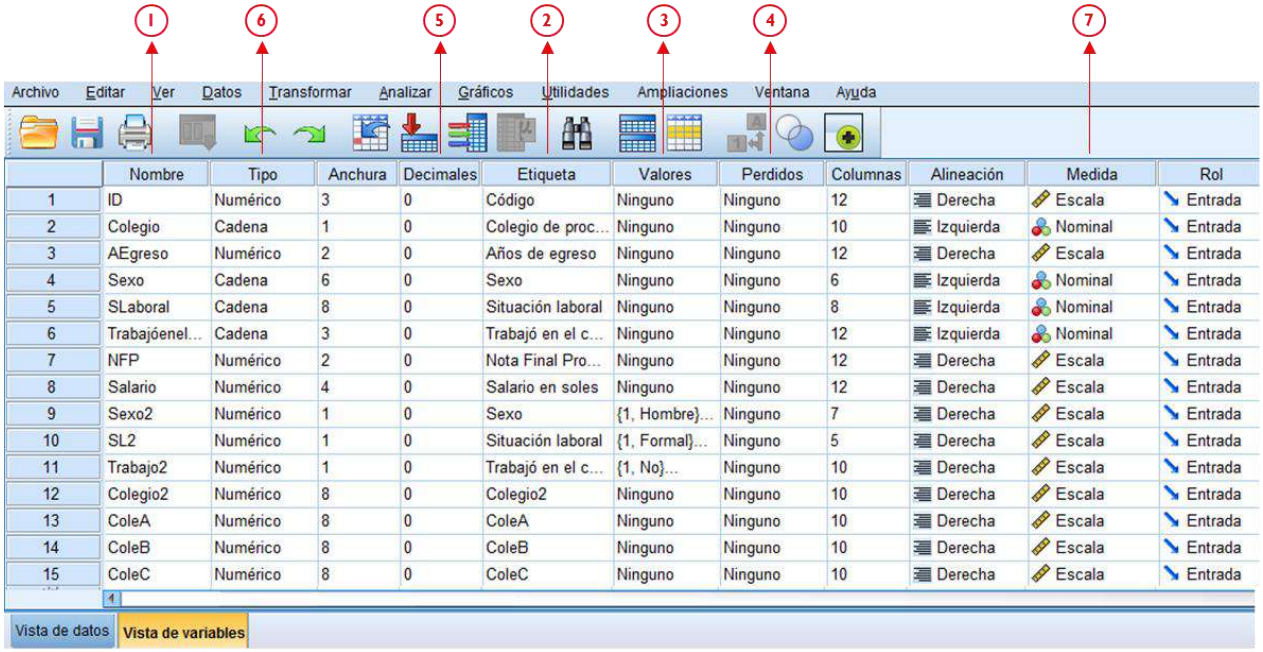
¿Cómo trabajar con una base de datos?

Independientemente del *software* que use, es importante que entienda las variables que le están entregando en la base de datos: qué significan y cuál es la unidad en la que están expresadas. A continuación, le damos algunos consejos que le serán de ayuda para iniciarse en la lectura de base de datos.

En SPSS:

SPSS tiene tres vistas con las que se familiarizará en cuanto comience a usar este *software*: "Vista de datos", "Vista de variables" y "Resultados". La primera vista es donde aparecerán todos los datos cargados a partir de la primera fila, es decir, las filas son los casos y las columnas son las variables. La segunda vista es donde aparecerá el detalle de las variables que están registradas en la vista de datos. La tercera vista aparecerá cuando se comience a ejecutar tareas; en esta aparecerá la sintaxis de las tareas que se ejecutan y los resultados de los análisis o pruebas que se corran en el *software*. Para entender cómo se trabaja con una base de datos, a continuación se tomarán de referencia los elementos que aparecen en la "Vista de variables".

Figura 1-1: Vista de variables en SPSS



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	ID	Numérico	3	0	Código	Ninguno	Ninguno	12	Derecha	Escala	Entrada
2	Colegio	Cadena	1	0	Colegio de proc...	Ninguno	Ninguno	10	Izquierda	Nominal	Entrada
3	AEgreso	Numérico	2	0	Años de egreso	Ninguno	Ninguno	12	Derecha	Escala	Entrada
4	Sexo	Cadena	6	0	Sexo	Ninguno	Ninguno	6	Izquierda	Nominal	Entrada
5	SLaboral	Cadena	8	0	Situación laboral	Ninguno	Ninguno	8	Izquierda	Nominal	Entrada
6	Trabajóenel...	Cadena	3	0	Trabajó en el c...	Ninguno	Ninguno	12	Izquierda	Nominal	Entrada
7	NFP	Numérico	2	0	Nota Final Pro...	Ninguno	Ninguno	12	Derecha	Escala	Entrada
8	Salario	Numérico	4	0	Salario en soles	Ninguno	Ninguno	12	Derecha	Escala	Entrada
9	Sexo2	Numérico	1	0	Sexo	{1, Hombre}...	Ninguno	7	Derecha	Escala	Entrada
10	SL2	Numérico	1	0	Situación laboral	{1, Formal}...	Ninguno	5	Derecha	Escala	Entrada
11	Trabajo2	Numérico	1	0	Trabajó en el c...	{1, No}...	Ninguno	10	Derecha	Escala	Entrada
12	Colegio2	Numérico	8	0	Colegio2	Ninguno	Ninguno	10	Derecha	Escala	Entrada
13	ColeA	Numérico	8	0	ColeA	Ninguno	Ninguno	10	Derecha	Escala	Entrada
14	ColeB	Numérico	8	0	ColeB	Ninguno	Ninguno	10	Derecha	Escala	Entrada
15	ColeC	Numérico	8	0	ColeC	Ninguno	Ninguno	10	Derecha	Escala	Entrada

¿Cómo trabajar con una base de datos en SPSS?

a. **Entienda las variables:**
Por lo general, las variables de la BD están explicadas o contextualizadas en un caso de estudio, por lo que es importante realizar una correcta lectura e interpretación del caso en mención. El objetivo de este primer paso es entender cómo se define esta variable y cuál es su unidad de medida. Los elementos que le ayudarán a lograr este primer objetivo son el nombre de la "Variable" (1) y su "Etiqueta" (2), en la cual aparece una pequeña descripción más precisa del significado de la variable. Para conocer cuál es la unidad de medida de la variable, tiene que identificar con qué categorías cuenta. Para esto, puede desglosar la casilla "Valores" (3) y editar en caso de que necesite cambiar los valores de su codificación.

b. **Valide si hay valores perdidos:**
Estos valores deben ser definidos por el usuario y son señalados en la columna de "Perdidos" (4), para que luego tengan un tratamiento especial y se excluyan de la mayoría de los cálculos en adelante. Entre los valores (3) que se podrían considerar como perdidos están las siguientes categorías y sus respectivos códigos: "Otros" (cod. 94), "Ninguno" (cod. 96), "No precisa" (cod. 99, 9), etc.

c. **Valide los decimales:**
Una vez entendido cuáles son las categorías que contemplan cada variable y su

unidad de medida, es conveniente revisar cuántos "Decimales" (5) está considerando cada variable. En caso de que haya más de los necesarios, se puede editar este valor por cada variable.

d. Valide el tipo y la medida de la variable:
Dependiendo de la naturaleza de la variable, se debe especificar en "Tipo" (6) si la variable es numérica, cadena, etc., mientras que en tipo de "Medida" (7) se ingresará el tipo de variable según su naturaleza: escalar, nominal u ordinal.

En Excel:

Excel es un programa con hojas de cálculo que te permite gestionar un volumen de data importante y que, además, cuenta con herramientas para analizarla. Por lo general, una base de datos empezará sus registros desde la celda "A1", en donde la primera fila tiene el nombre de las variables y, a partir de la segunda fila, se tiene a los casos, como se muestra en la siguiente figura. A continuación, se explica qué pasos podría tomar en cuenta para lograr una mejor lectura de bases de datos en Excel.

Figura 1-2: Vista de datos en Excel

	A	B	C	D
1	Encuesta	Sexo (M=0, H=1)	Días de estadía	Gasto Estimado
2	1	Hombre	6	165.6
3	2	Mujer	2	33.8
4	3	Hombre	5	235.5
5	4	Hombre	1	35.3
6	5	Mujer	2	87.6
7	6	Hombre	2	78.6
8	7	Hombre	2	92.4
9	8	Mujer	3	112.2
10	9	Hombre	6	117
11	10	Mujer	7	338.8
12	11	Hombre	7	250.6
13	12	Mujer	4	135.6
14	13	Hombre	5	249.5
15	14	Mujer	1	35.8
16	15	Mujer	1	50.6
17	16	Mujer	3	108.6
18	17	Hombre	5	172.5
19	18	Mujer	3	69

¿Cómo trabajar con una base de datos en Excel?

a. Entienda las variables:
Por lo general, las variables de la BD están explicadas o contextualizadas en un caso de estudio, por lo que es importante realizar una correcta lectura e interpretación del caso en mención. El objetivo de este primer paso es entender cómo se define

esta variable y cuál es su unidad de medida. Los elementos que le ayudarán a lograr este primer objetivo son el nombre de la variable y las categorías/valores con los que cuenta cada una.

b. Identifique cuáles son valores de cada variable:
Seleccione la fila donde aparecen los nombres de la variable > Datos > Filtro > Clic en la flecha de cada variable para que aparezca la lista de valores.
Otra alternativa es crear una tabla dinámica con todos registros de la BD y ver los valores en formato "Cuenta".

c. Identifique el tipo de variable:
Para facilitar el análisis del comportamiento de cada variable debe identificar cuál es su naturaleza (cualitativa o cuantitativa) y su tipo (nominal, ordinal, discreta o continua); esto le ayudará a decidir con qué herramientas puede trabajar cada una de ellas.

Con respecto de la parte del análisis de la base de datos propuesta en el ejercicio II, a continuación se muestran algunos comandos que le serán de utilidad.

En SPSS:

- **Crear tablas de frecuencia:**
  Analizar > Estadísticos descriptivos > Frecuencias > Inserte las variables a analizar.
- **Crear tablas cruzadas:**
  Analizar > Estadísticos descriptivos > Tablas cruzadas > Inserte las variables a analizar (fila y columna) > Aceptar.

En Excel:

- **Crear tablas dinámicas:**
  Seleccione la BD > Insertar > Tabla dinámica > Elija dónde desea colocar la tabla: Nueva hoja o en la hoja existente (seleccione un rango de celdas que ocupará la tabla) > Aceptar > Arrastre las variables a analizar en cuadrante columna, fila o filtro y determine cómo trabajará con los valores: cuenta, promedio, etc.

Las respuestas del ejercicio II son las siguientes. Para obtener más detalles sobre la solución de este ejercicio en Excel, tome de referencia la base "SOL_1" (Hoja Ejercicio 1.2) del *drive* compartido en la bibliografía de este documento.

a) El porcentaje de personas casadas con estudios superiores es mayor que el de solteros con estudios superiores. (V)

b) El ingreso promedio mensual de las mujeres es igual al de los hombres. (F)

c) Hay un mayor número de mujeres con estudios superiores que con primaria, secundaria y sin nivel de estudios. (V)

d) El ingreso promedio mensual de los divorciados es menor que el de los viudos. (F)

e) No se evidencian estadísticas de personas divorciadas sin nivel de estudios. (V)

Caso aplicado: Diagnóstico para la promoción de proyectos de desarrollo

a) Tipos de investigación

Tabla 1-5: Enfoque de investigación

Enunciado	Enfoque de investigación
"Se plantea recoger información sobre las horas de trabajo semanal de la/el jefe del hogar, el monto total de sus ingresos y cómo lo distribuye en sus gastos mensuales".	Cuantitativo, debido a que el enfoque requiere que se recoja información de variables cuantitativas.
"Para evaluar el impacto de la educación en las condiciones de vida de las familias, se desea conocer las historias de vida de los hijos, haciendo énfasis en su desarrollo académico (fortalezas y limitaciones) y la toma de decisiones de los padres respecto a la educación de sus hijos".	Cualitativo, debido a que se desea profundizar en el recojo de información de casos específicos.
"El objetivo principal es determinar, a través de entrevistas a profundidad, cuáles son los bienes más valorados por las familias y, así, tener una mejor aproximación a las condiciones de vida de las familias y posibles alternativas de mejora".	Cualitativo. En el enunciado se señala que se aplicarán entrevistas a profundidad, herramienta que es predominante en este enfoque de investigación.
"Una vez determinados los ingresos familiares y comparados con la media de la localidad, se busca saber cómo es percibida esa diferencia en las familias".	Mixto. En la parte inicial del enunciado se plantea el recojo de información cuantitativa. Luego se señala que se desea profundizar en los resultados, para lo que se usarían instrumentos cualitativos.
"Ya que se apunta a la creación de un perfil socioeconómico basado en las condiciones de vida, se aplicarán encuestas que identifiquen qué bienes se encuentran en el hogar, cuántas habitaciones hay y el número de personas residentes en la casa".	Cuantitativo, ya que se plantea el uso de una encuesta, instrumento predominante en este enfoque. Además, se señala que se recogerá información de variables cuantitativas.
"Esta parte del estudio incluirá un acompañamiento a las madres de familia en varios momentos del día, de manera que se observe cuál es la importancia de las labores domésticas para el hogar y su influencia en las	Cualitativo, ya que se plantea un proceso de recojo de información exhaustivo con el que se recogerá información a detalle. Esto sugiere un enfoque cualitativo.

Enunciado	Enfoque de investigación
condiciones de vida de la familia".	
"A través de una serie de indicadores, se pretende llevar a cabo una medición del aporte económico que supone el trabajo doméstico y su influencia en las condiciones de vida de la familia".	Cuantitativo. Si bien el enunciado es bastante sencillo, se podría sugerir que el enfoque es predominantemente cuantitativo ya que se va a recoger variables cuantitativas como "aporte económico".

b) Definición de población y muestra

b.1. ¿Quiénes son los elementos de estudio de la investigación (población)?

Según lo que indica el caso, los elementos de estudio o población de interés son el jefe de hogar. Al mencionar que la comunidad donde se planea recoger información cuenta con 10 000 familias, se asume que se habla de un total de 10 000 jefes de hogar.

b.2. ¿Quiénes van a proveer la información que necesitamos? ¿Es viable recoger información en toda la población? ¿Qué recursos nos demandaría esto?

Si se deseara recoger información de los 10 000 jefes de familia, el proceso podría ser muy costoso, además que el flujo de información podría ser muy tedioso para trabajar. En este caso se sugiere realizar muestreo. Un muestreo es un método estadístico que permite elegir un grupo representativo de la población de tal manera que no es necesario recoger la información de todos.

c) Tipos de variable y operacionalización[2]

Tabla 1-6: Operacionalización de variables

Variable general	Variable intermedia	Variable específica o indicador	Tipo y clasificación de variable
Condiciones de vida	Características de la vivienda	Número de miembros en el hogar	Cuantitativa. Discreta, Razón
		Acceso a los servicios básicos	Cualitativa. Nominal

[2] La operacionalización de una variable consiste en definir las variables específicas o indicadores que la componen, así como su tipo y clasificación.

Variable general	Variable intermedia	Variable específica o indicador	Tipo y clasificación de variable
		Número de habitaciones en el hogar	Cuantitativa. Discreta, Razón
		Tipo de material de la vivienda	Cualitativa. Nominal
Educación	Nivel educativo	Años de escolaridad de la/el jefe de hogar	Cuantitativa. Continua, Razón
		Último grado de instrucción alcanzado por el padre	Cualitativa. Ordinal
Solvencia económica	Empleo e ingresos	Ingreso promedio mensual	Cuantitativa. Continua, Razón
		Número de horas de trabajo semanal	Cuantitativa. Continua, Razón
		Nivel socioeconómico	Cualitativa. Ordinal
	Capacidad de gasto	Gasto promedio mensual en alimentos	Cuantitativa. Continua, Razón
		Gasto promedio mensual en artículos de lujo	Cuantitativa. Continua, Razón
		Rango de gasto promedio mensual en educación y salud	Cualitativa. Ordinal

Lecturas recomendadas

- Lind, D. A., Marchal, W. G., y Wathen, S. A. (2008). *Estadística aplicada a los negocios y la economía* (13.ª ed., cap. 1). México: McGraw-Hill Interamericana.
- Levin, R., y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed. revisada, cap. 1). México: Pearson Educación de México; Prentice Hall.
- Hernández, R., Fernández, C., y Baptista, P. (2010). *Metodología de la investigación* (cap. 1). México DF: McGraw-Hill.

2. Principales fuentes de información

Este capítulo tiene el objetivo de identificar cuáles son las principales fuentes de investigación y cuáles son sus técnicas de recojo de información. Asimismo, se hace particular énfasis en la necesidad de trabajar con técnicas de muestreo con la finalidad de optimizar los recursos con los que uno cuenta como investigador.

Palabras clave

- Muestreo aleatorio simple
- Fuentes de información primarias y secundarias
- Muestra y población

Ejercicio I: Muestreo aleatorio simple

Para este ejercicio, emplee el archivo en Excel "BD_2".

El Ministerio de Salud ha alertado a las diferentes regiones y unidades ejecutoras sobre la problemática de la alta tasa de embarazo adolescente en el país [3]. Como consecuencia, los Gobiernos regionales (en especial, los de las regiones con las mayores tasas[4]) han decidido invertir en la formulación de distintos programas sociales que ataquen la problemática en mención, tales como planes de educación sexual, mejora del servicio de atención y distribución de métodos anticonceptivos en establecimientos de salud, etc.

Por ejemplo, el Gobierno regional de Loreto ha decidido trabajar conjuntamente con un equipo de consultores para diseñar e implementar los programas antes mencionados. No obstante, debido al presupuesto limitado con el que cuenta la región, debe priorizar las provincias en las cuales deban ejecutarse dichos programas.

En ese sentido, con el fin de determinar las provincias a priorizar, el equipo de consultores al que usted pertenece ha revisado investigaciones preliminares sobre la

[3] Una tasa es un coeficiente que muestra la proporción que un evento representa del total. En este caso, la tasa de nacidos de madres adolescentes se halla dividiendo el número de nacidos de madres adolescentes sobre el total de nacidos en un determinado territorio.

[4] Loreto, Ucayali, Amazonas y San Martín con 26.66, 22.92, 21.33 y 20.12 %, respectivamente. Fuente: INEI-ENAHO (2012-2016), y ENDES (2010-2016). Extraído de INFOMIDIS. Consulta del 22 de agosto del 2017. Recuperado de http://sdv.midis.gob.pe/Infomidis/#/indicadoresEmblematicos

problemática en la región y, además, ha recopilado datos del INEI sobre el número de niños nacidos vivos de madres adolescentes por distritos[5].

Se sabe que una de las cuatro provincias seleccionadas es la provincia de Loreto, y que el número aproximado de mujeres entre 15 y 19 años es, aproximadamente, 3250.

Con el dato anterior, se le pide calcular el tamaño de la muestra para crear una línea de base sobre las mujeres adolescentes de la provincia de Loreto, considerando un error del 4 % y la máxima variabilidad en la distribución.

Ejercicio II: Muestreo aleatorio simple

Para este ejercicio, emplee el archivo en Excel "BD_2".

(Continúa del enunciado del Ejercicio I)

Se desea realizar un monitoreo continuo a las tasas de nacidos de madres adolescentes. **Como última tarea, se le encomienda que realice el muestreo adecuado para dicho propósito, teniendo en cuenta que debe ser lo más representativo posible de las provincias.** A continuación, se muestra una tabla con el número de mujeres adolescentes por provincia de Loreto:

Tabla 2-1: Madres adolescentes por provincia

Provincia	# de mujeres adolescentes
Maynas	28 592
Alto Amazonas	5548
Loreto	3369
Mariscal Ramón Castilla	3733
Requena	3610
Ucayali	3569
Datem del Marañón	3195

Caso aplicado: Líneas móviles

Para este caso, emplee el archivo en Excel "BD_2".

[5] INEI (2014). Maternidad en la adolescencia 2012. Lima. Consulta del 23 de agosto del 2017. Recuperado de https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1184/libro.pdf

En los últimos años, el Gobierno del Perú viene implementando las medidas necesarias para mejorar los servicios de las telecomunicaciones en todo el territorio nacional. En ese sentido, entre las principales reformas que viene trabajando desatacan aquellas enfocadas en la mejora de la infraestructura.

Un ejemplo de ello es el decreto legislativo que modifica la Ley N° 30477, "Ley que regula la ejecución de obras de servicios públicos autorizadas por las municipalidades en las áreas de dominio público", principalmente de telecomunicaciones, aprobado en octubre del 2016. De esta manera se busca simplificar y hacer más expeditos los procedimientos administrativos para la instalación de infraestructura necesaria para la prestación de servicios públicos de telecomunicaciones.

Además, desde el Ministerio de Transportes y Comunicaciones (MTC), el año pasado se implementaron tres grandes concesiones para implementar una Red Regional de Fibra Óptica en Lima, Ica y Amazonas. Estos proyectos se sumaron a las 8 concesiones ya otorgadas para las regiones de Apurímac, Ayacucho, Huancavelica, Lambayeque, Tumbes, Piura, Cajamarca y Cusco. Una vez implementados los 21 proyectos regionales que se tienen planeados, se espera beneficiar a más de 6000 localidades, lo que significará llegar a 3.9 millones de habitantes. De esta manera, estos proyectos permitirán extender 30 000 kilómetros de fibra óptica en el Perú.

Por otro lado, también se ha impulsado el ingreso de más empresas de telecomunicaciones móviles para generar más competitividad entre ellas y mejores opciones para los usuarios. De esta manera, el mercado de la telefonía móvil sigue creciendo en el Perú y ya reporta una penetración a nivel nacional de 80 % del territorio.

Ante este contexto favorable para que las empresas de telecomunicaciones inviertan en el Perú, desde inicios de este año la empresa de telefonía móvil "Dolphin Telecom", en la que usted viene trabajando, decidió iniciar operaciones el 2018. Para esto, viene realizando ya varios estudios del contexto y del mercado que le permitan identificar la situación más específica del sector, así como el comportamiento del cliente peruano.

Para analizar el contexto, a usted le han encargado presentar algunos gráficos y tasas que permitan conocer cómo se han desenvuelto las empresas de telefonía móvil que ya operan en el país. De esta forma, usted ha identificado que debe presentar y calcular lo siguiente:

a) Graficar la serie de tiempo de todas las líneas en servicio (sin diferenciar la modalidad contractual ni la empresa), interpretar el gráfico y extraer las principales ideas de lo encontrado
b) Calcular el crecimiento interanual de todo el sector y extraer las principales ideas de lo encontrado
c) Calcular el crecimiento acumulado de todo el sector y extraer las principales ideas de lo encontrado
d) Calcular la tasa de crecimiento promedio anual de todo el sector y extraer las principales ideas de lo encontrado

Para presentar lo que se propone, usted debe valerse de las fuentes de información oficiales. Así, usted se dirigió a la página web de OSIPTEL, el organismo regulador de las telecomunicaciones (http://www.osiptel.gob.pe/); allí descargó la base de datos que empleará para la resolución de este caso. Al descargar la información, identificó que esta se encontraba presentada en periodos mensuales y en tablas según el año, por lo que deberá ordenar la información para poder analizarla.

Debido a las características de "Dolphin Telecom", la empresa cuenta con infraestructura, personal y estrategias comerciales para competir directamente con "Telefónica Móviles" y "América Móvil", quienes son los principales operadores en el Perú. Por esto es que también debe realizar un análisis que le permita analizar el desempeño por empresa operadora. Así, debe realizar lo siguiente:

e) Graficar la serie de tiempo según empresa operadora, interpretar el gráfico y extraer las principales ideas de lo encontrado
f) Calcular la tasa de crecimiento promedio anual según empresa operadora y extraer las principales ideas de lo encontrado
g) Calcular y graficar la participación porcentual de cada una de las empresas operadoras al último periodo con el que cuenta con información

Una vez terminado el análisis de contexto, le han solicitado, además, dar soporte al equipo de *marketing* con sus conocimientos estadísticos para la investigación de mercados en el estudio que vienen realizando para conocer cómo se ha comportado la portabilidad entre empresas. Para esto, han elaborado una encuesta semiestructurada en la que se busca conocer los principales motivos por los que un cliente decide portar a otra empresa. El equipo cuenta con información desde julio de 2014, ya que desde esa fecha se cambió la norma para permitir que la portabilidad numérica tome tan solo un día, lo que ha generado que los clientes se sientan más animados a cambiarse de empresa operadora basándose, sobre todo, en la calidad de los servicios ofrecidos.

El equipo de *marketing* le ha comentado que desea que la muestra para la aplicación de la encuesta sea representativa para todos los departamentos del Perú y que se tome en cuenta la distribución poblacional tan variada con la que cuenta el territorio peruano. Para esto, le comentan que ya cuentan con la información de las líneas móviles portadas por departamento.

Así, usted ha identificado que debe establecer y calcular lo siguiente:

h) ¿Qué tipo de muestreo estadístico es el más adecuado para los requerimientos del equipo de *marketing*?
i) ¿Cuántos elementos tiene la población que se tomaría en cuenta para este estudio?
j) Considerando un nivel de confianza del 95 % y un error del 5 %, ¿cuánto sería la muestra que debería encuestarse para dicho estudio?
k) ¿Cuántos participantes por departamento deberían considerarse en la muestra? Presente una tabla de doble entrada con dichas cantidades.

Solucionario

Ejercicio I: Muestreo aleatorio simple

Para este ejercicio, emplee el archivo en Excel "SOL_2".

En este caso, nos encontramos con una población finita (menor a 100 mil individuos), por lo que se debe aplicar la fórmula de muestreo aleatorio simple para población finita:

[2-1] Fórmula de muestreo para poblaciones finitas

$$n = \frac{N * p * q * Z\alpha_{/2}{}^2}{(N-1) * e^2 + p * q * Z\alpha_{/2}{}^2}$$

Donde:

N:	3250
p:	0.5
q:	0.5
Z:	1.96
e:	0.04

Como se señala, la distribución se considera con una máxima dispersión ($p = 0.5$). Por otro lado, como no se señala el nivel de confianza, se asume 95 %. De esta manera, se necesita encuestar a 507 mujeres adolescentes como mínimo.

Ejercicio II: Muestreo aleatorio simple

Para este ejercicio, emplee el archivo en Excel "SOL_2".

(Continúa del enunciado del Ejercicio I)

En este caso, nos encontramos con una población finita (menor a 100 mil individuos), por lo que se debe aplicar la misma fórmula del ejercicio I:

Donde:

N: 3250
p: 0.251
q: 0.749
Z: 1.96
e: 0.05

Como ya existe un estudio previo, se toma esa distribución. El 25.1 % reporta que está o ha tenido un embarazo (p); el 74.9 % de las adolescentes no tiene o ha tenido un embarazo (q) y, como no se señala el nivel de confianza, se asume 95 %. De esta manera, se necesita encuestar, por lo menos, a 266 mujeres adolescentes.

Ahora se debe calcular el número de encuestas por provincia. Para ello, primero se calculará la participación (peso relativo) de cada una y luego se multiplicará por el número total de encuestas. Para más detalle sobre la solución de este ejercicio en Excel, tome de referencia la base "SOL_2" (Hoja Ejercicio 2.2) del *drive* compartido en la bibliografía al final de este documento.

En Excel:

- **Frecuencia relativa:**
  Para esto se usará tablas dinámicas.
  Seleccione la BD > Insertar > Tabla dinámica > Elija dónde desea colocar la tabla: Nueva hoja o en la hoja existente (seleccione un rango de celdas que ocupará la tabla) > Aceptar > Arrastre las variables a analizar en cuadrante columna, fila o filtro y determine cómo trabajará con los siguientes valores: cuenta, promedio, etc.
- **Cuotas muestrales:**
  Para calcular el número de encuestas por estrato se usarán las funciones de multiplicar (*) y redondear (REDONDEAR.MAS).

Tabla 2-2: Encuestas por provincia

Provincia	Número de mujeres adolescentes	Frecuencia relativa	Número de encuestas por provincia	Número de encuestas exactas
Maynas	28 592	55.39 %	147.00	148
Alto Amazonas	5548	10.75 %	28.53	29

Loreto	3369	6.53 %	17.32	18
Mariscal Ramón Castilla	3733	7.23 %	19.19	20
Requena	3610	6.99 %	18.56	19
Ucayali	3569	6.92 %	18.34	19
Datem del Marañón	3195	6.19 %	16.42	17
Total	51 616			270

Finalmente, es importante mencionar que este ejercicio tiene como objetivo el cálculo del tamaño de muestra, mas no piden que se realice la selección de la muestra de forma aleatoria en sí. Para esto último, SPSS nos da facilidad para ejecutar este tipo de tareas con los siguientes comandos.

En SPSS:

- **Selección de casos para muestra aleatoria:**
  Datos > Seleccionar casos > Seleccionar: Muestra aleatoria de casos > Ejemplo: Colocar el % casos a seleccionar según el "n" deseado > Continuar > Seleccionar "Copiar casos en nuevo conjunto de datos": Colocar nombre > Aceptar.

  Datos > Seleccionar casos > Muestra aleatoria de casos > Ejemplo: Exactamente > Continuar > Seleccionar "Copiar casos en nuevo conjunto de datos": Colocar nombre > Aceptar.

- **Selección de casos por condicional si:**
  Datos > Seleccionar casos > Selecciona: Si se satisface la condición: Si…> Ingrese la variable por la que desea estratificar y construya la expresión que necesite cumplir > Continuar.

- **Selección de casos por rango:**
  Datos > Seleccionar casos > Seleccione: Basándose en el rango de tiempo o de los casos > Ingrese el rango > Aceptar.

Caso aplicado: Líneas móviles

Para este caso, emplee el archivo en Excel "SOL_2".

a) Graficar la serie de tiempo de todas las líneas en servicio (sin diferenciar la modalidad contractual ni la empresa), interpretar el gráfico y extraer las principales ideas de lo encontrado.

Para realizar esto, se deben sumar todas las líneas y así obtener un total por periodo. Luego se procede a insertar un gráfico de series de tiempo. Luego se debe colocar los títulos y los nombres más adecuados para los ejes.

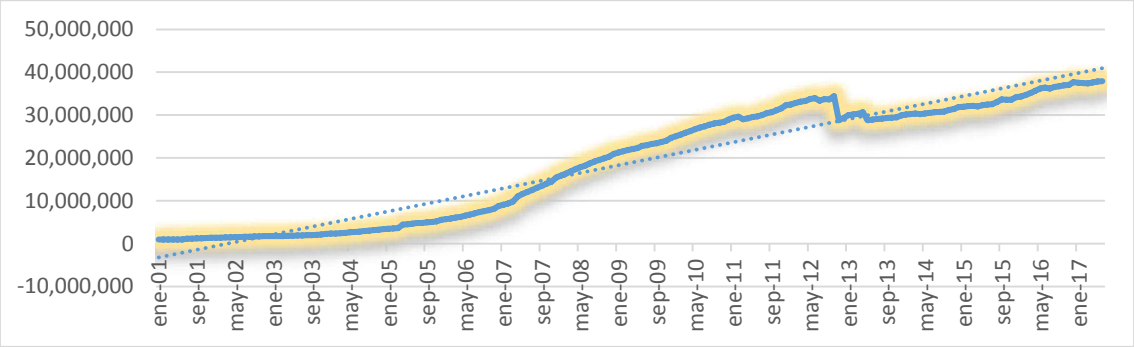
En Excel:

- **Crear gráfico de tendencia:**
  Seleccionar el rango de datos que desea analizar > Insertar > Gráficos > Seleccionar el tipo de gráfico que necesites (gráficos recomendados): Líneas > Aceptar.
- **Modificar gráfico:**
  Seleccionar el gráfico > Herramientas de gráfico > Diseño y Formato.

En SPSS:

- **Crear gráfico de tendencia:**
  Gráficos > Generador de gráficos > Aceptar (cuadro de diálogo) > Seleccionar el tipo de gráfico que desea y arrastre a la sección de vista previa: Línea > Arrastre las variables según correspondan al eje X o Y en la vista previa. También se puede usar alguna variable de filtro.

Figura 2-1: Evolución histórica de las líneas móviles en el Perú



b) Calcular el crecimiento interanual de todo el sector y extraer las principales ideas de lo encontrado.

Para calcular esto, necesitamos los datos anuales. Como el último mes del que tenemos información es junio del 2017, calcularemos el crecimiento interanual en función a todos los meses de junio de los años de los que contamos con información. En el solucionario SOL_2 (Hoja: Caso 2) que aparece en el *drive* compartido en la bibliografía al final de este material, se aprecia con mayor detalle cómo se calculó esta tasa; en este ejercicio, se creó la variable "Crecimiento interanual".

El crecimiento interanual nos permite observar cómo se va dando el desempeño del sector respecto a su rendimiento anterior. Según lo observado, el sector

estuvo en crecimiento durante todos los años de análisis a excepción del 2013, cuando los usuarios dejaron de adquirir líneas de telefonía móvil.

En Excel:

- **Crear variable crecimiento anual:**
  En una columna vacía nombrar la variable en la celda de la primera fila > Use el signo "=" y elija la celda con el primer dato de la variable sobre la cual hará el cálculo, y utilice los signos de operación que necesite: ($Total de líneas 2017$-Total de líneas)/Total de líneas.

En SPSS:

- **Variable calculada crecimiento anual:**
  Transformar > Calcular variable > Nombrar la nueva variable en "variable objetivo" > Ingrese la variable con las que trabajará y utilice los signos de operación que necesite para crear la siguiente sintaxis: (37 944 348 – Total de líneas)/Total de líneas.

c) Calcular el crecimiento acumulado de todo el sector y extraer las principales ideas de lo encontrado.

Para calcular esto necesitamos los datos anuales. Como el último mes del que tenemos información es junio del 2017, calcularemos el crecimiento acumulado en función a todos los meses de junio de los años de los que contamos con información.

El crecimiento acumulado nos permite analizar el desempeño del sector en un periodo de tiempo determinado. Para su cálculo, se debe contar con un dato inicial y un dato final; en función a esto se puede calcular. Según lo encontrado desde el 2001 a la fecha (2017), el sector ha crecido en 3242.9 %, es decir, actualmente existe 32.4 veces más que el 2001.

En el solucionario SOL_2 (Hoja Caso 2) que aparece en el *drive* compartido en la bibliografía al final de este material, se aprecia con mayor detalle cómo se calculó esta tasa; en este ejercicio, se creó la variable "Crecimiento acumulado".

En Excel:

- **Crear variable crecimiento acumulado:**
  En una columna vacía, nombre la variable en la celda de la primera fila > Use el signo "=" y elija la celda con el primer dato de la variable sobre la cual hará el cálculo y utilice los signos de operación que necesite.

En SPSS:

- **Variable calculada crecimiento acumulado:**
  Transformar > Calcular variable > Nombrar la nueva variable en "variable objetivo" > Ingrese la variable con la que trabajará y utilice los signos de operación que necesite.

d) Calcular la tasa de crecimiento promedio anual de todo el sector y extraer las principales ideas de lo encontrado.

Para calcular esto necesitamos los datos anuales. Como el último mes del que tenemos información es junio del 2017, calcularemos la tasa de crecimiento promedio anual en función a todos los meses de junio de los años de los que contamos con información. La tasa de crecimiento promedio anual nos permite encontrar un valor que representa el desempeño de todos los periodos que estamos analizando. Por ello, su cálculo considera el total de periodos de los que contamos con información. Sin embargo, pese a que no contemos con información de todos los periodos para todos los casos, esta tasa nos permite comparar el desempeño entre ellos. Así, para todo el periodo de tiempo con el que contamos con información, podemos indicar el que desempeño del sector, en promedio, ha sido de 24.53 %.

En el solucionario SOL_2 (Hoja Caso 2) que aparece en el *drive* compartido en la bibliografía al final de este material, se aprecia con mayor detalle cómo se calculó esta tasa. En este ejercicio, se creó las variables "diferencia de años" y "crecimiento promedio anual".

En Excel:

- **Crear variable diferencia de años y crecimiento anual promedio:**
  En una columna vacía nombre la variable en la celda de la primera fila > Use el signo "=" y elija la celda con el primer dato de la variable sobre la cual hará el cálculo y utilice los signos de operación que necesite.

En SPSS:

- **Variable calculada diferencia de años y crecimiento anual promedio:**
  Transformar > Calcular variable > Nombrar la nueva variable en "variable objetivo" > Ingrese la variable con las que trabajará y utilice los signos de operación que necesite.

En la siguiente tabla se presentan los resultados de las preguntas b, c y d:

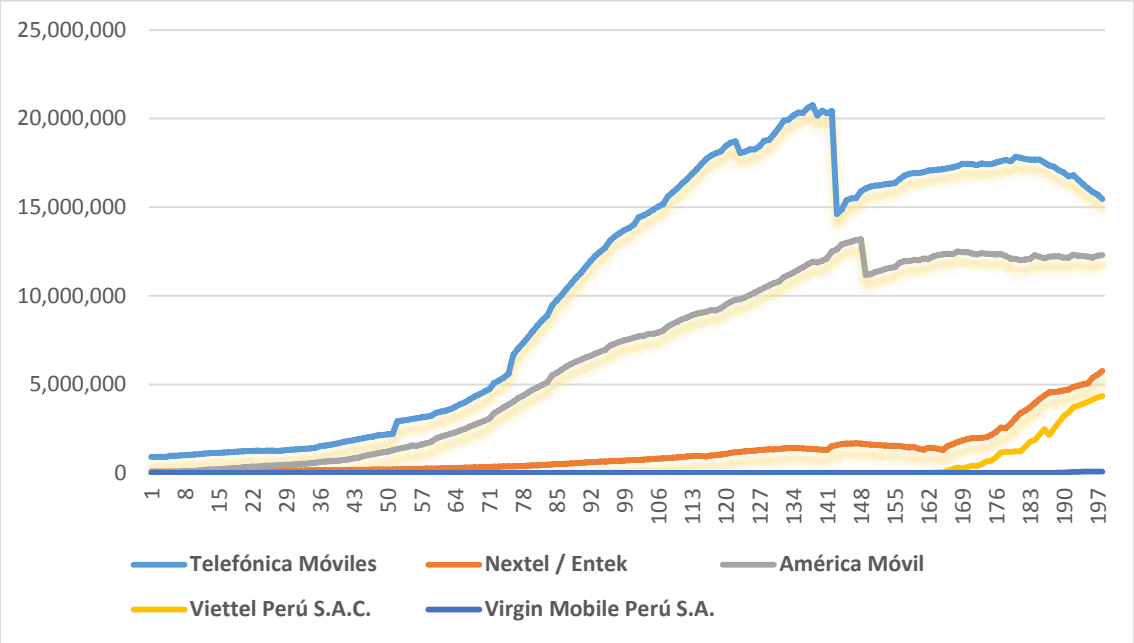
Tabla 2-3: Cálculo de la tasa de crecimiento interanual, tasa de crecimiento acumulado y tasa de crecimiento promedio anual

Mes	Año	Total de líneas	Crecimiento interanual	Crecimiento acumulado	Diferencia de años	Crecimiento promedio anual
Junio	2001	1 135 072		3,242.90 %	16	24.53 %
Junio	2002	1 560 913	37.52 %	2,330.91 %	15	23.70 %
Junio	2003	1 908 966	22.30 %	1,887.69 %	14	23.81 %
Junio	2004	2 731 758	43.10 %	1,289.01 %	13	22.43 %
Junio	2005	4 655 214	70.41 %	715.09 %	12	19.11 %
Junio	2006	6 758 254	45.18 %	461.45 %	11	16.98 %
Junio	2007	12 057 300	78.41 %	214.70 %	10	12.15 %
Junio	2008	18 233 529	51.22 %	108.10 %	9	8.48 %
Junio	2009	22 858 680	25.37 %	66.00 %	8	6.54 %
Junio	2010	27 099 375	18.55 %	40.02 %	7	4.93 %
Junio	2011	29 690 565	9.56 %	27.80 %	6	4.17 %
Junio	2012	34 010 443	14.55 %	11.57 %	5	2.21 %
Junio	2013	28 975 620	-14.80 %	30.95 %	4	6.97 %
Junio	2014	30 558 748	5.46 %	24.17 %	3	7.48 %
Junio	2015	32 469 361	6.25 %	16.86 %	2	8.10 %
Junio	2016	36 478 887	12.35 %	4.02 %	1	4.02 %
Junio	2017	37 944 348	4.02 %	0.00 %	0	

e) Graficar la serie de tiempo según empresa operadora, interpretar el gráfico y extraer las principales ideas de lo encontrado.

Para elaborar este gráfico es necesario calcular el total de líneas que tenía cada una de las empresas operadoras en todos los periodos de tiempo que queremos graficar.

Figura 2-2: Evolución de las líneas por empresa



La figura anterior fue creada en Excel, pero esta también se podría lograr en SPSS. A continuación se presentan los comandos de apoyo.

En Excel:

- **Crear gráfico de tendencia o evolución:**
  Seleccionar el rango de datos que desea analizar > Insertar > Gráficos > Seleccionar el tipo de gráfico que necesite (gráficos recomendados): Líneas > Aceptar.
- **Modificar gráfico:**
  Seleccionar el gráfico > Herramientas de gráfico > Diseño y Formato.

En SPSS:

- **Crear gráfico de tendencia o evolución:**
  Gráficos > Generador de gráficos > Aceptar (cuadro de diálogo) > Seleccionar el tipo de gráfico que desea y arrastre a la sección de vista previa: Línea > Arrastre las variables según correspondan al eje X o Y en la vista previa. También se puede usar alguna variable de filtro.

f) Calcular la tasa de crecimiento promedio anual según empresa operadora y extraer las principales ideas de lo encontrado.

Para calcular estas tasas se crearon las variables crecimiento acumulado y diferencia de años por cada operadora. De esto se tiene la siguiente tabla; a partir de esta, a nivel general se puede observar que quien ha tenido un mejor

desempeño hasta la fecha, respecto a los años que lleva operando (considerando información desde el 2001), es Viettel, la cual solo con 2 años de operaciones presenta una tasa de crecimiento promedio de 156.17 %. Para el caso de Viettel y de Virgin Mobile habría que estar a la expectativa de cómo se seguirán comportando en periodos posteriores.

En el solucionario SOL_2 (Hoja Caso 2) que aparece en el *drive* compartido en la bibliografía al final de este material, se tiene mayor detalle de cómo se calculó esta tasa.

En Excel:

- **Crear variable crecimiento promedio anual según operadora:**
  En una columna vacía nombre la variable en la celda de la primera fila > Use el signo "=" y elija la celda con el primer dato de la variable sobre la cual hará el cálculo y utilice los signos de operación que necesite.

En SPSS:

- **Variable calculada crecimiento promedio anual según operadora:**
  Transformar > Calcular variable > Nombrar la nueva variable en "variable objetivo" > Ingrese la variable con la que trabajará y utilice los signos de operación que necesite.

Tabla 2-4: Tasa de crecimiento promedio anual

Mes	Año	TM	CA	CPA	N/E	CA	CPA	AM	CA	CPA	VP	CA	CPA	VMP	CA	CPA
Junio	2001	973 065	1488.80 %	18.87 %	90 481	6254.93 %	29.63 %	71 526	17 087.45 %	37.94 %	0			0		
Junio	2002	1 180 549	1209.56 %	18.71 %	120 843	4658.25 %	29.37 %	259 521	4636.99 %	29.33 %	0			0		
Junio	2003	1 303 795	1085.77 %	19.32 %	140 065	4005.24 %	30.39 %	465 106	2543.16 %	26.35 %	0			0		
Junio	2004	1 795 468	761.06 %	18.01 %	164 579	3393.77 %	31.44 %	771 711	1493.02 %	23.73 %	0			0		
Junio	2005	2 987 235	417.54 %	14.68 %	212 729	2602.97 %	31.62 %	1 455 250	744.77 %	19.46 %	0			0		
Junio	2006	3 979 139	288.53 %	13.13 %	296 494	1839.33 %	30.94 %	2 482 621	395.18 %	15.65 %	0			0		
Junio	2007	7 314 938	111.35 %	7.77 %	392 929	1363.37 %	30.78 %	4 349 433	182.65 %	10.95 %	0			0		
Junio	2008	11 280 018	37.06 %	3.56 %	561 575	923.91 %	29.49 %	6 391 936	92.33 %	7.54 %	0			0		
Junio	2009	14 425 764	7.17 %	0.87 %	736 897	680.30 %	29.28 %	7 696 019	59.74 %	6.03 %	0			0		
Junio	2010	17 139 925	-9.80 %	-1.46 %	965 050	495.82 %	29.04 %	8 994 400	36.68 %	4.56 %	0			0		
Junio	2011	18 255 654	-15.31 %	-2.73 %	1 263 880	354.95 %	28.72 %	10 171 031	20.87 %	3.21 %	0			0		
Junio	2012	20 753 572	-25.51 %	-5.72 %	1 347 972	326.57 %	33.66 %	11 908 899	3.23 %	0.64 %	0			0		
Junio	2013	16 160 018	-4.33 %	-1.10 %	1 603 747	258.54 %	37.60 %	11 211 855	9.65 %	2.33 %	0			0		
Junio	2014	17 063 821	-9.40 %	-3.24 %	1 409 208	308.03 %	59.80 %	12 085 719	1.72 %	0.57 %	0			0		
Junio	2015	17 427 019	-11.29 %	-5.81 %	2 009 274	186.17 %	69.17 %	12 369 926	-0.62 %	-0.31 %	663 142	556.22 %	156.17 %	0		
Junio	2016	17 518 388	-11.75 %	-11.75 %	4 359 827	31.89 %	31.89 %	12 111 379	1.50 %	1.50 %	2 489 293	74.82 %	74.82 %	0		
Junio	2017	15 460 036	0.00 %		5 750 007	0.00 %		12 293 496	0.00 %		4 351 670	0.00 %		89 139	0.00 %	

Donde:

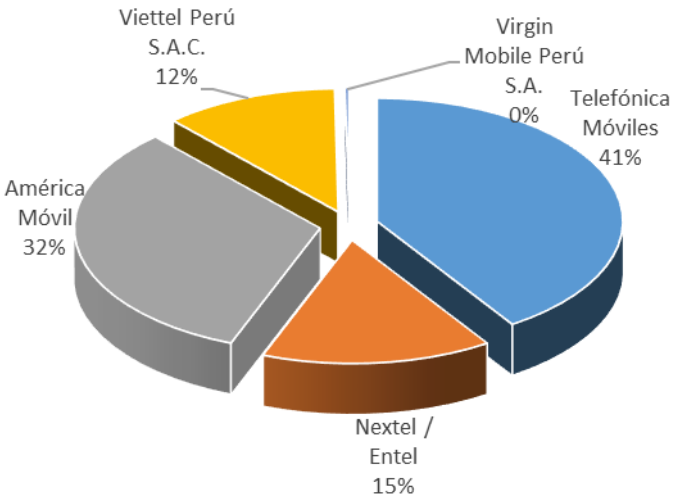
CA: Crecimiento acumulado
CPA: Crecimiento promedio anual
TM: Telefónica Móviles
N/E: Nextel/Entel
AM: América Móviles
VP: Viettel Perú
VMP: Virgin Mobile Perú

g) Calcular y graficar la participación porcentual de cada una de las empresas operadoras al último periodo con el que cuenta con información.

A junio del 2017, es Telefónica Móviles la que tiene mayor participación con un 41 %. Sin embargo, como se ha visto en el gráfico de series de tiempos, es la empresa que está perdiendo más clientes a la fecha.

Figura 2-3: Participación por empresa de telefonía móvil

Empresa	Total de líneas
Telefónica Móviles	15 460 036
Nextel / Entel	5 750 007
América Móvil	12 293 496
Viettel Perú S.A.C.	4 351 670
Virgin Mobile Perú S.A.	89 139



La figura anterior fue creada en Excel, pero esta también se podría lograr en SPSS. A continuación se presentan los comandos de apoyo para ambos *software*.

En Excel:

- **Crear gráfico para partición de mercado:**
  Seleccionar el rango de datos que desea analizar > Insertar > Gráficos > Seleccionar el tipo de gráfico: Gráfico circular 3D > Aceptar.

En SPSS:

- **Crear gráfico para partición de mercado:**
  Gráficos > Generador de gráficos > Aceptar (cuadro de diálogo) > Seleccionar el tipo de gráfico: Circular > Arrastre las variables según correspondan al eje X o Y en la vista previa.

h) ¿Qué tipo de muestreo estadístico es el más adecuado para los requerimientos del equipo de *marketing*?

Debido a que el equipo de *marketing* quiere una muestra que sea representativa de todos los departamentos, y considera a los departamentos como grupos homogéneos entre sí, estos departamentos pueden ser considerados como estratos.

De esta manera, se procedería a calcular la cantidad de la muestra y a repartir dicha cantidad proporcionalmente a la cantidad de elementos que tiene un estrato. Así, del total de la muestra, 75.17 % deberá ser de Lima y así sucesivamente por departamento.

i) ¿Cuántos elementos tiene la población que se tomaría en cuenta para este estudio?

Debido a que se quiere recoger la opinión de aquellos que cambiaron de empresa operadora desde que cambió la ley, ya que se presume que solo la calidad del servicio los motivó a hacerlo, la población sería de 4 227 369; de ellos es que se podría elegir la muestra. Esta cantidad es considerada una población infinita para efectos del uso de las fórmulas de cálculo de muestra.

j) Considerando un nivel de confianza del 95 % y un error del 5 %, ¿cuánto sería la muestra que debería encuestarse para dicho estudio?

[2-2] Fórmula de muestreo para poblaciones infinitas

$$n = \frac{Z^2 p(1-p)}{e^2}$$

Donde:

N:	4 274 808
p:	0.5
$1-p$:	0.5
Z:	1.96
e:	0.05

Aplicando la fórmula de muestreo aleatorio simple para poblaciones infinitas, el resultado es de 384 clientes para este estudio.

k) ¿Cuántos participantes por departamento debería considerarse en la muestra? Presentar una tabla de doble entrada con dichas cantidades.

Para crear la siguiente tabla es importante crear la variable distribución muestral (%) y, sobre esta, calcular la muestra respetando la representatividad original de las líneas portadas. Para ver más detalle sobre la solución de este ejercicio, revise la base SOL_2 que aparece en el *drive* compartido en la bibliografía al final de este material. También recuerde que para calcular variables en SPSS o Excel puede apoyarse en el uso de los siguientes comandos.

En Excel:

- **Crear variable distribución muestral:**
  En una columna vacía nombre la variable en la celda de la primera fila > Use el signo "=" y elija la celda con el primer dato de la variable sobre la cual hará el cálculo y utilice los signos de operación que necesite.

En SPSS:

- **Variable calculada distribución muestral:**
  Transformar > Calcular variable > Nombrar la nueva variable en "variable objetivo" > Ingrese la variable con la que trabajará y utilice los signos de operación que necesite.

Tabla 2-5: Muestra por departamento

Departamento	Muestra exacta
Amazonas	1
Ancash	5
Arequipa	23
Ayacucho	3
Apurímac	3
Cajamarca	3
Cusco	5
Huancavelica	2
Huánuco	3
Ica	4

Junín	4
La Libertad	17
Lambayeque	12
Lima y Callao	286
Loreto	1
Madre de Dios	1
Moquegua	1
Pasco	1
Piura	10
Puno	2
San Martín	1
Tacna	6
Tumbes	2
Ucayali	2
TOTAL	**398**

Lecturas recomendadas

- Anderson, D., Sweeney, D., y Williams, T. (2008). *Estadística para administración y economía* (10.ª ed., cap. 22). Cengage Learning Editores.
- Levin, R., y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed. revisada, cap. 6). México: Pearson Educación de México; Prentice Hall.

3. Estadística descriptiva: frecuencias y gráficos

El tercer capítulo tiene como objetivo que el alumno utilice estadística descriptiva para resumir el análisis de datos. Con la ayuda de *software* de ingreso matricial, el alumno podrá generar gráficos y tablas de frecuencia para presentar la información de manera efectiva. Esta primera parte busca que a través del uso de medidas de tendencia central y de variabilidad se logre mejorar la toma de decisiones en el contexto de la gestión de organizaciones.

Palabras clave

- Medidas de tendencia central
- Medidas de variabilidad
- Distribuciones de frecuencia

- Gráficos para estadística descriptiva

Ejercicio I: Uso de tablas de frecuencia y gráficos

Para este ejercicio, emplee el archivo en Excel "BD_3".

(Continúa del enunciado del Ejercicio I del Capítulo 2)

Con el fin de determinar las provincias a priorizar, el equipo de consultores al que usted pertenece ha revisado investigaciones preliminares sobre la problemática en la región; además, ha recopilado datos del INEI sobre el número de niños nacidos vivos de madres adolescentes por distritos[6]. Para tener un mapeo inicial de la problemática en la región, se le pide que analice lo siguiente:

a. Determinar la "tasa de nacidos de mujeres adolescentes" (TNMA) de cada distrito.
b. Además, se desea saber cuántos distritos tienen una TNMA menor al 20 % y cuántos superiores al 60 %. Al respecto, las autoridades de la región le han solicitado presentar los resultados de manera gráfica.

Ejercicio II: Tasas

Para este ejercicio, emplee el archivo en Excel "BD_3".

(Continúa del enunciado del Ejercicio I del Capítulo 2)

Debido a que también se dispone de información a nivel de cada provincia, se le pide calcular las TNMA por provincia. Una vez calculadas las tasas, debe mostrar un gráfico de dichas tasas y determinar cuáles son las 4 provincias con mayor prioridad para la implementación de los programas sociales en la región.

Caso aplicado: La percepción de docentes sobre la LRM

Para este caso, emplee el archivo en Excel "BD_3".

En el marco de la huelga docente protagonizada en el Perú durante el año 2017, se le pide como asesor del Ministerio de Educación (MINEDU) que realice un informe sobre la percepción de los docentes hacia la Ley de Reforma Magisterial (LRM) y dé otra información relevante que fortalezca los argumentos del MINEDU en la negociación.

[6] INEI (2014). Maternidad en la adolescencia 2012. Lima. Consulta del 23 de agosto del 2017. Recuperado de https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1184/libro.pdf

En ese sentido, usted dispone de información proveniente del MINEDU y de los resultados de la Encuesta Nacional Docente (ENDO) del año 2014.

La aprobación de la LRM es una de las variables a tomar en cuenta al momento de elaborar la comunicación gubernamental del Ejecutivo hacia el SUTEP central, como también a las sedes descentralizadas. Se sabe que la tasa de aprobación varía considerablemente según regiones. Una vez identificadas las regiones con mayor aprobación y desaprobación, es necesario analizar los principales grupos etarios del profesorado.

a. Utilizando la información de la pregunta 64 del cuestionario (ver la base BD_3), determine la región con mayor tasa de percepción negativa a la LRM. Explique los supuestos tomados.
b. Elabore un gráfico circular para los porcentajes de cada respuesta a la pregunta 64 en la región elegida.
c. Construya una tabla de frecuencias (absolutas y relativas, simples y acumuladas) de la edad de los docentes con este tipo de percepción en la región elegida. Construya el número de intervalos utilizando la fórmula de Sturges. Comente.
d. Grafique el histograma con la frecuencia relativa simple. Utilice escala vertical con saltos de 5 %.

Solucionario

Ejercicio I: Uso de tablas de frecuencia y gráficos

Para este ejercicio, emplee el archivo en Excel "SOL_3".

(Continúa del enunciado del Ejercicio I de la Unidad II)

a. Determinar la "tasa de nacidos de mujeres adolescentes" (TNMA) de cada distrito.

Para calcular la tasa, se necesita una división simple en donde el numerador es el total de nacidos de madres adolescentes y el denominador es el total de nacidos dentro de ese distrito. Para ello, se crea la siguiente tabla, la cual tiene la finalidad de agrupar los datos y observar cuál es el número de distritos que tienen una tasa entre 0 y 20 %, y entre 80 y 100 %.

Tabla 3-1: Tasa de nacidos de mujeres adolescentes (TNMA)

TNMA	Frecuencia
0 - 0.2	17
0.2 - 0.4	11
0.4 - 0.6	6
0.6 - 0.8	1
0.8 - 1	2
Total general	37

Una vez creada la variable "Tasa distrital de nacidos de madres adolescentes", la tabla de frecuencias anterior también se puede construir en SPPS mediante los siguientes comandos.

En SPSS:

- **Crear tablas de frecuencia:**
  Analizar > Estadísticos descriptivos > Frecuencias > Inserte la variable a analizar: TNMA.
- **Agrupar datos:**
  Ventana editor de datos > Transformar > Agrupación visual.

En Excel:

- **Crear tablas dinámicas:**
  Selecciona la BD > Insertar > Tabla dinámica > Elija dónde desea colocar la tabla > Aceptar > Arrastre la variable TNMA al cuadrante fila y elija la opción "cuenta" para el cálculo de valores.
- **Agrupar datos en una tabla dinámica:**
  Seleccione un dato de la tabla dinámica > Clic derecho > Agrupar > Llenar los datos "comenzar en", "terminar en" y "por" según corresponda.

b. Además, se desea saber cuántos distritos tienen una TNMA menor al 20 % y cuántos superiores al 60 %. Al respecto, las autoridades de la región le han solicitado presentar los resultados de manera gráfica.

Una vez que se tiene esa columna, es posible, a través de una tabla dinámica, agrupar los datos y observar cuál es el número de distritos que tienen una tasa entre 0 y 20 %, y entre 80 y 100 %.

Tabla 3-2: Tabla de frecuencias para la TNMA

TNMA	X	F	F	h	H
0 - 0.2	0.1	17	17	45.9 %	45.9 %
0.2 - 0.4	0.3	11	28	29.7 %	75.7 %
0.4 - 0.6	0.5	6	34	16.2 %	91.9 %
0.6 - 0.8	0.7	1	35	2.7 %	94.6 %
0.8 - 1	0.9	2	37	5.4 %	100.0 %

En Excel, a través de las opciones de "Insertar" y "Gráficos recomendados", es posible generar un gráfico de polígono. Asimismo, también se puede generar un gráfico de ojiva con la frecuencia acumulada. Visualmente, la primera permite comparar qué intervalos agrupan la mayor cantidad de distritos. Con el de ojiva, se puede notar fácilmente cómo a medida que aumenta la tasa, son menos los distritos en dichos grupos. Sobre estos mismos gráficos, ambos se pueden construir en SPSS con los siguientes comandos.

En SPSS:

- **Crear gráfico de frecuencias en SPSS:**
  Analizar > Estadísticos descriptivos > Frecuencias > Ingresar variables según correspondan > Gráficos > Marcar histogramas > *Check* en mostrar curva normal.

Figura 3-1: Niños nacidos de madres adolescentes

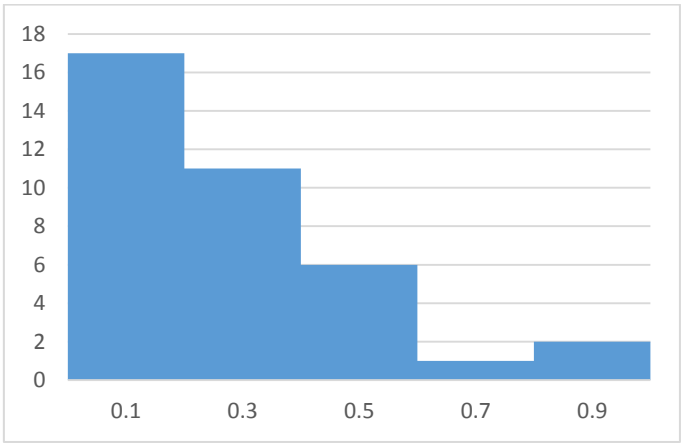
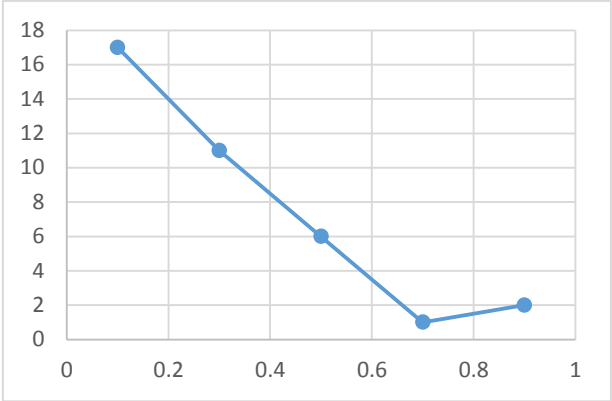


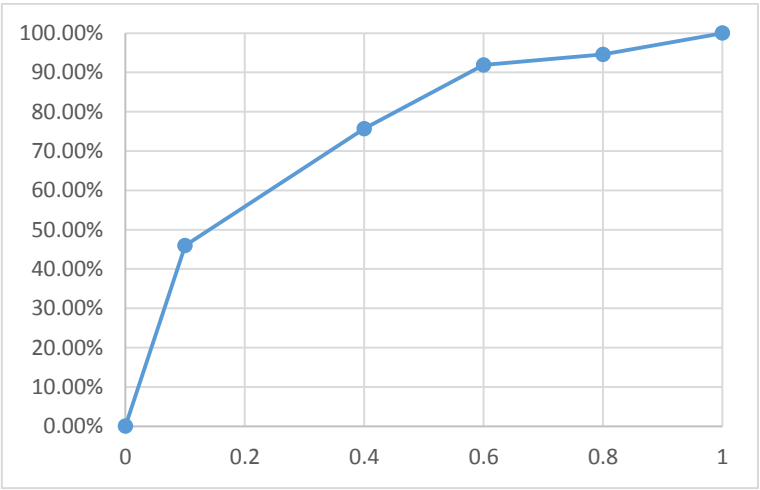
Figura 3-2: Tasa distrital de nacidos de madres adolescentes



x	F	H
0	0	0.00 %
0.1	17	45.95 %
0.4	28	75.68 %
0.6	34	91.89 %
0.8	35	94.59 %
1	37	100.00 %

Recuerde que, para la ojiva, debe ingresar una marca adicional cero (0) para que la ojiva tenga un inicio en el origen y cambiar las marcas de clase por el máximo de cada intervalo, de manera que el último punto de la ojiva sea 1 o 100 %.

Figura 3-3: Tasa distrital de nacidos de madres adolescentes (%)



Son 17 distritos de 37 los que tienen una tasa menor a 20 % de nacidos de mujeres adolescentes; mientras que son solo el 8.1 % o 3 distritos los que tienen una tasa mayor a 60 %.

Ejercicio II: Tasas

Para este ejercicio, emplee el archivo en Excel "SOL_3".

(Continúa del enunciado del Ejercicio I del Capítulo 2)

La tasa es un coeficiente que indica la proporción que representa un evento (fenómeno) del total. En ese sentido, para calcular la tasa de una provincia, primero se debe sumar todos los nacimientos de madres solteras (evento) y dividirlo entre la suma de los nacimientos totales de cada distrito (totales), todo agrupado por provincias.

[3-1] Fórmula de tasas

$$Tasa\ provincial(nac.madr.adoles) = \frac{\sum_{Distrito=1}^{5} Nac.madres.adoles}{\sum_{Distrito=1}^{5} Total\ de\ nacidos}$$

Con el fin de hacer esta operación más fácil, en Excel se usará tablas dinámicas para (a) calcular el total de nacidos de madres adolescentes por provincia, (b) el total de nacidos en una provincia y (c) dividir estas dos cantidades para hallar la tasa de nacimientos de madres adolescentes a nivel provincial; para ver más detalle de la solución de este ejercicio, ver la base SOL 3 (Hoja Ejercicio 3.2). Asimismo, es importante mencionar que la misma tabla, que aparece a continuación, también puede ser generada en SPSS mediante los siguientes comandos.

En SPSS:

- **Crear tablas cruzadas:**
  Analizar > Estadísticos descriptivos > Tablas cruzadas > Inserte las variables a analizar (fila y columna) > Aceptar.

En Excel:

- **Crear tablas dinámicas:**
  Seleccione la BD > Insertar > Tabla dinámica > Elija dónde desea colocar la tabla: Nueva hoja o en la hoja existente (selecciona un rango de celdas que ocupará la tabla) > Aceptar > Arrastre las variables a analizar en cuadrante columna (Sumatoria de valor), fila (Provincia) y determine cómo trabajará con los valores: suma de totales.

Tabla 3-3: Nacidos por provincia

Provincias	Total de nacidos de madres adolescentes	Total de nacidos del distrito
Alto Amazonas	203	5104
Datem del Marañón	4	845
Loreto	2000	5887
Mariscal Ramón Castilla	1677	4626
Maynas	5067	21 574
Requena	1902	5024
Ucayali	1307	3729
Total general	**12 160**	**46 789**

Una vez calculado el total de nacidos de madres adolescentes por provincia y el total de naciones por provincia, se procede a calcular la tasa de nacidos de madres adolescentes; este procedimiento se hizo en Excel con los siguientes comandos.

En Excel:

- **Crear variable TNMA por provincia:**
  En una columna vacía, nombrar la variable en la celda de la primera fila > Use el signo "=" y elija la celda con el primer dato de la variable sobre la cual hará el cálculo y utilice los signos de operación que necesite: = Nacidos de madres adolescentes/Nacidos total (formato de celda: porcentaje)

Tabla 3-4: TNMA por provincia

Provincia	Nacidos de madres adolescentes	Nacidos total	Tasa de nacidos de madres adolescentes
Requena	1902	5024	37.86 %
Mariscal Ramón Castilla	1677	4626	36.25 %
Ucayali	1307	3729	35.05 %
Loreto	2000	5887	33.97 %
Maynas	5067	21 574	23.49 %
Alto Amazonas	203	5104	3.98 %
Datem del Marañón	4	845	0.47 %

Se sabe que la prioridad se establece a partir de la comparación de los pesos relativos (tasas, porcentajes, proporciones) y no de los pesos absolutos. Por ejemplo, de la tabla anterior se tiene que Maynas presenta el mayor número de nacidos de madres adolescentes. Sin embargo, ese número solo representa el 23.5 % de casos de esa provincia. La situación es más alarmante en el caso de Requena, Mariscal Ramón Castilla, Ucayali y Loreto; en ese orden.

Caso aplicado: La percepción de docentes sobre la LRM

Para este caso, emplee el archivo en Excel "SOL_3".

a. Utilizando la información de la pregunta 64 del cuestionario, determine la región con mayor tasa de percepción negativa a la LRM (explique los supuestos tomados).

Se consideró la respuesta "Lo perjudica" como indicador de percepción negativa a la LRM.

Usando Excel, se creó una tabla dinámica en donde se filtró el código 2 "Lo perjudica" de la respuesta a la pregunta 64 y se colocó la variable "Región" como etiqueta de fila. Como se requiere conocer la frecuencia absoluta como valores, se utilizó cuenta. Así se obtuvo la siguiente tabla; para ver más detalle, ver la base SOL_3 (Hoja Caso 3).

Tabla 3-5: Frecuencia de respuestas a la p64 de la ENDO 2014 por región

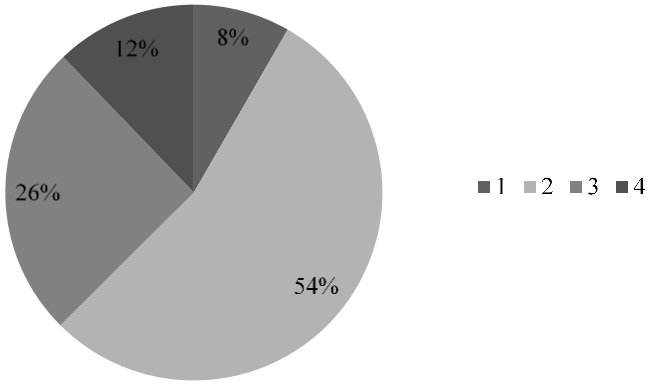
Región	Lo perjudica	Total	Porcentaje
1	175	354	49.4 %
2	119	301	39.5 %
3	157	335	46.9 %
4	93	183	50.8 %
5	157	333	47.1 %
6	196	339	57.8 %
7	61	154	39.6 %
8	112	306	36.6 %
9	173	349	49.6 %
10	115	324	35.5 %
11	88	223	39.5 %
12	121	274	44.2 %

Región	Lo perjudica	Total	Porcentaje
13	124	269	46.1 %
14	83	225	36.9 %
15	113	259	43.6 %
16	125	257	48.6 %
17	130	347	37.5 %
18	124	287	43.2 %
19	139	266	52.3 %
20	133	304	43.8 %
21	103	268	38.4 %
22	131	313	41.9 %
23	139	345	40.3 %
24	114	248	46.0 %
25	120	279	43.0 %
26	116	323	35.9 %

La región de Cajamarca tiene el mayor número de docentes con percepción negativa y también el mayor porcentaje sobre el total de docentes de la región: 57.8 %.

b. Elabore un gráfico circular para los porcentajes de cada respuesta a la pregunta 64 en la región elegida.

Figura 3-4: Percepción de los docentes de la región Cajamarca sobre la LRM



Solo un 8 % considera que la LRM lo beneficia, mientras que un 12 % desconoce los contenidos de la LRM. Para la construcción de gráfico, apóyese en los siguientes comandos en Excel y SPSS.

En Excel:

- **Crear gráfico circular:**
  Seleccionar el rango de datos que desea analizar > Insertar > Gráficos > Seleccionar el tipo de gráfico: Circular > Aceptar

En SPSS:

- **Crear gráfico circular:**
  Gráficos > Generador de gráficos > Aceptar (cuadro de diálogo) > Seleccionar el tipo de gráfico: Circular > Arrastre las variables según correspondan al eje X o Y en la vista previa.

c. Construya una tabla de frecuencias (absolutas y relativas, simples y acumuladas) de la edad de los docentes con este tipo de percepción en la región elegida. Construya el número de intervalos utilizando la fórmula de Sturges. Comente.

Recordando la fórmula de Sturges:

[3-2] Fórmula de Sturges

$$K = 1 + 3.3logN$$

Donde:

K: Número de intervalos
log: Logaritmo decimal
N: Total de frecuencias

La frecuencia que necesitamos es el número de docentes que considera que la LRM los perjudica.

$$K = 1 + 3.3\log(196)$$
$$K = 8.56$$

Para calcular el ancho de cada intervalo, utilizamos la fórmula:

[3-3] Fórmula de ancho de intervalo

$$\frac{Valor\ máximo - Valor\ mínimo}{Número\ de\ intervalos}$$

$$\frac{68 - 25}{8.56} = 5.02$$

Con redondeo simple, obtenemos 9 intervalos con un ancho de 5. Utilizando Excel y la función agrupar de la tabla dinámica, se obtiene la siguiente tabla; para ver más detalle de la solución, ver el archivo SOL_3 (Hoja Caso 3).

Tabla 3-6: Tabla de frecuencias según rango de edad

Rango de edad	Absoluta	Relativa	Abs. acumulada	Relativa acumulada
25-29	4	2.04 %	4	2.04 %
30-34	7	3.57 %	11	5.61 %
35-39	20	10.20 %	31	15.82 %
40-44	56	28.57 %	87	44.39 %
45-49	48	24.49 %	135	68.88 %
50-54	37	18.88 %	172	87.76 %
55-59	20	10.20 %	192	97.96 %
60-64	2	1.02 %	194	98.98 %
65-69	2	1.02 %	196	100.00 %
Total general	**196**	**100.00 %**		

En Excel:

- **Crear tablas dinámicas:**
  Selecciona la BD > Insertar > Tabla dinámica > Elija dónde desea colocar la tabla: Nueva hoja o en la hoja existente (seleccione un rango de celdas que ocupará la tabla) > Aceptar > Arrastre las variables a analizar en cuadrante columna (Sumatoria de valores), fila (Edad) o filtro (P64 y región) y determine cómo trabajará con los valores "cuentas".

En SPSS:

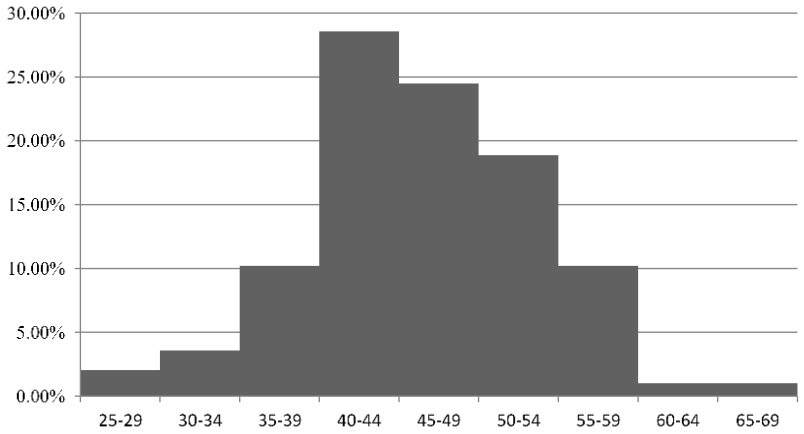
- **Crear tablas de frecuencia:**
  Analizar > Estadísticos descriptivos > Frecuencias > Inserte las variables a analizar.

- **Crear tablas cruzadas:**
  Analizar > Estadísticos descriptivos > Tablas cruzadas > Inserte las variables a analizar (fila y columna) > Aceptar.

De acuerdo con la tabla, se tiene que el mayor porcentaje de docentes se encuentra en el rango de edad de 40 a 44 años (28.57 %). Sumando esta cantidad con el siguiente rango de edad con mayor frecuencia relativa (24.49 %), se puede concluir que el 53.06 % de docentes en la región Cajamarca con una percepción negativa hacia la LRM se encuentra entre los 40 a 49 años de edad.

d. Grafique el histograma con la frecuencia relativa simple. Utilice escala vertical con saltos de 5 %.

Figura 3-5: Histograma de frecuencias relativas. Rango de edad de docentes con percepción negativa hacia la LRM en la región de Cajamarca (porcentaje)



En Excel:

- **Crear histograma:**
  Seleccionar el rango de datos que desea analizar > Insertar > Gráficos > Seleccionar el tipo de gráfico: Columna agrupada > Aceptar

En SPSS:

- **Crear histograma:**
  Gráficos > Generador de gráficos > Aceptar (cuadro de diálogo) > Seleccionar el tipo de gráfico: Barras agrupadas > Arrastre las variables según correspondan al eje X o Y en la vista previa.

Lecturas recomendadas

- Anderson, D., Sweeney, D., y Williams, T. (2008). *Estadística para administración y economía* (10.ª ed., cap. 2). Cengage Learning Editores.
- Levin, R., y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed. revisada, cap. 2). México: Pearson Educación de México; Prentice Hall.

4. Estadística descriptiva: Medidas de tendencia central

Es importante reconocer los usos aplicativos de la estadística descriptiva dentro de la investigación cuantitativa, pues no solo facilita la lectura de resultados numéricos de forma visual mediante el uso de gráficos. El objetivo de este capítulo es comprender la interpretación de las medidas de tendencia central y de dispersión o variabilidad.

Palabras clave

- Medidas de tendencia central
- Estadísticos resúmenes
- Medidas de dispersión

Ejercicio I: Medidas de tendencia central

A continuación, se presenta un cuadro con ciertas características de 10 participantes de un *focus group* realizado para el lanzamiento de un nuevo producto al mercado por parte de una compañía de confección de calzado.

Determinar las medidas de tendencia central que expliquen mejor a las variables del cuadro. Justifique su elección.

Tabla 4-1: Determine las medidas de tendencia central

Variable	Datos	Medida
Talla de zapatos	35, $35_{1/2}$, 36, 36, 36, 36, 36, 37, $37_{1/2}$, 38, 39	
Ingreso mensual (miles de S/)	2.5, 3, 3.5, 3.8, 4.2, 4.4, 5, 6, 11.2, 12	
Edad	22, 24, 25, 26, 26, 26, 28, 30, 31, 32	
Color de zapatos preferido	Azul, Azul, Blanco, Blanco, Blanco, Blanco, Negro, Negro, Rojo, Rojo	
Número de compras de calzado en un año	1, 2, 2, 3, 4, 4, 5, 5, 6, 7	

Ejercicio II: Medidas de dispersión

Para este ejercicio, emplee el archivo en Excel "BD_4".

Una empresa distribuidora de bebidas gaseosas se encuentra realizando un estudio de las ventas que ha venido realizando en la temporada de verano de este año en la ciudad de Lima y sus balnearios. Para este fin, le ha proporcionado a usted la información de las ventas de dos marcas de gaseosas ("Coca Soda" y "Popsy") que esta empresa distribuye, en el periodo que va desde el 01 de diciembre del año 2017 hasta el 31 de enero del año 2018.

Con la información proporcionada se le pide calcular las medidas de tendencia central para las ventas y el ingreso en Lima y sus balnearios, y las medidas de dispersión para las ventas y el ingreso en los meses en los que se desarrolló este estudio.

Caso aplicado: Al rescate de Arkham

Para este caso, emplee el archivo en Excel "BD_4".

Arkham Technology S.A. es una empresa del sector de tecnología que inició sus operaciones en el año 1995. Dicha empresa comenzó como una pequeña empresa que distribuía *hardware* y *software* a personas naturales y en poco tiempo empezó a crecer. Actualmente, se ha constituido como una mediana empresa que factura ventas mayores de S/7 000 000.00 y ha expandido su cartera de clientes clasificándolos en tres tipos consumidores: individuales, microempresas y pequeñas empresas.

En los planes de Arkham se encuentra no detener el crecimiento que han venido llevando y el directorio de esta empresa cree que esto solo será posible si es que aseguran la calidad de sus procesos de producción y entrega de productos. A raíz de esto, se contrató a una consultora para evaluar los procesos de la empresa y, posteriormente, se aprobó la constitución de un departamento de calidad.

La comisión encargada de poner en marcha esta política en la empresa determinó que serán necesarias 3 personas para iniciar el trabajo y usted ha sido contratado para trabajar en este equipo a cargo de Bruno Hernández, el nuevo director de Calidad.

El analista de información del área, Ricardo Mendoza, reporta que aproximadamente el 10 % de los pedidos presenta pérdidas monetarias para la empresa por concepto de fallas del producto o en la entrega.

Hernández, alarmado por la situación, decide proporcionarle a usted una base de datos que contiene los datos de las órdenes de compra de los productos en los que se han experimentado pérdidas a partir del año 2014. Con esta información se le solicita lo siguiente:

a. Calcular el monto de pérdida por cada una de las órdenes de compra de la compañía.
b. Calcular los estadísticos resumen correspondientes para esta variable. Interprete los resultados.

Al proporcionarle los datos al director de Calidad, este intuye que algo no va bien con estos resultados. Para complementar la información, le solicita lo siguiente:

c. Calcule la cantidad de datos atípicos para la variable "Pérdida".
d. De existir datos atípicos, ¿cuál estadístico debe emplear para resumir mejor los datos? ¿Por qué no escoger los otros?

En vista de los resultados obtenidos, el analista de información le solicita lo siguiente:

e. Discriminar los datos atípicos (*outliers*, extremos) y todos los anteriores al año 2017 de la base de datos.
f. Calcular las medidas de distribución central y las medidas de dispersión de los nuevos datos obtenidos. Interprete los resultados.
g. Calcular el coeficiente de variación e interpretar los resultados obtenidos.

Para finalizar, se le solicita realizar algunos cruces de información para determinar posibles causales de pérdida en las órdenes de compra. Para ello, usted debe:

h. Crear un informe en el cual se determine la pérdida promedio por zona y distrito a la que se distribuyen los productos de la compañía.
i. Crear un informe en el cual se determine la pérdida promedio por categoría de producto y segmento de cliente.
j. Realizar conclusiones respecto a los resultados obtenidos.

Solucionario

Ejercicio I: Medidas de tendencia central

Tabla 4-2: Justificación para las medidas de tendencia central

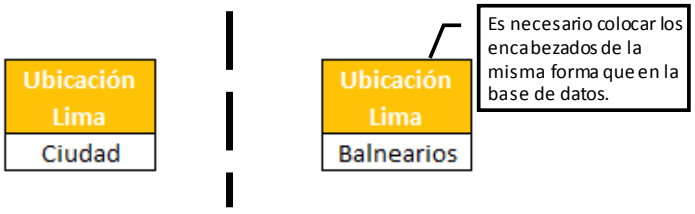
Variable	Datos	Medida	Justificación
Talla de zapatos	35, $35_{1/2}$, 36, 36, 36, 36, 36, 37, $37_{1/2}$, 38, 39	Mediana o moda	La variable "Talla de zapatos", al ser cualitativa-ordinal, puede ser explicada a través de la mediana. Existe un dato que se repite constantemente en el conjunto de datos; es posible emplear la moda.
Ingreso mensual (miles de S/)	2.5, 3, 3.5, 3.8, 4.2, 4.4, 5, 6, 11.2, 12	Mediana	Existen datos extremos que pueden distorsionar la interpretación de todo el conjunto.
Edad	22, 24, 25, 26, 26, 26, 28, 30, 31, 32	Media	Los datos no presentan valores extremos que puedan distorsionar su interpretación.
Color de zapatos preferido	Azul, azul, blanco, blanco, blanco, negro, negro, rojo, rojo	Moda	La moda puede utilizarse tanto para datos cualitativos, como cuantitativos.
Número de compras de calzado en un año	1, 2, 2, 3, 4, 4, 5, 5, 6, 7	Media o mediana	La media toma todos los datos del conjunto. Asimismo, puede ser comparada en distintos conjuntos de datos.

Ejercicio II: Medidas de dispersión

Para este ejercicio, emplee el archivo en Excel "SOL_4".

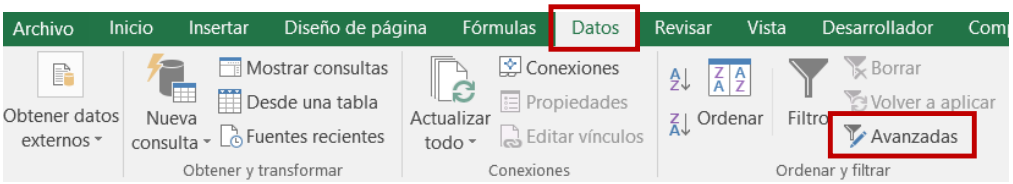
Para realizar el cálculo de las medidas de tendencia central, primero deben hacerse depuraciones en la base de datos de tal manera que se pueda tener una base de datos de los registros de las ventas hechas en la ciudad y los balnearios. Para esto, se emplearán filtros avanzados en Excel. El primer paso para hacerlo es definir los criterios de segmentación; esto se hace copiando el encabezado de la variable a filtrar y en la siguiente celda, debajo, el criterio de segmentación que, para este caso, sería el valor que se desea filtrar, ya sea "Ciudad" o "Balnearios".

Figura 4-1: Definiendo criterios para la aplicación de filtros avanzados



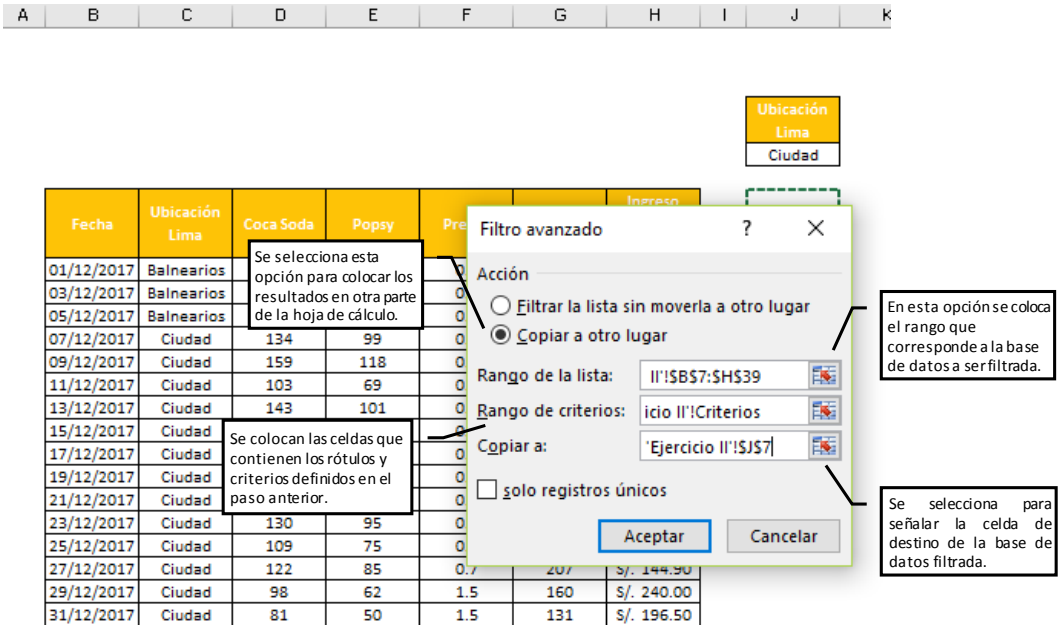
Una vez definidos los criterios, se procede a aplicar los filtros avanzados a la base de datos de la manera que se presenta a continuación:

Figura 4-2: Aplicación de la herramienta filtros avanzados



Posteriormente, se configura la herramienta filtros avanzados, como puede verse en la siguiente figura. Para ver más detalle de este solucionario, ver el archivo SOL_4 (Hoja Ejercicio 4.2).

Figura 4-3: Configuración de filtros avanzados



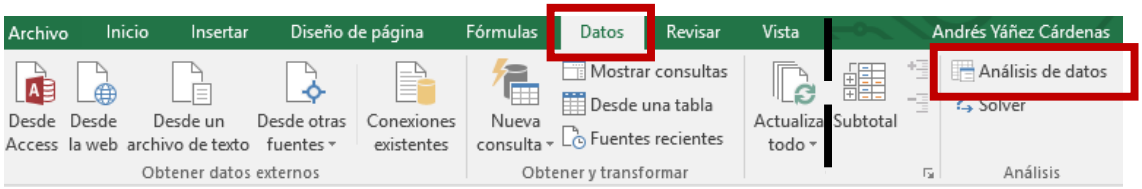
El resultado final será la base de datos filtrada de acuerdo con los criterios establecidos (registros que correspondan únicamente a Lima Ciudad), de manera que el análisis se realizará con 17 datos en vez de 32. A continuación, se aplicará de la misma manera la herramienta filtros avanzados para los "Balnearios"; como resultado se obtendrá una base de datos filtrada de 15 registros en lugar de 32. Hasta aquí es importante mencionar que esta segmentación de la base también podría ejecutarse en SPPS siguiendo estos comandos.

En SPSS:

- **Segmentar archivo:**
  Datos > Seleccionar casos > Seleccionar: Si se satisface la condición > Clic en "Si…" > Ingresar las condiciones de filtro (si son más de una, usar el botón &) > Continuar > Aceptar.

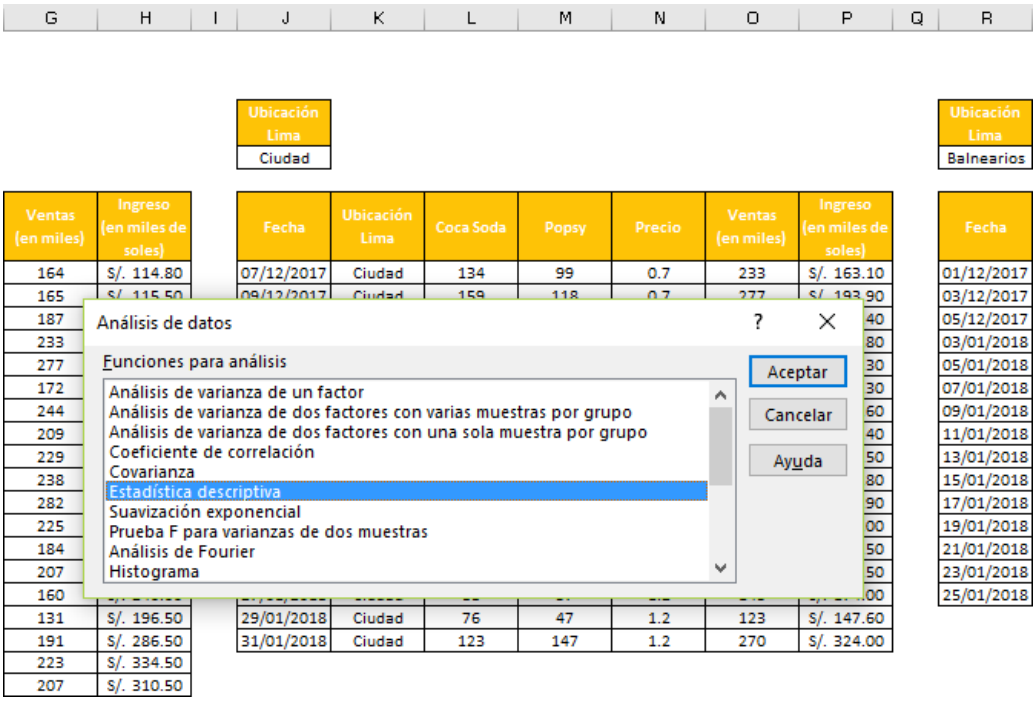
Una vez que se tienen ambas bases de datos, lo siguiente en Excel es aplicar la herramienta "Análisis de datos" [7] a cada una para obtener un reporte con los estadísticos descriptivos que se solicitan; esta herramienta se encuentra en la cinta de opciones en la pestaña datos, como se muestra a continuación:

Figura 4-4: La herramienta Análisis de datos



Seguidamente, se selecciona la opción "Estadística descriptiva" y se da clic en "Aceptar".

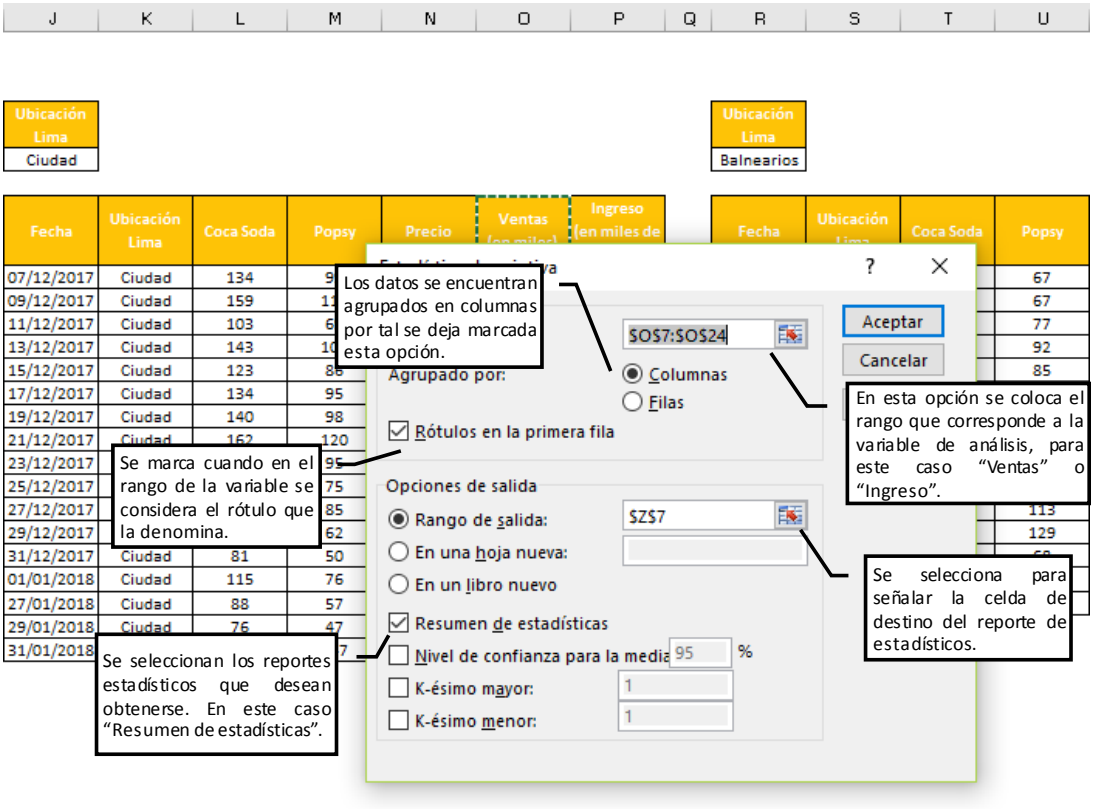
Figura 4-5: Análisis de datos - Estadística descriptiva



[7] Para utilizar la herramienta Análisis de datos, es necesario activarla en la cinta de opciones de Microsoft Excel. En la página de soporte de Microsoft, se presenta un tutorial para hacerlo, titulado "Inicio rápido: activar y usar un complemento de Excel 2016 para Windows" (https://support.office.com/es-es/article/inicio-r%C3%A1pido-activar-y-usar-un-complemento-de-excel-2016-para-windows-3da7ea04-888a-4b32-b064-87de0061f123).

Luego se seleccionan las opciones deseadas para configurar el reporte de estadísticos descriptivos.

Figura 4-6: Configuración de estadísticos descriptivos



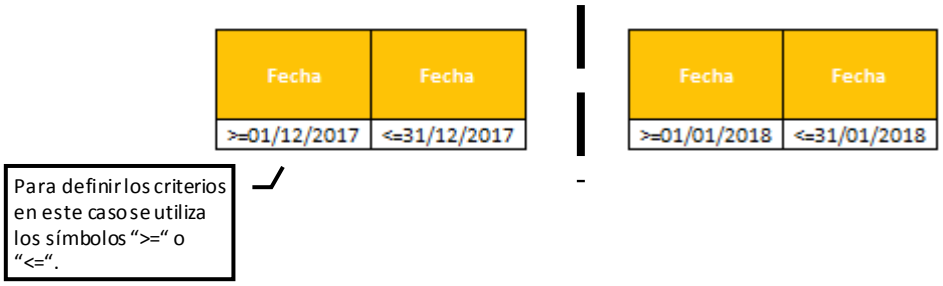
Al finalizar se obtiene como resultado estos reportes en donde se calcula las medidas de tendencia central de los ingresos y las ventas.

Figura 4-7: Medidas de tendencia central de ingresos y ventas

Ubicación Lima: Ciudad	*Ventas (en miles)*	*Ingreso (en miles de soles)*	Ubicación Lima: Balnearios	*Ventas (en miles)*	*Ingreso (en miles de soles)*
Media	207.06	183.45	Media	190.33	241.94
Error típico	12.01	13.19	Error típico	12.63	24.02
Mediana	209	166.60	Mediana	187	206.40
Moda	#N/A	#N/A	Moda	187	#N/A
Desviación estándar	49.50	54.38	Desviación estándar	48.93	93.01
Varianza de la muestra	2450.43	2957.23	Varianza de la muestra	2393.81	8651.76
Curtosis	-0.93	2.02	Curtosis	1.24	-1.19
Coeficiente de asimetría	-0.18	1.52	Coeficiente de asimetría	0.92	0.17
Rango	159	203.6	Rango	192	288.7
Mínimo	123	120.4	Mínimo	113	114.8
Máximo	282	324	Máximo	305	403.5
Suma	3520	3119	Suma	2855	3629
Cuenta	17	17	Cuenta	15	15

Para realizar el cálculo de las medidas de dispersión de las ventas e ingresos mensuales en este ejercicio nos piden depurar la base de datos por meses, para esto se necesita aplicar nuevamente la herramienta "Filtros avanzados". Sin embargo, esta vez se emplean criterios distintos, de la forma que se muestra a continuación:

Figura 4-8: Filtros avanzados para fechas



Empleando estos criterios se ordenará que *"sean filtrados todos los registros mayores o iguales al 01/12/2017 y menores o iguales del 31/12/2017"* esto se hace para filtrar todos los registros del mes de diciembre, de manera similar se hace para el mes de enero. Es así que en el mes de diciembre se obtienen 16 de 32 registros y enero la misma cantidad.

Recuerda que esta segmentación también se puede ejecutar desde SPSS con la ayuda de los siguientes comandos:

En SPSS:

- **Segmentar archivo:**
  Datos > Seleccionar casos > Seleccionar: Si se satisface la condición > Clic en "Si…" > Ingresar las condiciones de filtro (si son más de una, usar el botón &) > Continuar > Aceptar.

Una vez obtenidas las bases de datos filtradas, se aplica la herramienta "Análisis de datos" obteniendo un reporte idéntico al que se vio anteriormente. Para medidas de dispersión, además con los datos del reporte se calcula el "Coeficiente de variación", esto se realiza aplicando esta fórmula a los datos obtenidos:

[4-1] Fórmula del coeficiente de variación

$$CV = \frac{Desviación\ estándar}{Media}$$

Como resultado se obtienen estos valores como medidas de dispersión para cada periodo señalado:

Figura 4-9: Medidas de tendencia central de ingresos y ventas para el 2017

Fecha	Dic-17			Fecha	Ene-18		
Ventas (en miles)		*Ingreso (en miles de soles)*		*Ventas (en miles)*		*Ingreso (en miles de soles)*	
Media	206.69	Media	159.23	Media	191.75	Media	262.50
Error típico	10.82	Error típico	8.73	Error típico	13.70	Error típico	19.59
Mediana	208	Mediana	158.90	Mediana	189	Mediana	283.5
Moda	#N/A	Moda	#N/A	Moda	#N/A	Moda	#N/A
Desviación estándar	43.28	Desviación estándar	34.90	Desviación estándar	54.82	Desviación estándar	78.36
Varianza de la muestra	1872.90	Varianza de la muestra	1218.29	Varianza de la muestra	3004.73	Varianza de la muestra	6141.00
Curtosis	-0.69	Curtosis	0.28	Curtosis	-0.20	Curtosis	-1.22
Coeficiente de asimetría	0.14	Coeficiente de asimetría	0.70	Coeficiente de asimetría	0.58	Coeficiente de asimetría	0.11
Rango	151	Rango	125.20	Rango	192	Rango	255.9
Mínimo	131	Mínimo	114.80	Mínimo	113	Mínimo	147.6
Máximo	282	Máximo	240	Máximo	305	Máximo	403.5
Suma	3307	Suma	2547.70	Suma	3068	Suma	4200
Cuenta	16	Cuenta	16	Cuenta	16	Cuenta	16
CV	20.94%	CV	21.92%	CV	28.59%	CV	29.85%

Sobre el análisis de estadísticos descriptivos, es importante precisar que también se pueden obtener estos resultados desde SPSS, tanto para medidas de tendencia central como las de dispersión.

En SPSS:

- **Medidas de tendencia central y dispersión:**
  Analizar > Estadísticos descriptivos > Frecuencias > Ingresar variables > *Check* en Mostrar tablas de frecuencias > Gráficos: Gráficos de barras y *check* en frecuencias > Continuar > Estadísticos > *Check* en Tendencia central: media, mediana, moda. *Check* en dispersión: desviación estándar, varianza, rango, etc. > Continuar > Aceptar.

Caso aplicado: Al rescate de Arkham

Para este caso, emplee el archivo en Excel "SOL_4".

a. Calcular el monto de pérdida por cada una de las órdenes de compra de la compañía.

Para calcular el monto de pérdida en las órdenes de compra en la compañía, en Excel se procede a crear la variable "Venta proyectada" ("$Cantidad$" × "$Precio\ Unitario$").

En SPSS:

- **Variable calculada venta proyectada:**
  Transformar > Calcular variable > Nombrar la variable venta proyectada en "variable objetivo" > Ingrese la variable con la que trabajará y utilice los signos de operación que necesite: Cantidad*Precio unitario

A continuación, se debe restar el valor de la "Venta real" menos el valor de "Venta proyectada". Para ver más detalle de esta solución, ver el archivo SOL_4 (Hoja Caso 4).

B	C	D	E	F	L	M
N° Orden	Fecha de compra	Cantidad	Precio unitario	Zona	Venta real	Venta proyectada
667	02/01/2014	43	S/ 15.04	Lima Moderna	S/ 614.80	=E3*D3
676	03/01/2014	4	S/ 212.60	Lima Sur	S/ 698.00	S/ 850.40
691	10/01/2014	31	S/ 70.98	Callao	S/ 2,137.10	S/ 2,200.38
710	17/01/2014	40	S/ 30.53	Lima Este	S/ 1,143.49	S/ 1,221.20
711	19/01/2014	18	S/ 207.48	Lima Sur	S/ 3,568.45	S/ 3,734.64
713	21/01/2014	40	S/ 19.98	Lima Norte	S/ 751.94	S/ 799.20
715	21/01/2014	19	S/ 19.98	Lima Norte	S/ 354.13	S/ 379.62
725	26/01/2014	49	S/ 90.98	Lima Sur	S/ 4,321.63	S/ 4,458.02
738	26/01/2014	24	S/ 376.13	Lima Moderna	S/ 7,332.09	S/ 9,027.12
773	06/02/2014	16	S/ 71.37	Lima Este	S/ 950.46	S/ 1,141.92
793	08/02/2014	20	S/ 105.98	Lima Norte	S/ 2,026.01	S/ 2,119.60
812	11/02/2014	17	S/ 60.98	Lima Sur	S/ 1,002.73	S/ 1,036.66
815	14/02/2014	34	S/ 236.97	Lima Moderna	S/ 6,686.34	S/ 8,056.98
854	15/02/2014	46	S/ 417.40	Lima Centro	S/ 18,824.42	S/ 19,200.40
877	15/02/2014	34	S/ 78.69	Lima Moderna	S/ 2,548.30	S/ 2,675.46
903	26/02/2014	15	S/ 55.48	Lima Moderna	S/ 803.04	S/ 832.20
908	05/03/2014	25	S/ 107.53	Lima Sur	S/ 2,553.84	S/ 2,688.25
909	06/03/2014	26	S/ 60.98	Lima Este	S/ 1,523.50	S/ 1,585.48
916	14/03/2014	34	S/ 62.18	Lima Moderna	S/ 1,932.58	S/ 2,114.12
928	20/03/2014	30	S/ 160.98	Lima Norte	S/ 4,620.05	S/ 4,829.40
944	24/03/2014	23	S/ 138.75	Lima Norte	S/ 2,527.79	S/ 3,191.25
949	27/03/2014	31	S/ 55.98	Lima Norte	S/ 1,685.05	S/ 1,735.38
954	10/04/2014	20	S/ 400.98	Lima Norte	S/ 6,449.06	S/ 8,019.60
962	11/04/2014	44	S/ 30.56	Lima Moderna	S/ 1,317.34	S/ 1,344.64
963	11/04/2014	47	S/ 20.98	Lima Sur	S/ 887.45	S/ 986.06
996	14/04/2014	14	S/ 100.98	Lima Moderna	S/ 1,386.65	S/ 1,413.72

Se calcula la "Venta proyectada" multiplicando la "Cantidad" por el "Precio unitario" (Columna "D" x Columna "E")

Figura 4-10: Cálculo de la variable "Venta proyectada"

Figura 4-11: Cálculo de la variable "Pérdida"

B	K	L	M	N
N° Orden	Fecha de entrega	Venta real	Venta proyectada	Pérdida
667	02/01/2014	S/ 614.80	S/ 646.72	=L3-M3
676	04/01/2014	S/ 698.00	S/ 850.40	-S/ 152.40
691	11/01/2014	S/ 2,137.10	S/ 2,200.38	-S/ 63.28
710	17/01/2014	S/ 1,143.49	S/ 1,221.20	-S/ 77.71
711	21/01/2014	S/ 3,568.45	S/ 3,734.64	-S/ 166.19
713	23/01/2014	S/ 751.94	S/ 799.20	-S/ 47.26
715	22/01/2014	S/ 354.13	S/ 379.62	-S/ 25.49
725	28/01/2014	S/ 4,321.63	S/ 4,458.02	-S/ 136.39
738	28/01/2014	S/ 7,332.09	S/ 9,027.12	-S/ 1,695.03
773	07/02/2014	S/ 950.46	S/ 1,141.92	-S/ 191.46
793	11/02/2014	S/ 2,026.01	S/ 2,119.60	-S/ 93.59
812	12/02/2014	S/ 1,002.73	S/ 1,036.66	-S/ 33.93
815	15/02/2014	S/ 6,686.34	S/ 8,056.98	-S/ 1,370.64
854	16/02/2014	S/ 18,824.42	S/ 19,200.40	-S/ 375.98
877	16/02/2014	S/ 2,548.30	S/ 2,675.46	-S/ 127.16
903	28/02/2014	S/ 803.04	S/ 832.20	-S/ 29.16
908	06/03/2014	S/ 2,553.84	S/ 2,688.25	-S/ 134.41
909	07/03/2014	S/ 1,523.50	S/ 1,585.48	-S/ 61.98
916	16/03/2014	S/ 1,932.58	S/ 2,114.12	-S/ 181.54
928	24/03/2014	S/ 4,620.05	S/ 4,829.40	-S/ 209.35
944	25/03/2014	S/ 2,527.79	S/ 3,191.25	-S/ 663.46
949	29/03/2014	S/ 1,685.05	S/ 1,735.38	-S/ 50.33
954	12/04/2014	S/ 6,449.06	S/ 8,019.60	-S/ 1,570.54
962	11/04/2014	S/ 1,317.34	S/ 1,344.64	-S/ 27.30
963	13/04/2014	S/ 887.45	S/ 986.06	-S/ 98.61
996	19/04/2014	S/ 1,386.65	S/ 1,413.72	-S/ 27.07

El cálculo de la "Pérdida" se realiza a partir de la resta entre la "Venta proyectada" y la "Venta real"

b. Calcular los estadísticos resumen correspondientes para esta variable. Interprete los resultados.

Para calcular los estadísticos resumen para la variable "Pérdida", en Excel se emplea la herramienta "Análisis de datos". Una vez aplicada, se obtiene como resultado este reporte:

Tabla 4-3: Estadísticos para "Pérdida"

Pérdida		
Media	-S/	330.80
Error típico		16.06295124
Mediana	-S/	156.37
Moda	-S/	26.93
Desviación estándar		507.9551187
Varianza de la muestra		258018.4027
Curtosis		64.31394549
Coeficiente de asimetría		-5.845469244
Rango		8092.76
Mínimo		-8117.84
Máximo		-25.08
Suma		-330799.507
Cuenta		1000

En SPSS:

- **Análisis de estadísticos descriptivos:**
  Analizar > Estadísticos descriptivos > Frecuencias > Ingresar variables > Estadísticos > *Check* en tendencia central: media, mediana, moda. *Check* en dispersión: desviación estándar, varianza, rango > Continuar > Aceptar.

De acuerdo con la tabla anterior, en promedio se ha presentado una pérdida de S/330.80 producto de las fallas en la producción y distribución de los productos.

El 50 % de las órdenes de compra que presentan pérdidas en 2017 alcanzan un valor de hasta S/156.37.

El valor de pérdidas más recurrente durante los periodos analizados es de S/26.93.

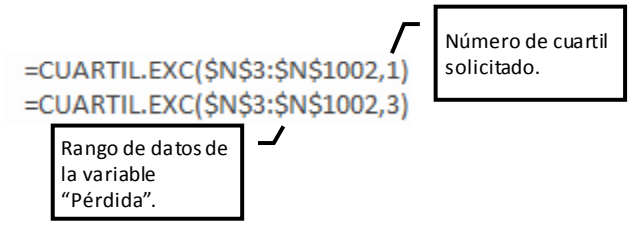
c. Calcule la cantidad de datos atípicos para la variable "Pérdida".

Para calcular los datos atípicos, primero se debe calcular los valores de los cuartiles 1 y 3; ambos cálculos se realizan con la fórmula de Excel CUARTIL.EXC que tiene la siguiente sintaxis:

[4-2] Fórmula de cuartil en Excel

$$= CUARTIL.EXC(Rango\ de\ datos, Número\ de\ Cuartil)$$

Figura 4-12: Aplicando el cálculo de los cuartiles



Una vez calculados los cuartiles, se calcula el rango intercuartil, que es la resta del valor del tercer cuartil menos el primer cuartil. Se obtienen los siguientes resultados:

Tabla 4-4: Rango intercuartil

Q1	-S/	390.18
Q3	-S/	62.23
RIC	S/	327.95

En SPSS:

- **Cálculo de cuartiles:**
  Analizar > Estadísticos descriptivos > Frecuencias > Ingresar variable (escalar) > Estadísticos > *Check* en Percentiles > Añadir 25, 50, 75 > Continuar > Aceptar.

Posteriormente, es necesario hacer el cálculo del límite inferior y superior de los datos; esto quiere decir que cualquier dato que se encuentre por debajo del límite superior o por encima del límite inferior será considerado como atípico. A continuación, se muestran las fórmulas respectivas para el cálculo de cada uno que se procederá a hacer en Excel.

[4-3] Fórmula de límite inferior y superior

$$LI = Q_1 - (1.5 \times RIC)$$
$$LS = Q_3 + (1.5 \times RIC)$$

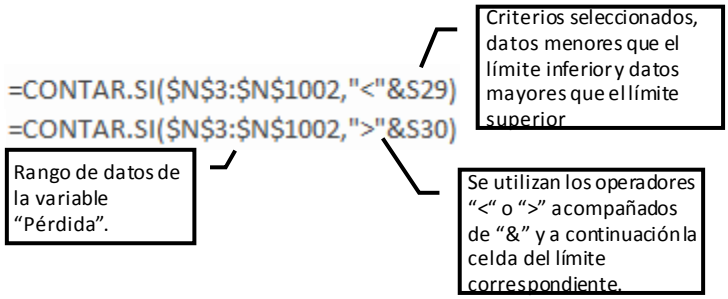
Para finalizar con la estimación de los datos atípicos en la variable "Pérdida", se debe calcular la cantidad de datos que se encuentran por debajo o por encima de

los límites según corresponda; para esto, se emplea la fórmula de Excel CONTAR.SI. Esta tiene la siguiente sintaxis:

[4-4] Fórmula Contar.Si en Excel

$$= CONTAR.SI(Rango\ de\ datos, Criterio\ para\ el\ conteo)$$

Figura 4-13: Aplicando el conteo por criterios



Se obtienen los siguientes resultados al aplicar la fórmula:

Tabla 4-5: Calculando los datos atípicos

Q1	-S/	390.18
Q3	-S/	62.23
RIC	S/	327.95
LI	-S/	882.11
LS	S/	429.70
ATIP INF		95
ATIP SUP		0

En total, se encontraron 95 datos atípicos menores que el límite inferior en la variable "Pérdida".

En SPSS:

- **Identificación de datos atípicos:**
  Datos > Identificar casos atípicos > Ingresar variable (escalar) > Aceptar.

d. De existir datos atípicos, ¿cuál estadístico debe emplear para resumir mejor los datos? ¿Por qué no escoger los otros?

Para este caso, la mejor medida de tendencia central para utilizar sería la MEDIANA, esto debido a la gran cantidad de datos atípicos en el conjunto de datos, lo que afecta el resultado que presenta el valor de la MEDIA.

Por otro lado, la MODA calcula el dato con mayor frecuencia en el conjunto de datos. Sin embargo, el resultado no permite interpretar la tendencia que estos toman.

e. Discriminar los datos atípicos (*outliers*, extremos) y todos los anteriores al año 2017 de la base de datos.

Se emplearán los filtros avanzados de Excel para discriminar los datos atípicos y los datos del año 2017 de la base de datos; para esto, se necesita definir los siguientes criterios:

Tabla 4-6: Criterios para calcular los datos atípicos en años anteriores al 2017

Fecha de compra	Pérdida
>31/12/2016	>-882.11

Una vez definidos los criterios, se procede a aplicar los filtros avanzados a la base de datos, tal y como se realizó en los ejercicios anteriores en esta unidad. Como resultado final, se obtendrá la base de datos filtrada de acuerdo con los criterios establecidos (no cuenta con datos atípicos y todos los anteriores al año 2017), de manera que el análisis se realizará con 283 datos en vez de 1000.

Recuerde que este procedimiento también puede hacerse en SPSS con los siguientes comandos.

En SPSS:

- **Segmentar archivo:**
  Datos > Seleccionar casos > Seleccionar: Si se satisface la condición > Clic en "Si…" > Ingresar las condiciones de filtro (si son más de una, usar el botón &) > Continuar > Aceptar.

f. Calcular las medidas de distribución central y las medidas de dispersión de los nuevos datos obtenidos. Interprete los resultados.

Al aplicar "Análisis de datos" en Excel, se obtienen las medidas de distribución central y dispersión en la base de datos filtrada en el paso anterior.

Tabla 4-7: Nuevos estadísticos para "Pérdida"

Pérdida		
Media	-S/	224.82
Error típico		12.25351974
Mediana	-S/	140.58
Moda		#N/A
Desviación estándar		206.14
Varianza de la muestra		42492.10
Curtosis		0.363125617
Coeficiente de asimetría		-1.162996271
Rango		808.2075
Mínimo		-833.6075
Máximo		-25.4
Suma		-63624.166
Cuenta		283

En SPSS:

- **Medidas de tendencia central y dispersión:**
  Analizar > Estadísticos descriptivos > Frecuencias > Ingresar variables > Estadísticos > *Check* en tendencia central: media, mediana, moda. *Check* en dispersión: desviación estándar, varianza, rango > Continuar > Aceptar.

En 2017 se ha presentado una pérdida de S/224.82 en promedio, producto de las fallas en la producción y distribución de los productos. El 50 % de las órdenes de compra que presentan pérdidas en 2017 alcanzan un valor de S/140.58.

La desviación estándar de los valores de pérdida de las órdenes de compra presenta una variación de S/206.14 con respecto al promedio.

g. Calcular el coeficiente de variación e interpretar los resultados obtenidos.

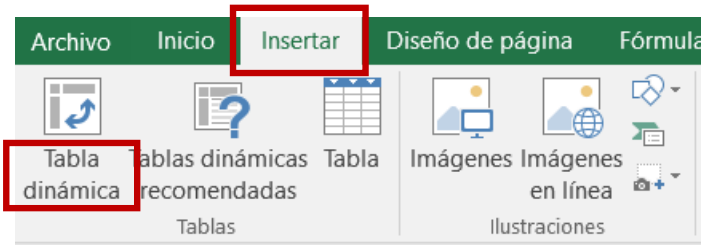
Al aplicar la fórmula, se obtiene que el coeficiente de variación toma un valor de 91.69 %[8], lo que demuestra que existe una alta heterogeneidad en los datos.

[8] Al tenerse un valor de media negativa, se ha colocado este resultado en valor absoluto para la aplicación correcta del coeficiente de variación.

h. Crear un informe en el cual se determina la pérdida promedio por zona y distrito a la que se distribuyen los productos de la compañía.

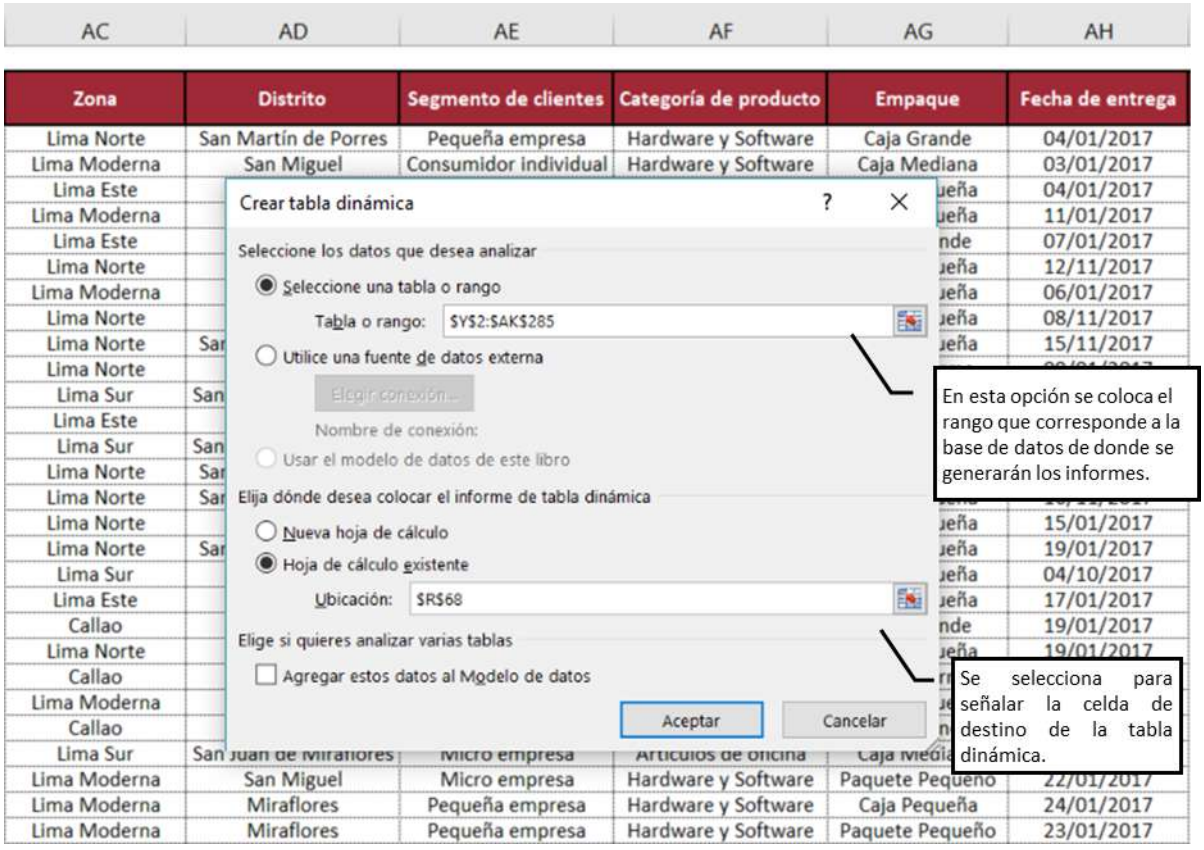
La mejor forma de trabajar reportes con datos cruzados entre variables en Excel es a través de las "Tablas dinámicas". Primero, debe seleccionarse alguna de las celdas que se encuentran dentro de esta base de datos. Una vez realizada esta acción, se realizan los siguientes pasos:

Figura 4-14: Elaborando informes de tablas dinámicas



Seguidamente, se colocan las especificaciones para generar la tabla dinámica en la hoja de cálculo.

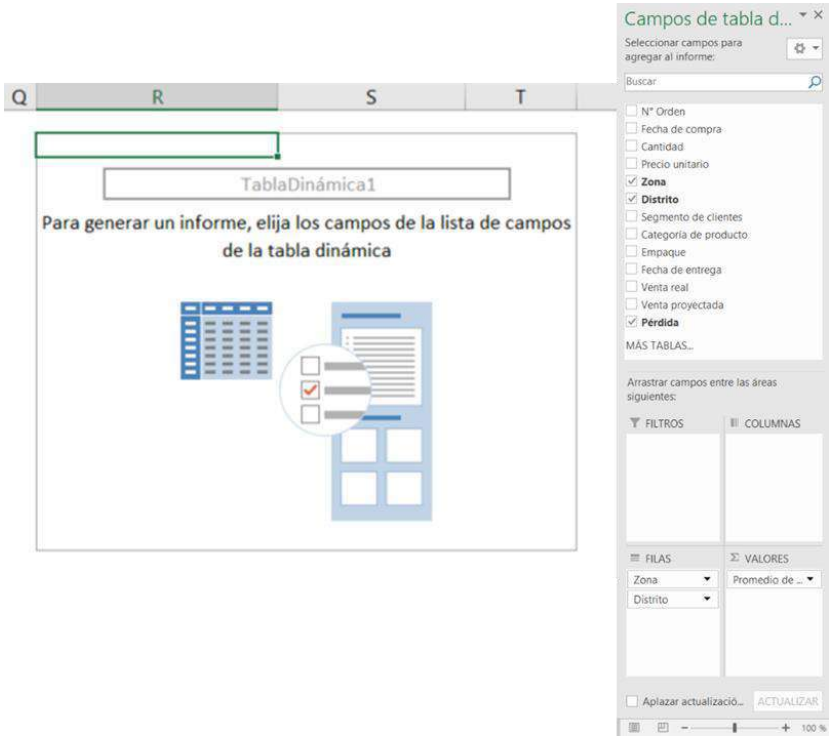
Figura 4-15: Creando una tabla dinámica



AC	AD	AE	AF	AG	AH
Zona	Distrito	Segmento de clientes	Categoría de producto	Empaque	Fecha de entrega
Lima Norte	San Martín de Porres	Pequeña empresa	Hardware y Software	Caja Grande	04/01/2017
Lima Moderna	San Miguel	Consumidor individual	Hardware y Software	Caja Mediana	03/01/2017
Lima Este				ueña	04/01/2017
Lima Moderna				ueña	11/01/2017
Lima Este				nde	07/01/2017
Lima Norte				ueña	12/11/2017
Lima Moderna				ueña	06/01/2017
Lima Norte				ueña	08/11/2017
Lima Norte	Sar			ueña	15/11/2017
Lima Norte					
Lima Sur	San				
Lima Este					
Lima Sur	San				
Lima Norte	Sar				
Lima Norte	Sar				
Lima Norte				ueña	15/01/2017
Lima Norte	Sar			ueña	19/01/2017
Lima Sur				ueña	04/10/2017
Lima Este				ueña	17/01/2017
Callao				nde	19/01/2017
Lima Norte				ueña	19/01/2017
Callao					
Lima Moderna					
Callao					
Lima Sur	San Juan de Miraflores	Micro empresa	Artículos de oficina	Caja Mediana	
Lima Moderna	San Miguel	Micro empresa	Hardware y Software	Paquete Pequeño	22/01/2017
Lima Moderna	Miraflores	Pequeña empresa	Hardware y Software	Caja Pequeña	24/01/2017
Lima Moderna	Miraflores	Pequeña empresa	Hardware y Software	Paquete Pequeño	23/01/2017

Una vez generada la tabla dinámica, es necesario especificar las variables que se emplearán para trabajar los datos; para este ejemplo se configurarán de la siguiente forma. Para más detalle de la solución del ejercicio, ver el archivo SOL_4 (Hoja Caso 4).

Figura 4-16: Los cuadrantes de una tabla dinámica



En Excel:

- **Crear tablas dinámicas:**
  Selecciona la BD > Insertar > Tabla dinámica > Elija dónde desea colocar la tabla: Nueva hoja o en la hoja existente (seleccione un rango de celdas que ocupará la tabla) > Aceptar > Arrastre las variables a analizar en cuadrante columna, fila o filtro y determine cómo trabajará con los siguientes valores: cuenta, promedio, etc.

En SPSS:

- **Crear tablas cruzadas:**
  Analizar > Estadísticos descriptivos > Tablas cruzadas > Inserte las variables a analizar (fila y columna) > Aceptar.

Al finalizar se obtendrá este resultado.

Tabla 4-8: Tabla dinámica para "Pérdida" según zona y distrito

	Promedio de Pérdida	
⊟ Callao	-S/	205.69
Bellavista	-S/	205.69
⊟ Lima Centro	-S/	443.77
Breña	-S/	443.77
⊟ Lima Este	-S/	206.93
Ate - Vitarte	-S/	235.65
Santa Anita	-S/	112.56
⊟ Lima Moderna	-S/	212.62
Jesús María	-S/	121.29
Lince	-S/	113.64
Miraflores	-S/	233.12
San Miguel	-S/	215.23
⊟ Lima Norte	-S/	217.83
Independencia	-S/	216.86
Los Olivos	-S/	206.10
San Martín de Porres	-S/	226.72
⊟ Lima Sur	-S/	240.42
Chorrillos	-S/	257.49
San Juan de Miraflores	-S/	235.30
Total general	**-S/**	**224.82**

i. Crear un informe en el cual se determine la pérdida promedio por categoría de producto y segmento de cliente.

Para trabajar el informe, se siguen los mismos pasos del ejercicio anterior; solo se reemplazan las variables "Zona" y "Distrito" por "Categoría de producto" y "Segmento de clientes" respectivamente.

Tabla 4-9: Tabla dinámica por producto y cliente segmento

	Promedio de Pérdida	
⊟ Artículos de oficina	-S/	130.45
Consumidor	-S/	147.26
Micro empresa	-S/	144.87
Pequeña empresa	-S/	98.40
⊟ Hardware y Software	-S/	290.24
Consumidor	-S/	292.31
Micro empresa	-S/	254.29
Pequeña empresa	-S/	335.10
⊟ Mobiliario	-S/	184.11
Consumidor	-S/	167.92
Micro empresa	-S/	219.60
Pequeña empresa	-S/	161.93
Total general	**-S/**	**224.82**

j. Realizar conclusiones respecto a los resultados obtenidos.

En la zona Lima Centro, en el distrito de Breña, se está presentando la mayor cantidad de pérdidas: S/443.77 en promedio, debido a fallas en la distribución. Este monto representa casi el doble del promedio de los valores de pérdida.

Por otro lado, los productos de *Hardware* y *Software* presentan pérdidas promedio de S/290.24. En esta línea de productos, la situación es más crítica para aquellos dirigidos a pequeñas empresas debido a que estos presentan mayores pérdidas: en promedio S/335.10.

Lecturas recomendadas

- Anderson, D., Sweeney, D., y Williams, T. (2008). *Estadística para administración y economía* (10.ª ed., cap. 3). Cengage Learning Editores.
- Levin, R., y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed. revisada, cap. 3). México: Pearson Educación de México; Prentice Hall.

5. Ideas introductorias de probabilidad, distribución normal y estimación

En este capítulo se busca poner en práctica los conceptos relacionados con los tipos de probabilidad, la distribución de probabilidad, valor esperado y varianza. Asimismo, se hace una revisión de la función de distribución normal estándar, y su aplicación mediante el uso de la tabla de distribución normal.

Palabras clave

- Variable aleatoria
- Probabilidad
- Valor esperado
- Varianza
- Tabla de distribución normal

Ejercicio I: Probabilidades

Para este ejercicio, emplee el archivo en Excel "BD_5".

Según los resultados de la Práctica Calificada 1 del curso de Métodos de Investigación Cuantitativa, se sabe que las notas se distribuyen de forma normal con una media de 10.6 y desviación típica de 4.8.

En función a estos resultados, halle la probabilidad de que un estudiante tomado al azar tenga una nota superior a 14 y esté comprendida entre 10 y 14.

Ejercicio II: Coeficiente de variación

Para este ejercicio, emplee el archivo en Excel "BD_5".

Usted se encuentra a cargo de la cartera de inversiones del banco donde trabaja y, como parte de sus labores, debe decir en qué proyectos invertir para el tercer trimestre del 2016. Asimismo, se le pide que su decisión considere el resultado de la segunda vuelta de las elecciones presidenciales que se realizarán en el Perú; esto último debido al impacto directo que los resultados tendrán en las inversiones.

Se le presentan 3 proyectos: A, B y C, que dependen de quién gane las elecciones presidenciales (candidato 1 o candidato 2). El proyecto A tiene ganancias de 2500 en caso de que gane el candidato 1 y de 2000 en caso de que gane el candidato 2. El proyecto B tiene ganancias de 3000 en caso gane el candidato 1 y pérdidas de 3000 en caso de que gane el candidato 2. Por último, el proyecto C tiene ganancias por 20 000 en caso de que gane el candidato 1 y pérdidas por 8000 en caso de que gane el candidato 2.

Si la probabilidad de que gane el candidato 1 es de 40 %, se le pide que halle la ganancia esperada para cada proyecto. Considerando solamente la ganancia esperada, ¿qué proyecto elegiría? Además, se le pide que halle la varianza y la desviación estándar de cada proyecto. Tomando en cuenta solamente las medidas de dispersión, ¿qué proyecto elegiría?

Caso aplicado: El despegue de Kichwa Wasi[9]

Para este caso, emplee el archivo en Excel "BD_5".

Kichwa Wasi, cuyo nombre puede ser traducido al español como "conociendo el quechua", era una organización civil sin fines de lucro que buscaba contribuir a la revaloración de la cultura andina y lengua quechua en el país. Nació por el sueño de un grupo de jóvenes que identificaba una problemática alrededor de la lengua quechua y la cultura andina al ver que estas eran relegadas frente a otros idiomas y culturas más comerciales. Este sueño se inició en febrero de 2012 con tres integrantes: dos de ellos

[9] Departamento Académico de Ciencias de la Gestión (2016). Apuntes de clase: estudios de caso #1.

gestores sociales egresados de la PUCP y el tercero un educador bilingüe intercultural egresado de la Universidad Nacional Santiago Antúnez de Mayolo (UNASAM).

Ricardo Hernández Palacios, quien había tenido una conexión muy fuerte con el mundo andino durante su juventud temprana al haber viajado por casi todo el país durante un año entero, era a quien desde hacía varios años le rondaba la idea de iniciar un emprendimiento social relacionado con la cultura andina. Hasta entonces, había estado esperando conocer a las personas que pudieran acompañarlo en esta aventura.

Kichwa Wasi llegó a contar con tres tipos de intervención. Primero, la realización de clases de lengua quechua; segundo, la implementación de turismo quechua; tercero, servicios de traducción al idioma quechua. En el primer caso, las clases se realizaban con el apoyo indispensable de la Facultad de Educación de la PUCP, el cual consistía en el otorgamiento de aulas para impartir las clases, lo que permitió que los costos se redujeran y se vieran reflejados en una mayor participación de personas.

El despliegue de estas intervenciones requería que se utilizaran distintas herramientas estadísticas para su gestión. En el caso de la realización de clases de quechua, los integrantes debían predecir los niveles de matrícula a fin de solicitar las aulas adecuadas. La experiencia les llevó a sistematizar los distintos niveles de matrícula de los primeros 24 meses consecutivos de clases.

a) Construir una distribución de probabilidad con los datos ofrecidos.
b) Calcular el valor esperado de alumnos matriculados al mes.
c) Calcular la varianza de la distribución.

La implementación del componente "turismo quechua", por su parte, se llevó a cabo en el distrito de Chacas, provincia de Asunción (Áncash). El plan consistió en generar una sinergia entre el potencial turístico del distrito, los servicios de la zona (como hospedajes, restaurantes, visitas guiadas, entre otros), y la celebración de festividades en la localidad. Para esto, el equipo de trabajo realizó múltiples visitas a la zona, recopilando información sobre el tema.

El equipo decidió indagar sobre la posibilidad de contratar servicios particulares de transporte que trasladaran a las personas interesadas desde Huaraz a Chacas, a fin de aprovechar la carretera recientemente inaugurada por el Gobierno regional. La carretera permitía reducir el tiempo de viaje a Chacas en un 50 %, por lo que resultaba estratégico fomentar su uso. El Gobierno local, entusiasmado por la idea, ofreció

solventar la compra de la movilidad. Sin embargo, el equipo debía incluir en su plan el coste del combustible.

Para esto, se registró el costo de combustible para cada traslado entre Chacas y Huaraz durante la etapa de trabajo de campo (32 traslados). Con el paso del tiempo, el equipo se percató de que los datos tenían una distribución normal denotada por x~N (145;19.6).

d) Calcule la probabilidad de que el consumo de gasolina durante el traslado de Huaraz a Chacas sea:

- inferior a S/150;
- comprendida entre S/138 y S/150.

Finalmente, otro aspecto fundamental para articular la oferta y demanda de estos servicios de manera sostenible y rentable, fue el análisis del gasto estimado de los turistas durante la visita a Chacas. Para esto, el equipo levantó información de 300 personas que visitaron el distrito durante los meses de enero, febrero y marzo de 2013.

e) Hallar la estimación puntual de lo siguiente:

- Gasto estimado diario promedio en la muestra.
- Varianza de la muestra.
- Desviación estándar de la muestra.
- Error estándar de la muestra.

f) Calcule los intervalos estimados que incluyan la media de la población el 95 % y el 99 % de veces.

Solucionario

Ejercicio I: Probabilidades

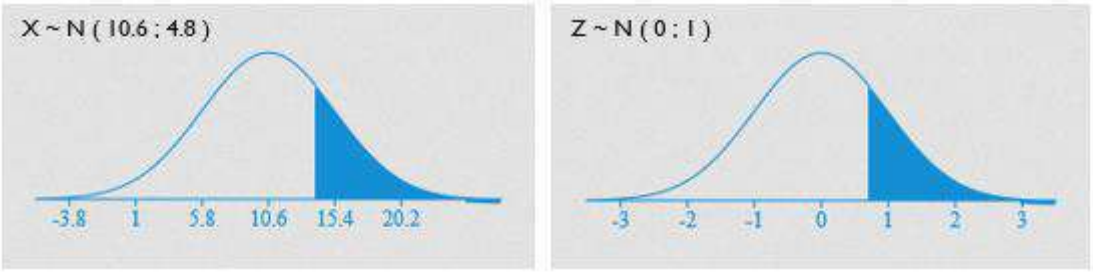
Para este ejercicio, emplee el archivo en Excel "SOL_5".

De acuerdo con el enunciado, los resultados de la Práctica Calificada I son parte de una distribución de probabilidad representada por $N(10.6; 4.8)$ en donde el primer valor corresponde a la media y el segundo valor a la desviación estándar.

Para hallar la probabilidad de un determinado conjunto de valores (en este caso, la probabilidad de todos los valores superiores a 14) es necesario convertir la

distribución normal del ejercicio en una distribución normal estandarizada $N(0;1)$, como se muestra a continuación.

Figura 5-1: Estandarización de una distribución normal



La distribución de la izquierda es la propuesta por la premisa del ejercicio, mientras que la de la derecha es la versión estandarizada de esta. Las áreas azules de ambas distribuciones son equivalentes, pero es más fácil hallar el área de la derecha ya que se trata de una estandarización. Se sabe que el punto de corte en el primer gráfico es 14 y se necesita saber cuál es el punto de corte estandarizado. Para esto, se utiliza la siguiente fórmula:

[5-1] Fórmula de la estandarización de una variable

$$Z = \frac{(\mathrm{x} - \mu)}{\sigma}$$

Donde:

x: Media muestral

μ: Error estándar

σ: Desviación típica

Reemplazando, se obtiene que el valor equivalente a $x = 14$ es $z = 0.71$. Para hallar el área a la derecha del punto de corte estandarizado 0.71, usamos la fórmula Excel:

[5-2] Fórmula para hallar la probabilidad normal en Excel

$$= 1 - DISTR.NORM.ESTAND(z)$$[10]

Donde:

"z" corresponde a valores estandarizados asociados a los puntos de corte.

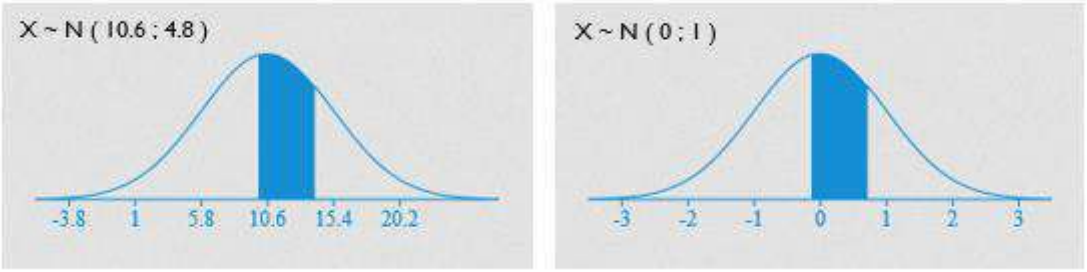
Como resultado se obtiene que la probabilidad de que la media sea mayor a 14 es de 23.89 %.

De la misma forma, para hallar la probabilidad de un determinado conjunto de valores (en este caso los valores entre 10 y 14), se debe transformar la distribución (y los

[10] La fórmula $DISTR.NORM.ESTAND(z)$ calcula la probabilidad a la izquierda del punto de corte.

puntos de corte) del ejercicio en una distribución normal estandarizada tal y como se muestra a continuación.

Figura 5-2: Distribución normal estandarizada



Se sabe que las áreas azules son equivalentes y que es más fácil calcularla en una distribución normal estandarizada (derecha). Para hacer este cálculo, se debe transformar los puntos de corte $x_1 = 10$ y $x_2 = 14$ en valores estandarizados; para ello, se usa la fórmula de estandarización mostrada anteriormente. Así, $x_1 = 10$ es equivalente a $z_1 = -0.13$, y $x_2 = 14$ es equivalente a $z_2 = 0.71$.

El área azul corresponde al área que se encuentra a la izquierda de 0.71, si se le resta el área que se encuentra a la izquierda de -0.13, utilizando la fórmula Excel:

[5-3] Probabilidad de un intervalo en Excel

$$= DISTR.NORM.ESTAND(z_2) - DISTR.NORM.ESTAND(z_1)$$

Donde:

"z_1" y "z_2" corresponden a valores estandarizados asociados a los puntos de corte.

De esta forma, la probabilidad de que la media se encuentre entre 10 y 14 es 31.29 %.

Ejercicio II: Coeficiente de variación

Para este ejercicio, emplee el archivo en Excel "SOL_5".

El Ejercicio II propone los casos de tres posibles inversiones, donde cada una de estas posee una determinada cantidad en caso de que gane el candidato 1 o el candidato 2. Asimismo, se sabe que cada candidato cuenta con una probabilidad determinada de ganar.

Tabla 5-1: Resultados por proyectos y candidatos

Proyecto	Candidato 1 (x_1)	Candidato 2 (x_2)
A	2500	2000
B	3000	-3000
C	20 000	-8000
Probabilidad	0.4	0.6

Para hallar el valor esperado de cada caso se debe seguir la siguiente fórmula.

[5-4] Valor esperado por escenario

$$E(x) = (x_1 \times p_1) + (x_2 \times p_2) \ldots + (x_n + p_n)$$

De esta forma, el valor esperado de cada caso corresponde a lo siguiente:

Tabla 5-2: Valor esperado por proyecto

Proyecto	E(x)
A	2200
B	-600
C	3200

De acuerdo con los resultados de la tabla anterior, se elegiría el proyecto C, ya que es el que maximiza el posible beneficio para el inversionista. Sin embargo, existen otros aspectos que deben ser considerados para tomar esta decisión, los cuales se explican a continuación.

Se debe recordar las siguientes fórmulas para la obtención de las medidas de dispersión para el caso de una distribución de probabilidad:

[5-5] Varianza

$$Var(x) = \left[p_1 \times \left(x_1 - E(x)\right)\right]^2 + [p_2 \times (x_2 - E(x))]^2 + \ldots + [p_n \times (x_n - E(x))]^2$$

[5-6] Desviación estándar

$$Desv(x) = \sqrt{Var(x)}$$

Lo cual genera los siguientes resultados:

Tabla 5-3: Varianza y desviación por proyecto

Proyecto	Var(x)	Desv(x)
A	60 000.00	244.95
B	8 640 000.00	2939.39
C	188 160 000.00	3717.14

Para poder elegir uno de los proyectos, es importante comparar la dispersión de cada caso. Asimismo, se debe recordar que las desviaciones estándar de distintas distribuciones no son comparables entre sí; para poder realizar una comparación, es necesario utilizar el coeficiente de variación. Los resultados para cada caso son los siguientes:

Tabla 5-4: Coeficiente de variación por proyecto

Proyecto	Coef(x)
A	0.11
B	- 4.90[11]
C	4.29

De esta forma, a pesar de que el proyecto C nos ofrece una maximización del beneficio (como se observó en la tabla 5-2), el proyecto A nos ofrece un resultado más estable de acuerdo con el coeficiente de variación. La decisión dependerá de la intención del inversionista, si es que este decide asumir el riesgo asociado a la variabilidad a fin de conseguir un mayor beneficio o no.

Caso aplicado: El despegue de Kichwa Wasi

Para este caso, emplee el archivo en Excel "SOL_5".

a) Construir una distribución de probabilidad con los datos ofrecidos.

Para poder solicitar los salones adecuados, primero se debe saber el número aproximado (esperado) de alumnos que tomará las clases de quechua. Si no se cuenta con probabilidades exactas, se puede usar una tabla de frecuencias relativas. En este caso, se procede a contar el número de veces que se repite cada cantidad de alumnos matriculados en cada mes.

[11] Siempre que el coeficiente de variación adopte un valor negativo (debido a un promedio negativo), su interpretación se encuentra distorsionada, por lo que no se debe tomar en cuenta en el análisis.

Este procedimiento en Excel se puede realizar a través de una tabla dinámica seleccionando la opción "cuenta". También se puede hacer de forma manual o usando la herramienta filtros. En este caso, la frecuencia relativa cumple la función de probabilidad (p). Puede hallarse dividiendo la frecuencia absoluta entre el número total de datos. Por ejemplo, para el caso de 25 alumnos, del total de 24 meses, se ha repetido 7; entonces, 7/24 es aproximadamente 0.29 o lo que es equivalente: 29 %.

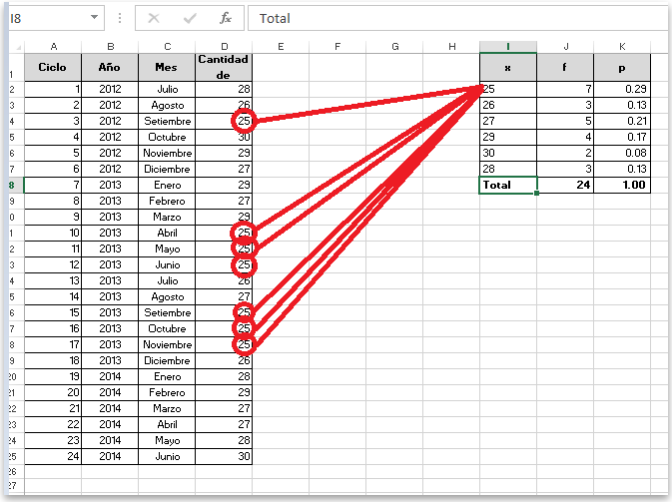
En Excel:

- **Crear tablas dinámicas:**
  Seleccione la BD > Insertar > Tabla dinámica > Elija dónde desea colocar la tabla: Nueva hoja o en la hoja existente (seleccione un rango de celdas que ocupará la tabla) > Aceptar > Arrastre las variables a analizar en cuadrante fila (cantidad de alumnos) y determine cómo trabajará con los valores: cuenta.

En SPSS:

- **Crear tablas de frecuencia:**
  Analizar > Estadísticos descriptivos > Frecuencias > Inserte las variables a analizar.

Figura 5-3: Frecuencias relativas



b) Calcular el valor esperado de alumnos matriculados al mes.

El valor esperado, en este caso, es la cantidad de alumnos que se espera que se matricularán en ese mes. Para hallar el resultado, se debe multiplicar cada cantidad por su respectiva probabilidad y luego sumar.

[5-7] Valor esperado

$$E = x_1 \times p_1 + x_2 \times p_2 + x_3 \times p_3 \ldots$$

En Excel es posible realizar la operación rápidamente, pues es una hoja de cálculo de uso práctico. Se agrega una nueva columna en la que cada fila represente la multiplicación de las cantidades por sus respectivas probabilidades. Por ejemplo, para la primera fila, 25 por 0.29 es igual a 7.29. Luego, se procede a sumar todos los resultados. Se tiene, entonces, que el valor esperado es 27 alumnos matriculados en el mes de julio del 2014.

Figura 5-4: Esperado de alumnos matriculados

x	f	p	x * p
25	7	0.29	7.29
26	3	0.13	3.25
27	5	0.21	5.63
29	4	0.17	4.83
30	2	0.08	2.50
28	3	0.13	3.50
Total	**24**	**1.00**	**27.00**

c) Calcular la varianza de la distribución.

La varianza es un indicador de dispersión de los datos. Mientras mayor sea la dispersión, menor será la posibilidad de hacer estimaciones precisas, por lo que los riesgos incrementan. Para evaluar qué tanta variabilidad existe en la estimación que se hizo en la pregunta anterior, es necesrio calcular la varianza. Para el caso de probabilidades, siga la misma lógica que cuando se halla la varianza poblacional, es decir, al valor esperado se le resta la cantidad "x", se eleva al cuadrado y luego se multiplica por su respectiva probabilidad. Finalmente, se suman todos los resultados.

De manera similar al ejercicio anterior, es posible realizar los cálculos de manera más rápida si se crea una o dos columnas en las que se realice las operaciones paso a paso. Por ejemplo, para la primera línea, en la primera columna se observa que al valor esperado (E(x)=27) se le resta la cantidad (25). El resultado (2) se eleva al cuadrado en la siguiente columna y se multiplica por la probabilidad: 4 por 0.29 resulta 1.17. Finalmente, se suman todos los resultados de esa columna y se tiene que la varianza es de 2.83.

Figura 5-5: Varianza de la distribución

x	f	p	x * p	(E(x)-x))	p*(E(x)-x))^2
25	7	0.29	7.29	2	1.17
26	3	0.13	3.25	1	0.13
27	5	0.21	5.63	0	0.00
29	4	0.17	4.83	-2	0.67
30	2	0.08	2.50	-3	0.75
28	3	0.13	3.50	-1	0.13
Total	**24**	**1.00**	**27.00**		**2.83**

d) Calcule la probabilidad de que el consumo de gasolina durante el traslado de Huaraz a Chacas sea:

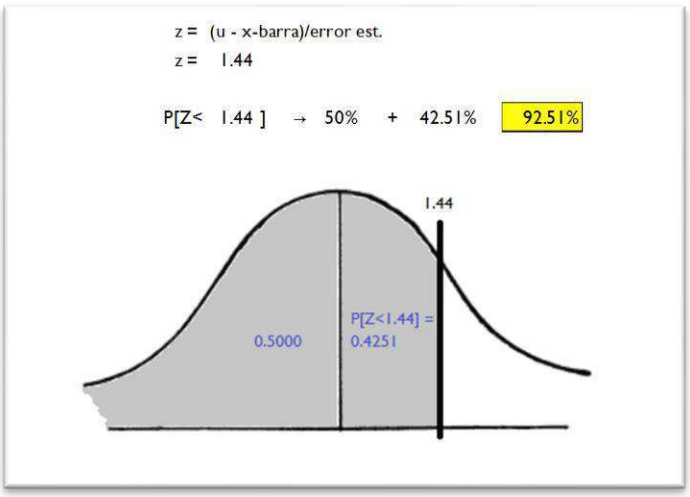
- inferior a S/150;
- comprendida entre S/138 y S/150.

En este caso, no se entrega una tabla con cada uno de los datos, a diferencia del ejercicio anterior. Sin embargo, se señalan los datos más importantes y útiles para los cálculos: 32 casos; una media muestral de 145 y una desviación muestral de 19.6. Con ello se puede calcular el error estándar: simplemente se divide la desviación muestral (19.6) entre la raíz cuadrada del número de elementos (√32). Los datos ordenados serían los siguientes:

$x \sim N(145.0; 19.6)$
$x - barra = 145.0$
$s = 19.6$
$n = 32$
$error\ est. = 3.5$

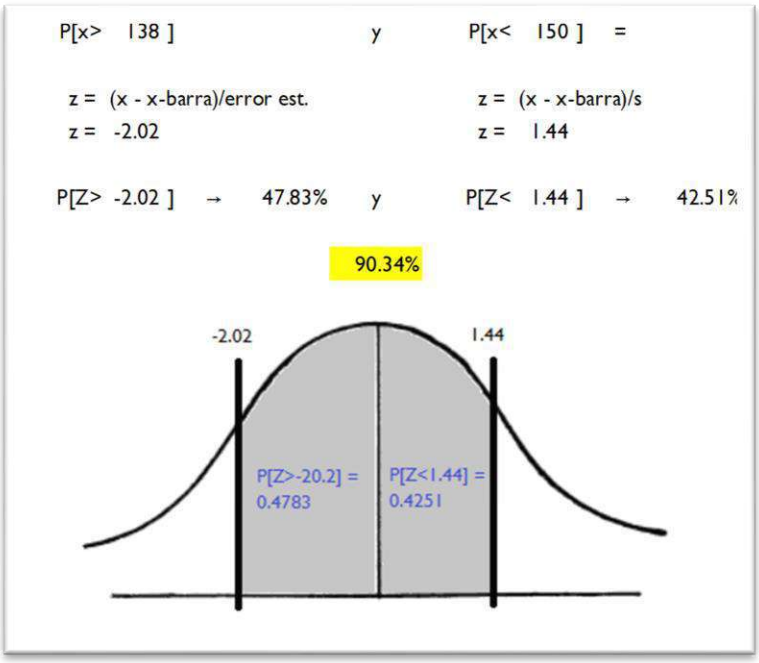
Entonces, primero piden hallar la probabilidad de que el consumo de gasolina sea inferior a S/150. Para esto, el siguiente paso sería calcular el puntaje Z correspondiente al parámetro que se desea comparar, en este caso 150. El procedimiento consiste en restarle al parámetro (150) el promedio muestral (145) y dividirlo entre el error estándar (3.5). El resultado, 1.44, es el puntaje Z. Se puede usar la función de Excel para determinar la probabilidad acumulada desde un inicio hasta el punto 1.44 de una distribución normal o revisar una tabla Z.

Figura 5-6: Probabilidad para consumo inferior a S/50



Lo segundo que piden hallar es la probabilidad de que el consumo de gasolina esté comprendido entre S/138 y S/150. Para saber cuál es la probabilidad de que el costo se encuentre dentro de un rango definido, se debe calcular los dos puntajes Z correspondientes a los extremos de tal rango. Por ejemplo, para el primer caso (138), se le resta el promedio muestral (145) y se divide entre el error estándar (3.5). El resultado (-2.02) debe buscarse en una tabla Z. Una vez que se tengan ambos resultados, se procede a dibujar y calcular el área total. Es importante recordar que se usa una distribución Z, dado que se tiene más de 30 datos.

Figura 5-7: Probabilidad para consumo comprendido entre S/138 y S/150



e) Hallar la estimación puntual de lo siguiente:

- Gasto estimado diario promedio en la muestra.
- Varianza de la muestra.
- Desviación estándar de la muestra.
- Error estándar de la muestra.

Se requiere calcular el gasto promedio diario de un turista. Para ello, primero se debe calcular el gasto diario de cada uno de los 300 turistas que formaron parte de la muestra. Para esto, se divide el gasto total entre el número de días que estuvo cada uno de los turistas. Con esa información, se puede proceder a pedir los estadísticos descriptivos en Excel. Así, se obtendrá el promedio que vendría a ser el gasto promedio de un turista en Chacas.

Figura 5-8: Gastos diarios estimados

Encuesta	Sexo (M=0, H=1)	Días de estadía	Gasto Estimado	Gasto Diario Estimado
1	Hombre	6	165.6	27.6
2	Mujer	2	33.8	16.9
3	Hombre	5	235.5	47.1
4	Hombre	1	35.3	35.3
5	Mujer	2	87.6	43.8
6	Hombre	2	78.6	39.3
7	Hombre	2	92.4	46.2
8	Mujer	3	112.2	37.4
9	Hombre	6	117	19.5
10	Mujer	7	338.8	48.4
11	Hombre	7	250.6	35.8
12	Mujer	4	135.6	33.9
13	Hombre	5	249.5	49.9
14	Mujer	1	35.8	35.8
15	Mujer	1	50.6	50.6
16	Mujer	3	108.6	36.2

Gasto Diario Estimado (Estadísticos Descriptivos)		*Con fórmula*
Media	33.27	33.27
Error típico	0.60	0.60
Mediana	34.55	
Moda	27.60	
Desviación estándar	10.43	10.43
Varianza de la muestra	108.75	108.75
Curtosis	-1.17	
Coeficiente de asimetría	-0.12	
Rango	35.90	
Mínimo	15.00	
Máximo	50.90	
Suma	9,981.10	
Cuenta	300.00	

Recuerde que el análisis de estadísticos descriptivos también se puede ejecutar desde SPSS.

En SPSS:

- **Medidas de tendencia central y dispersión:**
  Analizar > Estadísticos descriptivos > Frecuencias > Ingresar variables > *Check* en Mostrar tablas de frecuencias > Gráficos: Gráficos de barras y *check* en frecuencias > Continuar > Estadísticos > *Check* en Tendencia central: media, mediana, moda. *Check* en dispersión: desviación estándar, varianza, rango, etc. > Continuar > Aceptar.

f) Calcule los intervalos estimados que incluyan la media de la población el 95 % y el 99 % de veces.

Para hacer una estimación con intervalos de confianza es necesario, igual que en los ejercicios anteriores, calcular el error estándar. Los límites del intervalo de confianza están dados por la siguiente fórmula:

[5-8] Límites de un intervalo de confianza

$$IC = [x - error\ est. \times Z; x + erro\ est. \times Z]$$

En donde el valor de Z depende del nivel de confianza. Es útil recordar que el valor de Z para un nivel de confianza del 95 % es 1.96. A continuación, se muestran las dos tablas de resumen para los dos casos propuestos por el ejercicio.

Tabla 5-5: Cálculo del intervalo de confianza

Media muestral (x-barra)	33.27
Cuenta (n)	300
Desviación estándar (s)	10.43
Nivel de confianza	95 %
Estadístico (Z)	1.96
Intervalo de confianza	
Límite inferior	32.09
Límite superior	34.45

Media muestral (x-barra)	33.27
Cuenta (n)	300
Desviación estándar (s)	10.43
Nivel de confianza	99 %
Estadístico (Z)	2.575
Intervalo de confianza	
Límite inferior	31.72
Límite superior	34.82

Con un 95 % de confianza, la media poblacional del Gasto Diario estimado se encuentra en el rango de S/32.09 y S/34.45 soles. Mientras que, con un 99 % de confianza, la media poblacional del gasto diario estimado se encuentra en el rango de S/31.72 y S/34.82 soles.

Lecturas recomendadas

- Anderson, D., Sweeney, D., y Williams, T. (2008). *Estadística para administración y economía* (10.ª ed., caps. 4 y 5). Cengage Learning Editores.

- Levin, R., y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed. revisada, caps. 4 y 5). México: Pearson Educación de México; Prentice Hall.

6. Inferencia estadística: Test de hipótesis

El presente capítulo tiene como objetivo lograr que el alumno sea competente en el uso de datos muestrales para identificar características de una población. Además, se espera que el alumno sea capaz de definir la hipótesis nula y alternativa, el tipo de distribución adecuado para cada caso, y la interpretación de los resultados obtenidos en el análisis estadístico.

Palabras clave

- Distribución Normal
- Distribución t-Student
- Pruebas de hipótesis de 1 cola
- Pruebas de hipótesis de 2 colas

Ejercicio I: Prueba de hipótesis para la media poblacional

Para este ejercicio, emplee el archivo en Excel "BD_6".

La Oficina de Análisis Económico del Instituto de Estadísticas e Informática (INEI) informó que la media del ingreso anual de un residente de Lima, en los últimos cinco años, fue de S/18 688. Un investigador social decide probar lo siguiente:

$$H_0: \mu = S/.18{,}688$$
$$H_1: \mu \neq S/.18{,}688$$

Donde:
Siendo μ la media del ingreso anual de un residente de Lima.

¿Cuál es la conclusión de la prueba de hipótesis del investigador social si en una muestra de 400 residentes de Lima se obtuvo un ingreso medio anual de S/16 860 y una desviación estándar de S/14 624? Emplee un nivel de significancia de 0,05.

Ejercicio II: Prueba de hipótesis de proporciones

Para este ejercicio, emplee el archivo en Excel "BD_6".

Una famosa sala de cine del medio planea una oferta especial que permita a los clientes con tarjetas de socio comprar vasos de diseño especial con conocidas películas nacionales que se estrenarán pronto. Así que, si más del 15 % de los clientes compran

esos vasos, se implantará la promoción. En una prueba preliminar en varios locales, 88 de 500 clientes los compraron.

¿Se debe implantar la promoción especial? Lleve a cabo una prueba de hipótesis que apoye su decisión.

Caso aplicado: Análisis de un segmento de negocio de la empresa YITEL Telefonía Móvil

Para este caso, emplee el archivo en SPSS "BD_6".

Como parte de sus funciones como analista comercial de la división "Personas" en "YITEL Telefonía Móvil", usted tiene acceso periódicamente a información de los clientes según el plan de telefonía a la que se encuentran afiliados. En esta última parte del año, le corresponde analizar el comportamiento de los clientes del plan "Ahorro es progreso" para que en función a los resultados pueda ajustar la nueva propuesta del plan en precio, minutos extras, promociones, entre otras cosas, en miras al 2019. Cabe mencionar que "Ahorro es progreso" es un plan dirigido a personas del sector socioeconómico C y D, bajo la hipótesis de que son clientes que usan su plan sobre todo para sociabilizar, por lo que prefieren tener más megas para acceso a internet y redes sociales. Además, prefieren a YITEL por los precios accesibles que tienen sus planes; sin embargo, estarían dispuestos a migrar a otro operador siempre que les ofrezca alguna opción más adecuada a sus preferencias. En base a esta información, usted debe resolver las siguientes situaciones:

a) El plan "Ahorro es progreso" es una propuesta adaptada a la que ofrece YITEL en Colombia y fue diseñado considerando que el promedio de ingresos familiares anuales de los sectores C y D del país al que se aplique sea equivalente a $18 000 (S/60 000 al tipo de cambio actual). En función a la información con la que cuenta y considerando un nivel de confianza de 90 %, ¿es conveniente su aplicación en el Perú?

b) En el informe de análisis del trimestre anterior se mencionaba que si bien el precio del plan "Ahorro es progreso" es de 50 soles mensuales, los datos de facturación evidencian que los usuarios suelen hacer recargas adicionales por lo que en las recomendaciones se sugiere realizar un ajuste al precio. En función a la evidencia estadística, ¿qué podría afirmar sobre este tema?

c) El plan "Ahorro es progreso" ofrece megas de internet ilimitado y 200 minutos para llamadas a cualquier destino a nivel nacional. Desde hace meses su equipo viene discutiendo la opción de que el plan ofrezca llamadas ilimitadas en vez de un paquete limitado de minutos. Usted sabe que eso solo es conveniente cuando los usuarios consumen más de 300 minutos en llamadas. En función a la información con la que cuenta, ¿el plan debería ofrecer llamadas ilimitadas? En función a los resultados, ¿qué otra cosa podría proponer a la división a la que pertenece?

d) El primer semestre del año, su equipo trabajó muy duro para crear el índice "Propensión al abandono". Este considera variables como antigüedad del usuario en la empresa, ingreso familiar promedio anual, grado de identificación con la marca, entre otras cosas. El índice tiene valores numéricos continuos y oscila entre 0 y 100, siendo 0 el nivel más bajo de propensión al abandono y 100 el más alto. Usted sabe que si más del 15 % de clientes estaría dispuesto a abandonar a YITEL, su división debería plantear una estrategia comercial aún más agresiva para fidelizarlos. Según la información con la que cuenta, ¿será necesario el diseño de una nueva estrategia comercial?

Solucionario

Ejercicio I: Prueba de hipótesis para la media poblacional

Para este ejercicio, emplee el archivo en Excel "SOL_6".

El planteamiento de la hipótesis está establecido en la premisa del Ejercicio I. Como se recuerda, el interés del investigador se encuentra representado por la Hipótesis Alternativa (H_1). En este caso, el interés del investigador es cotejar si es que la media del ingreso anual de un residente de Lima es distinta a la identificada por el INEI hace cinco años. Los datos provistos por el problema son los siguientes:

$\mu = 18{,}688$
$\bar{x} = 16{,}860$
$s = 14{,}624$
$n = 400$
$\alpha = 0.05$

En primer lugar, debemos establecer el intervalo de confianza a través del cual se identifica el valor de el/los z-teórico(s). Para esto se identifican dos factores importantes:

i) La prueba, de acuerdo con el planteamiento, es de dos colas.
ii) *z-teórico* se determina como el punto de corte que contiene al $\alpha/2$ de casos en la distribución, es decir, 0.05/2 = 0.025. Este punto de corte corresponde a ± 1.96.

Ahora que ya se conoce el z-teórico, se debe calcular el *z-prueba* que representa al valor hipotético. Este valor se halla con la fórmula:

[6-1] Z-prueba

$$z - prueba = \frac{(\bar{x} - x)}{\sigma_{\bar{x}}}$$

Donde:
En donde $\bar{x}$ es la media muestral, y $\sigma_{\bar{x}}$ es el error estándar.

Sin embargo, en este caso no se cuenta con el error estándar como dato. En estos casos se requiere hallar el error estándar a partir de la fórmula:

[6-2] Error estándar

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Donde:
En donde σ es la desviación estándar poblacional. Pero al no contar con la desviación estándar poblacional como dato, se puede utilizar en su reemplazo la desviación estándar proveniente de la muestra (s).

En este caso, el valor de *z-prueba* corresponde a -2.50. Por lo que la evaluación final corresponde a la del siguiente gráfico:

Figura 6-1: Distribución normal estándar Z de dos colas para un nivel de confianza del 95 %

Como se observa, el valor de *z-prueba* cae dentro de la zona de rechazo de la hipótesis nula, por lo que se cuenta con suficiente información estadística para determinar que el ingreso anual de los residentes de Lima difiere a los S/18 688.

Ejercicio II: Prueba de hipótesis de proporciones

Para este ejercicio, emplee el archivo en Excel "SOL_6".

Mientras que en el ejercicio pasado se ha revisado pruebas de hipótesis relacionadas con el valor de una media poblacional, en el Ejercicio II se propone una prueba de hipótesis de proporciones. El tratamiento que se les da a estas pruebas es bastante similar, pero es necesario considerar las siguientes diferencias:

i) Para efectos prácticos, la proporción es el nuevo parámetro que debe ser cotejado, es decir, reemplaza a la media.
ii) El error estándar en el caso de una proporción está determinado por la siguiente fórmula:

[6-3] Error estándar para una proporción

$$\sigma_p = \frac{\sqrt{p(1-p)}}{n}$$

Donde:
p es la proporción, y n es el tamaño de la muestra.

Teniendo en consideración estas diferencias, se procede con los pasos ya descritos. Se consideran los siguientes datos:

$P = 0.15$
$p = 0.176\ (88/500)$
$n = 500$
$\alpha = 0.01$

El primer paso es plantear las hipótesis:

$$H_0: P \leq 0.15$$
$$H_1: P > 0.15$$

Donde:
H_1 representa el interés del investigador; en este caso, el investigador desea conocer si es que la proporción de clientes que compran los vasos de diseño especial es mayor a 0.15. Con esto, inferimos que se trata de una prueba de una cola hacia la derecha.

A continuación, se debe establecer el intervalo de confianza teórico a partir de los datos muestrales. Se usa el valor de alfa para determinar dicho valor de *z-teórico*, el cual equivale a 2.33.

Por último, se halla el valor del z-prueba de acuerdo con la siguiente fórmula:

[6-4] Z-prueba para una proporción

$$z - prueba = \frac{(p - P)}{\sigma_p}$$

Donde:
p es la proporción media muestral, y σ_p es el error estándar de la proporción.

De acuerdo con esto, el valor de *z-prueb*a es 1.53, por lo que la evaluación sería la siguiente:

Figura 6-2: Distribución normal estándar Z de una cola para un nivel de confianza del 95 %

Al caer dentro de la zona de aceptación de la hipótesis nula, podemos decir que no se cuenta con información suficiente para descartar H_0, por lo que la promoción no debería ser implementada.

Caso aplicado: Análisis de un segmento de negocio de la empresa YITEL Telefonía Móvil

Para este caso, emplee el archivo en SPSS "SOL_6".

Antes de pasar a desarrollar las preguntas del caso, siempre se debe corroborar que las variables del modelo tengan distribución normal; para esto se procede a validar el siguiente supuesto mediante dos métodos en SPSS.

Verificar supuesto de normalidad: Corroborar que variable aleatoria se distribuya normalmente en la muestra.

[6-5] Pruebas para validar normalidad

Kolmogorov - Smirnov (si n ≥ 30)

Shapiro - Wilk (si n < 30)

En donde:

H0: La distribución es normal

H1: La distribución no es normal

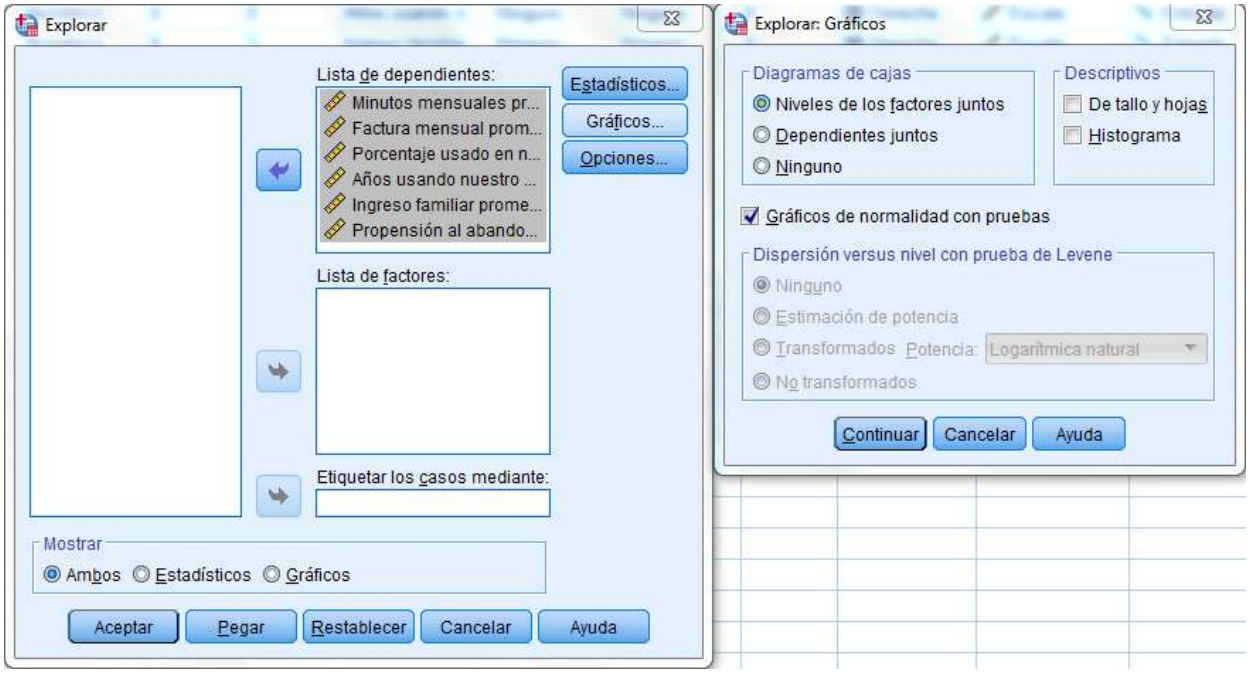
Regla universal: P-VALUE (SIG. ASINTÓTICA) < ALFA → RECHAZO H0

Método 1:

- **Prueba de normalidad (método 1):**
  Analizar > Estadísticos descriptivos > Explorar > Ingresar variables > Gráficos > *Check* en gráficos de normalidad con pruebas > Continuar > Aceptar.

A continuación se muestran las variables que se ingresaron al *software* y el detalle de los comandos ya descritos en la casilla anterior.

Figura 6-3: Prueba de normalidad I con SPSS



Una vez ingresados los datos, SPSS arroja la siguiente tabla con las significancias tanto para la prueba K-S como Shapiro-Wilk.

Tabla 6-1: Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Minutos mensuales promedio	.056	250	.052	.981	250	.002
Factura mensual promedio	.037	250	.200*	.995	250	.673
Porcentaje usado en negocios	.031	250	.200*	.995	250	.635
Años usando nuestro servicio	.034	250	.200*	.997	250	.916
Ingreso familiar promedio anual (Miles de soles)	.043	250	.200*	.996	250	.791
Propensión al abandono	.175	250	.000	.902	250	.000

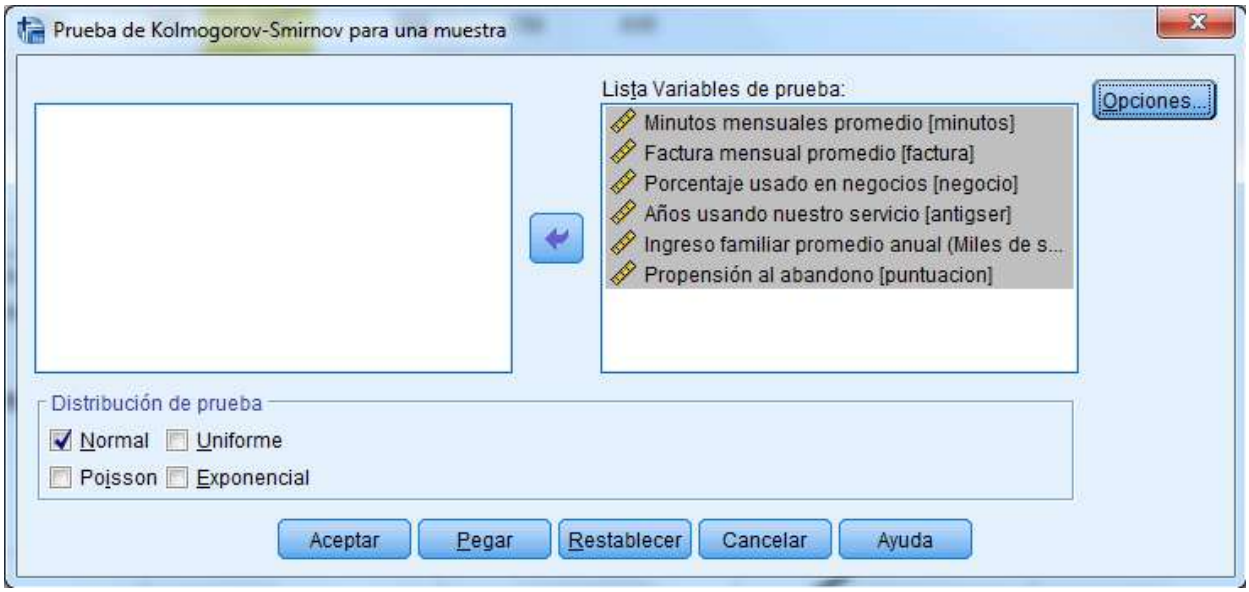
*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Método II:

- **Prueba de normalidad (método 11):**
  Analizar > Pruebas No Paramétricas > K-S de 1 muestra > Ingresar variables > *Check* en "Normal" > Aceptar.

Figura 6-4: Prueba de normalidad II con SPSS



Una vez ingresados los datos, SPSS arroja la siguiente tabla con las significancias para la prueba K-S.

Tabla 6-2: Prueba de Kolmogorov-Smirnov para una muestra

		Minutos mensuales promedio	Factura mensual promedio	Porcentaje usado en negocios	Años usando nuestro servicio	Ingreso familiar promedio anual (Miles de soles)	Propensión al abandono
N		250	250	250	250	250	250
Parámetros normales[a,b]	Media	162.1856	63.3963	32.6847	2.68	61.5896	41.5395
	Desv. Desviación	46.57060	19.79981	9.06560	.604	11.11588	13.32429
Máximas diferencias extremas	Absoluto	.056	.037	.031	.034	.043	.175
	Positivo	.056	.025	.030	.030	.027	.175
	Negativo	-.029	-.037	-.031	-.034	-.043	-.073
Estadístico de prueba		.056	.037	.031	.034	.043	.175
Sig. asintótica(bilateral)		.052[c]	.200[c,d]	.200[c,d]	.200[c,d]	.200[c,d]	.000[c]

a. La distribución de prueba es normal.

b. Se calcula a partir de datos.

c. Corrección de significación de Lilliefors.

d. Esto es un límite inferior de la significación verdadera.

Lectura de resultados para ambos métodos:

Si se utiliza un alfa (nivel de significancia) tradicional de 0.05, se tienen las siguientes conclusiones siguiendo la regla de que para rechazar la Ho, el p-value debe ser menor al alfa. Para esto, se podría utilizar pruebas de hipótesis (de medias) de 1 muestra en todas las variables de nuestra muestra, menos "Propensión al abandono", pues esta última no cumple el supuesto de normalidad.

Tabla 6-3: Interpretación de la significancia

Variable	p-value		Alfa	Interpretación	Distribución
Minutos mensuales promedio	0.052	>	0.050	No rechazo H0	Normal
Factura mensual promedio	0.200	>	0.050	No rechazo H0	Normal
Porcentaje usado en negocios	0.200	>	0.050	No rechazo H0	Normal
Años usando nuestro servicio	0.200	>	0.050	No rechazo H0	Normal
Ingreso familiar promedio	0.200	>	0.050	No rechazo H0	Normal
Propensión al abandono	0.000	<	0.050	Rechazo H0	No Normal

a) El plan "Ahorro es progreso" es una propuesta adaptada a la que ofrece YITEL en Colombia y fue diseñado considerando que el promedio de ingresos familiares anuales de los sectores C y D del país al que se aplique sea equivalente a $18 000 (S/60 000 al tipo de cambio actual). En función a la información con la que cuenta y considerando un nivel de confianza de 90 %, ¿es conveniente su aplicación en el Perú?

Tipo de prueba:

Prueba de Medias de 1 muestra (dos colas)

H0: u = 60

H1: u ≠ 60

Alfa: 0.10 (0.05 en cada cola)

N.C.: 90 %

Figura 6-5: Distribución normal estándar Z de dos colas para un NC del 90 %

Una vez definidas las hipótesis, se procede a ingresar la variable en análisis para que pase la prueba T para una muestra en SPSS con los siguientes comandos.

En SPSS:

- **Prueba T para una muestra:**
  Analizar > Comparar medias > Prueba T para 1 muestra

Figura 6-6: Prueba T para una muestra en SPSS

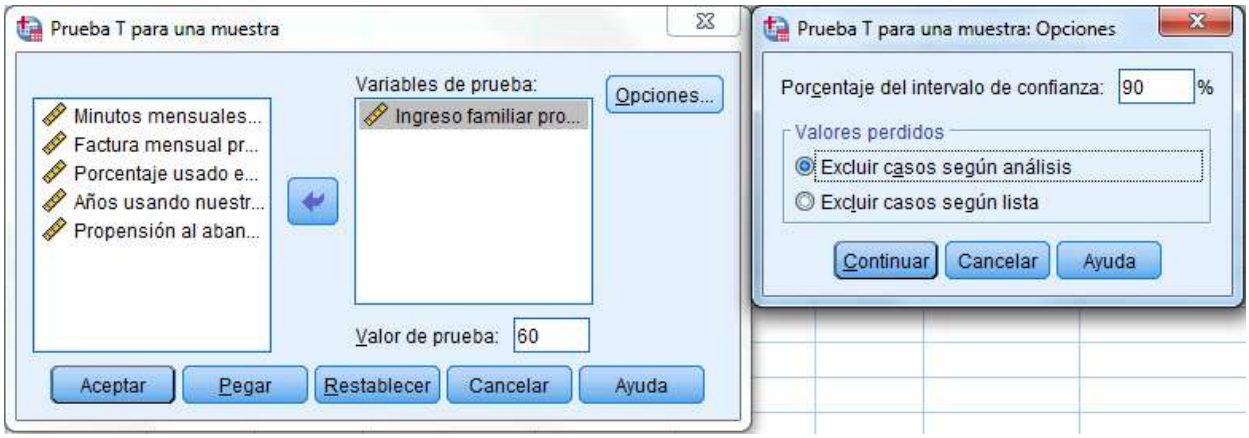


Tabla 6-4: Prueba para una muestra

	Valor de prueba = 60					
	t	gl	Sig. (bilateral)	Diferencia de medias	90% de intervalo de confianza de la diferencia	
					Inferior	Superior
Ingreso familiar promedio anual (Miles de soles)	2.261	249	.025	1.58962	.4289	2.7503

Resultados:

Regla universal: P-VALUE (SIG. ASINTÓTICA) < ALFA → RECHAZO H0

Tabla 6-5: Interpretación de la significancia

Variable	p-value		Alfa	Interpretación
Ingreso familiar promedio	0.025	<	0.100	Rechazo H0

Entonces, una vez hallada la significancia, se puede asegurar que el ingreso familiar promedio de la población es distinto a S/60 000 (H1). Esto implicaría que la aplicación del programa en el Perú no es conveniente.

b) En el informe de análisis del trimestre anterior se mencionaba que si bien el precio del plan "Ahorro es progreso" es de 50 soles mensuales, los datos de facturación evidencian que los usuarios suelen hacer recargas adicionales, por lo que en las recomendaciones se sugiere realizar un ajuste al precio. En función a la evidencia estadística, ¿qué podría afirmar sobre este tema?

Tipo de prueba:

Prueba de Medias de 1 muestra (una cola, derecha)

H0: $u \leq 50$

H1: $u > 50$

Alfa: 0.05 (0.10 en SPSS)

N.C.: 95 % (90 % en SPSS)

Figura 6-7: Distribución normal estándar Z de una cola para un NC del 95 %

Una vez definidas las hipótesis, se procede a ingresar la variable en análisis para que pase la prueba T para una muestra en SPSS con los siguientes comandos.

En SPSS:

- **Prueba T para una muestra:**
  Analizar > Comparar medias > Prueba T para 1 muestra

Importante:

SPSS solo puede reproducir pruebas de dos colas, por lo que se modifica el alfa (y el nivel de confianza) para simular una prueba de este tipo, sabiendo que se tiene interés solo en cola derecha.

Figura 6-8: Prueba T para una muestra en SPSS

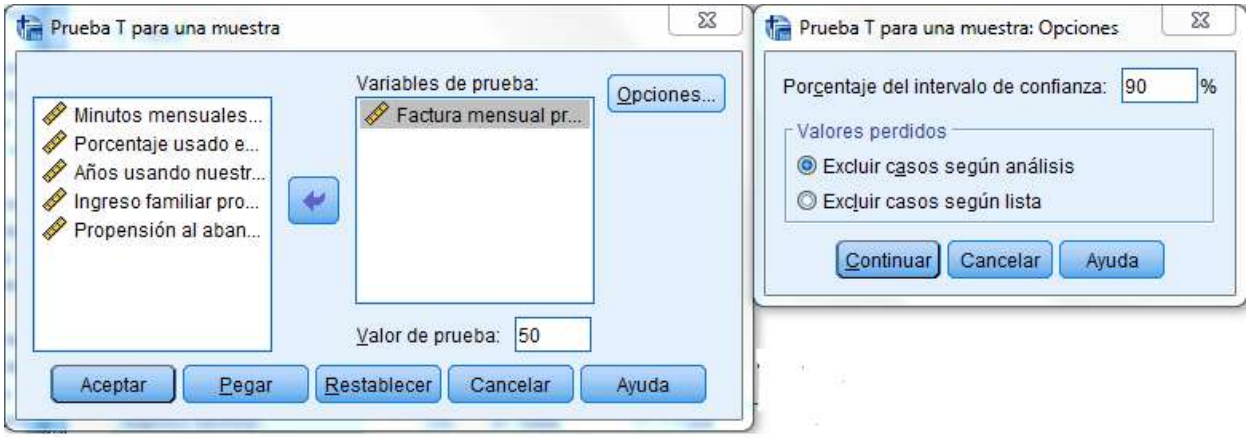


Tabla 6-6: Prueba para una muestra

	Valor de prueba = 50					
	t	gl	Sig. (bilateral)	Diferencia de medias	90% de intervalo de confianza de la diferencia	
					Inferior	Superior
Factura mensual promedio	10.698	249	.000	13.39634	11.3289	15.4638

Resultados:

Se verifica que el valor t se encuentra en cola de análisis. Al ser positivo, se encuentra en cola derecha. Se divide el p-value entre dos como arreglo para la prueba de una cola.

Tabla 6-7: Interpretación de la significancia

Variable	p-value		Alfa	Interpretación
Factura mensual promedio	0.000 / 2	<	0.050	Rechazo H0

Regla universal: P-VALUE (SIG. ASINTÓTICA) < ALFA → RECHAZO H0

De la regla se puede asegurar que la facturación mensual promedio de la población es mayor a S/50 (H1). Esto implicaría que el ajuste de precio del plan sí es conveniente.

c) El plan "Ahorro es progreso" ofrece megas de internet ilimitado y 200 minutos para llamadas a cualquier destino a nivel nacional. Desde hace meses, su equipo viene discutiendo la opción de que el plan ofrezca llamadas ilimitadas en vez de un paquete limitado de minutos. Usted sabe que eso solo es conveniente cuando los usuarios consumen más de 300 minutos en llamadas. En función a la información con la que cuenta, ¿el plan debería ofrecer llamadas ilimitadas? En función a los resultados, ¿qué otra cosa podría proponer a la división a la que pertenece?

Tipo de prueba:

Prueba de medias de 1 muestra (una cola, derecha)

H0: $u \leq 300$

H1: u > 300

Alfa: 0.05 (0.10 en SPSS)

N.C.: 95 % (90 % en SPSS)

Figura 6-9: Distribución normal estándar Z de una cola para un NC del 95 %

Una vez definidas las hipótesis, se procede a ingresar la variable en análisis para que pase la prueba T para una muestra en SPSS con los siguientes comandos.

En SPSS:

- **Prueba T para una muestra:**
  Analizar > Comparar medias > Prueba T para 1 muestra

Importante:

SPSS solo puede reproducir pruebas de dos colas, por lo que se modifica el alfa (y el nivel de confianza) para simular una prueba de este tipo, sabiendo que se tiene interés solo en cola derecha.

Figura 6-10: Prueba T para una muestra en SPSS

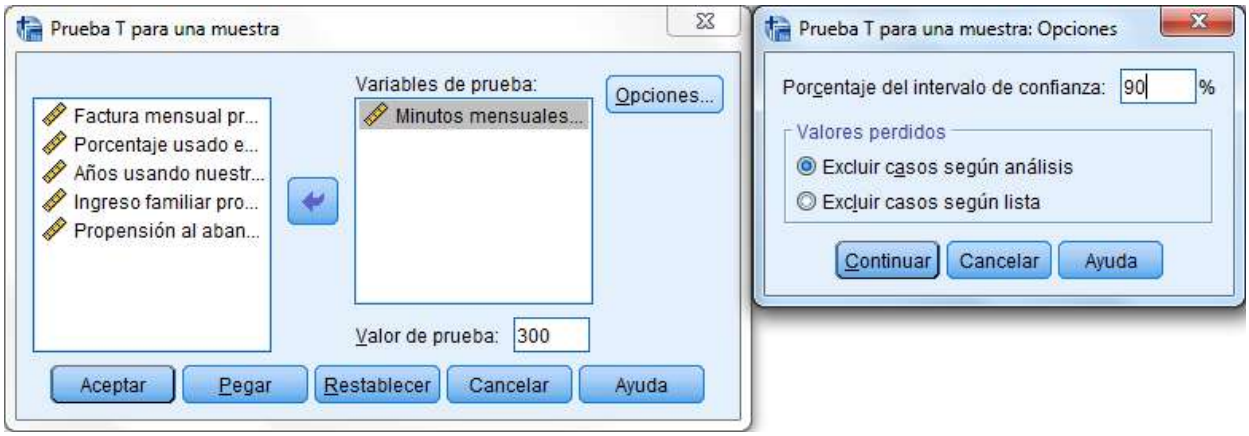


Tabla 6-8: Prueba para una muestra

	Valor de prueba = 300					
					90% de intervalo de confianza de la diferencia	
	t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
Minutos mensuales promedio	-46.790	249	.000	-137.81438	-142.6772	-132.9516

Resultados:

Regla universal: P-VALUE (SIG. ASINTÓTICA) < ALFA → RECHAZO H0

Se verifica que el valor t se encuentra en cola de análisis. Al ser negativo, se encuentra en cola izquierda, es decir, la media muestral cae automáticamente en la zona de aceptación de H0: no se puede rechazar la H0. Entonces, no se puede asegurar que los minutos mensuales promedio de la población sean mayores a 300. Esto implicaría que ofrecer un plan de minutos ilimitado no es conveniente.

d) El primer semestre del año, su equipo trabajó muy duro para crear el índice "Propensión al abandono"; este considera variables como antigüedad del usuario en la empresa, ingreso familiar promedio anual, grado de identificación con la marca, entre otras cosas. El índice tiene valores numéricos continuos y oscila entre 0 y 100, siendo 0 el nivel más bajo de propensión al abandono y 100 el más alto. Usted sabe que si más del 15 % de clientes estaría dispuesto a abandonar a YITEL, su división debería plantear una estrategia comercial aún más agresiva para fidelizarlos. Según la información con la que cuenta, ¿será necesario el diseño de una nueva estrategia comercial?

Tipo de prueba:

Prueba de proporciones de 1 muestra (una cola, derecha)

H0: $\pi \leq 0.15$

H1: $\pi > 0.15$

Alfa: 0.05 (0.10 en SPSS)

N.C.: 95 % (90 % en SPSS)

Figura 6-11: Distribución normal estándar Z de una cola para un NC del 95 %

Importante:

Las pruebas de proporciones buscan analizar la "proporción de ocurrencias" que existen en una muestra y en una población hipotética. Las pruebas de medias analizan variables cuantitativas (escalares), mientras que la proporción de ocurrencias se expresa mediante variables cualitativas (por ejemplo: proporción de mujeres, proporción de niños, proporción de pobres extremos, etc.). El ejercicio pide hallar la proporción de clientes "propensos a abandonar YITEL". Cuando el valor del índice es mayor a 50, se puede catalogar a la persona como "propensa a abandonar". Esto servirá para construir la siguiente variable cualitativa:

Variable: 0: no propenso al abandono ("puntuación" ≤ 50)

1: propenso al abandono ("puntuación" > 50)

El valor (1) siempre debe responder a la ocurrencia que se quiere analizar. En este caso se necesita analizar a los clientes propensos al abandono.

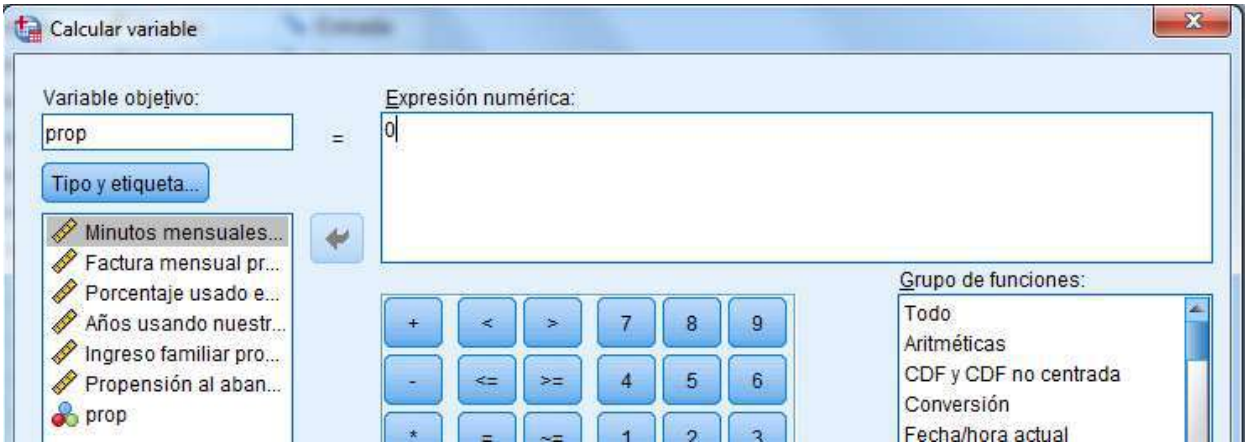
Método 1:

Crear variable "prop" con valor base (0).

En SPSS:

- **Calcular variable:**
  Transformar > Calcular variable: prop

Figura 6-12: Creación de la variable "prop", base 0



Luego, sobreescribir variable "prop" con valor (1) cuando cumple condición (puntuación > 50)

Figura 6-13: Creación de la variable "prop", valor 1

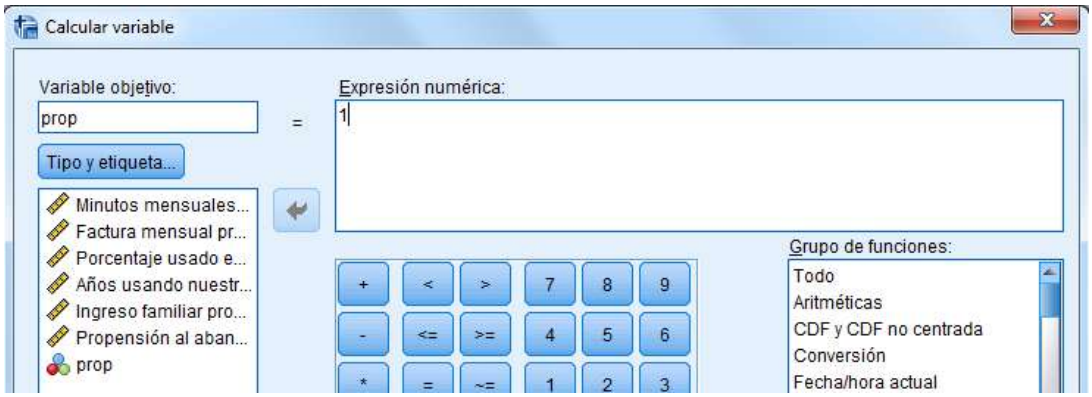
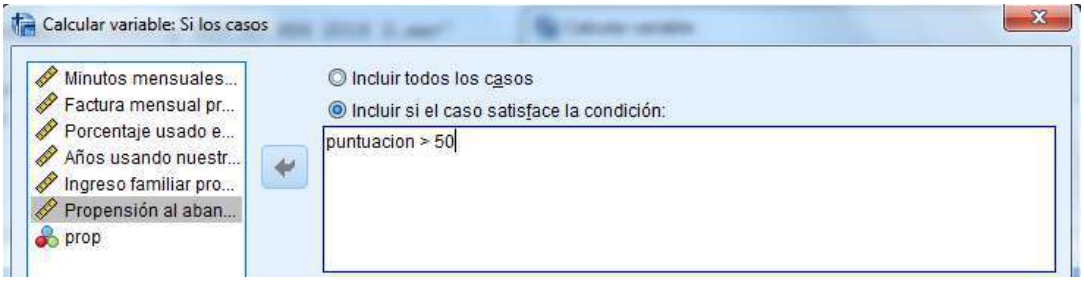


Figura 6-14: Creación de la variable "prop", valor 1



Método II:

Recodificar la variable "puntuación".

En SPSS:

- **Recodificación en distintas variables:**
  Transformar > Recodificar en distintas variables

Figura 6-15: Creación de la variable "prop"

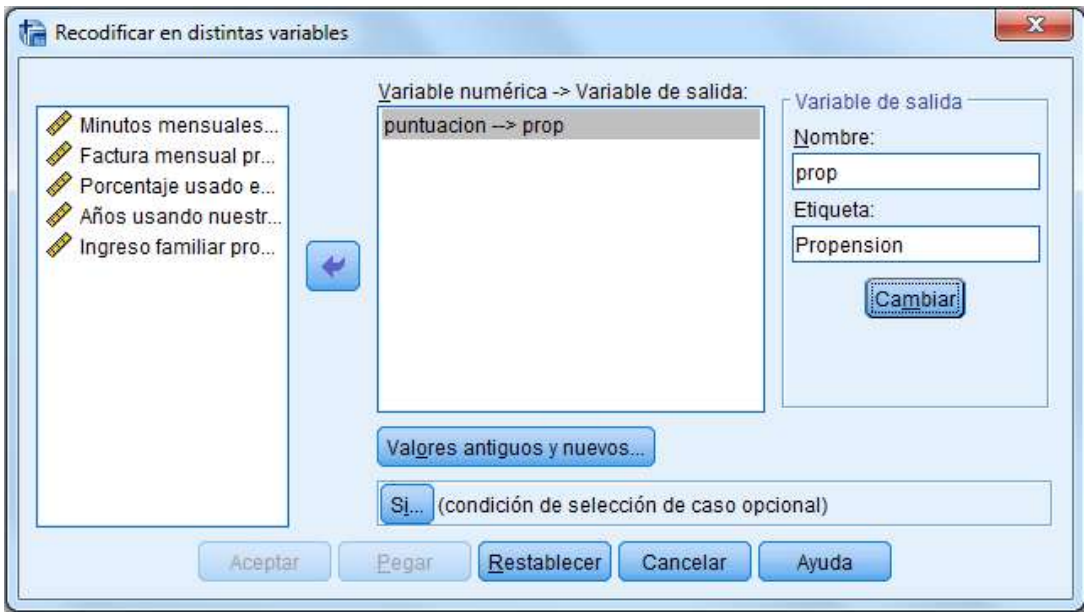
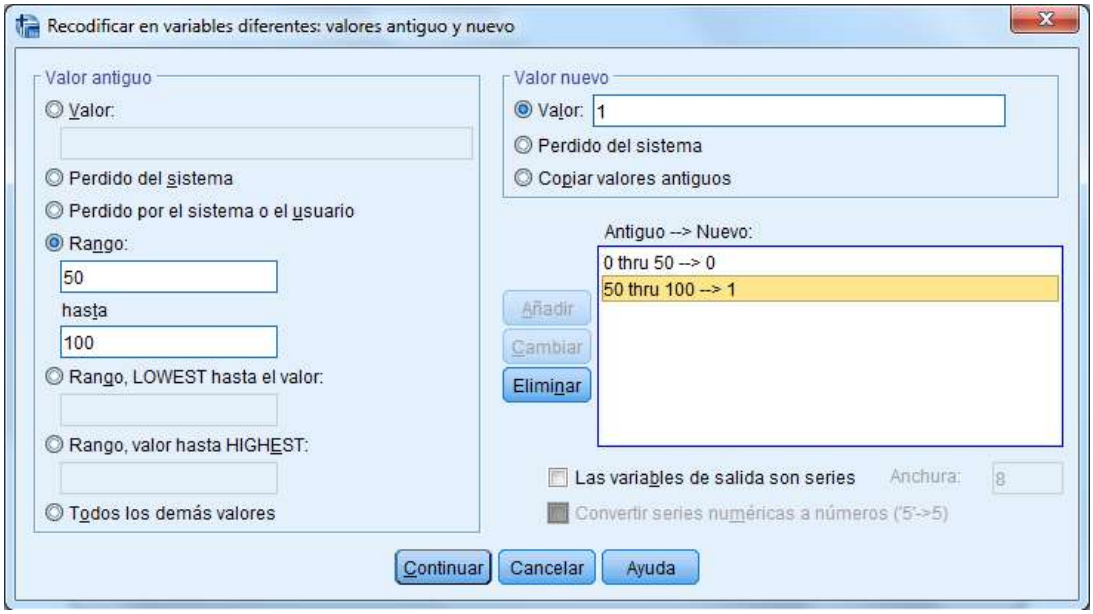


Figura 6-16: Asignación de valor a la variable "prop"



Una vez codificada la variable, se verifica la distribución de frecuencias de la nueva variable "prop" en SPSS mediante el uso de estadísticos descriptivos.

En SPSS:

- **Análisis de estadísticos descriptivos:**
  Analizar > Estadísticos descriptivos > Frecuencias

Tabla 6-9: Frecuencia de la variable prop

prop

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	.00	200	80.0	80.0	80.0
	1.00	50	20.0	20.0	100.0
	Total	250	100.0	100.0	

El valor muestral de la proporción de clientes "propensos a abandonar" es de 0.20.

Importante:

SPSS solo puede reproducir pruebas de dos colas, por lo que se modifica el alfa (y el nivel de confianza) para simular una prueba de este tipo, sabiendo que se tiene interés solo en cola derecha.

En SPSS:

- **Prueba T para una muestra:**
  Analizar > Comparar medias > Prueba T para 1 muestra

Figura 6-17: Prueba T para una muestra

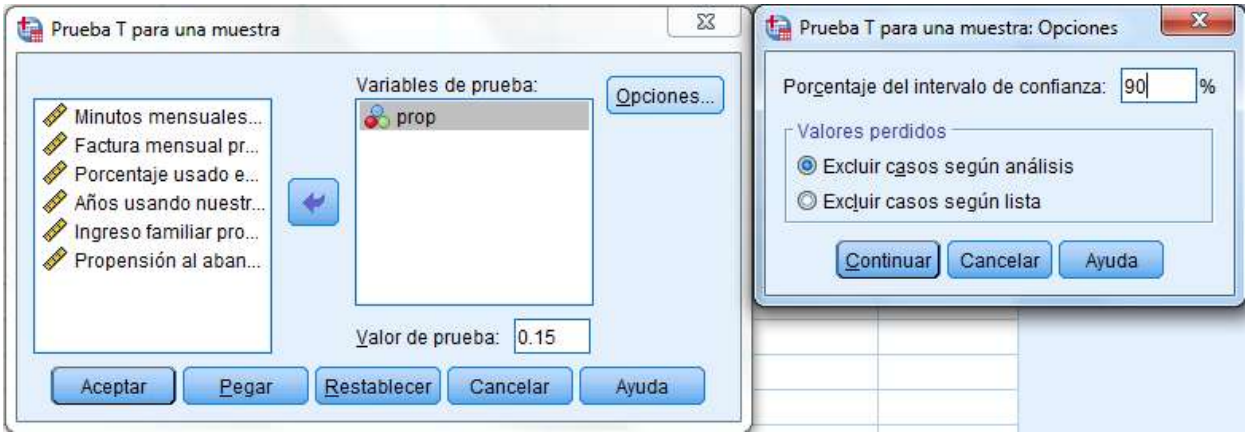


Tabla 6-10: Prueba para una muestra

Valor de prueba = 0.15

	t	gl	Sig. (bilateral)	Diferencia de medias	90% de intervalo de confianza de la diferencia	
					Inferior	Superior
prop	1.972	249	.050	.05000	.0081	.0919

Resultados:

Se verifica que el valor t se encuentra en cola de análisis. Al ser positivo, se encuentra en cola derecha. Se divide el p-value entre dos como arreglo para la prueba de una cola. Entonces, de esto se puede asegurar que la proporción de clientes propensos al abandono es mayor a 0.15. Esto implicaría que diseñar una estrategia agresiva de fidelización sí es conveniente.

Tabla 6-11: Interpretación de la significancia

Variable	p-value		Alfa	Interpretación
Proporción de clientes propensos al abandono	0.050 / 2	<	0.050	Rechazo H0

Lecturas recomendadas

- Anderson, D., Sweeney, D., y Williams, T. (2008). *Estadística para administración y economía* (10.ª ed., caps. 9 y 10). Cengage Learning Editores.

- Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4.ª ed., caps. 6 y 9). SAGE Publications.
- Levin, R., y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed. revisada, caps. 8 y 9). México: Pearson Educación de México; Prentice Hall.

7. Pruebas T para dos muestras independientes y pareadas

En este capítulo se busca que el alumno entienda y sea capaz de realizar un contraste de hipótesis para dos muestras, así como reconocer la pertinencia de cuándo usar un análisis para dos muestras independientes y dos muestras relacionadas. Se reforzarán aspectos vinculados al planteamiento de hipótesis estadísticas, nivel de medición de variables, entre otros.

Palabras clave:

- Pruebas T para dos muestras
- Muestras independientes
- Muestras relacionadas

Ejercicio I: Prueba de dos muestras independientes

Un docente de un curso de estadística siempre analiza el rendimiento de sus estudiantes según criterios de edad, escala de pago, número de cursos y sexo, para tener una idea de cuál es el perfil del estudiante que mejor rindió en el curso. Uno de estos análisis se centra en evaluar si existen diferencias significativas entre la nota promedio de los hombres y mujeres en el curso de estadística. La información con la que se cuenta es la siguiente:

Tabla 7-1: Notas promedio de Estadística

Notas promedio de Estadística por estudiantes												
11	16	11	14	09	15	15	11	13	17	15	16	13
H	M	H	M	H	M	H	H	H	M	M	M	H

**H: Hombre, M: Mujer.*

Ejercicio II: Prueba de dos muestras relacionadas

El mismo docente del curso de Estadística quiere comparar el desempeño de sus estudiantes con otros cursos afines en donde también se desempeña como docente. Para ello, recurre a sus notas promedio del ciclo anterior.

Tabla 7-2: Notas promedio de Estadística vs. curso afín

Notas promedio de Estadística y curso afín de los estudiantes en el último ciclo													
Estadística	11	16	11	14	09	15	15	11	13	17	15	16	13
Curso afín	16	11	14	17	09	18	16	15	16	18	17	15	15

Caso aplicado: Factores de empleo. Una mirada a los niveles de ingreso y satisfacción laboral en dos distritos de Lima Metropolitana

Para este ejercicio, emplee el archivo SPSS "BD_7".

El observatorio laboral CHAMBEA.pe es una institución fundada por egresados de la Facultad de Gestión y Alta Dirección, y tiene como objetivo proporcionar información actualizada y confiable sobre la dinámica y evolución del mercado laboral.

En el marco del presente estudio, se recogió información muestral de cuatro industrias (servicios, agroalimentación, industrial y formación) en dos distritos representativos de la clase media de Lima Metropolitana. La selección se basó en que se cuenta con información de que la población de alumnos de la Facultad de Gestión y Alta Dirección se desenvuelve en estos sectores tras su egreso.

El objetivo del estudio es caracterizar los principales factores involucrados en el comportamiento del mercado laboral, con especial énfasis en los niveles de ingreso y satisfacción laboral percibidos, además de otros elementos de interés como el nivel de especialización (expresado por los años de estudio), y la existencia de brechas de género.

Considerando los conocimientos obtenidos en el curso de Métodos de Investigación Cuantitativa, Ud. ha sido contratado por el observatorio para realizar el procesamiento y análisis estadístico de los datos, lo cual incluye la transformación de la base de datos "BD_7.sav", la creación de las variables que considere pertinentes, la validación de la muestra y el uso de las pruebas necesarias para cumplir con los

objetivos establecidos por el estudio. De acuerdo con el análisis del caso, responda lo siguiente:

a) De acuerdo con la Encuesta Nacional de Hogares (ENAHO), los distritos elegidos para el análisis son muy parecidos entre sí. Por esto, la muestra seleccionada debe representar dicha homogeneidad. ¿Existen diferencias significativas en los promedios de edad y años de estudio, de acuerdo con el distrito de residencia?

b) Según el Instituto Nacional de Estadística e Informática (INEI), "al año 2016, las mujeres ganan en promedio 29.2 % menos que sus pares masculinos", siendo una de las principales razones de esto la disparidad en los años de estudio entre ambos. ¿La muestra es representativa de esta brecha en años de estudio y salario?

c) Las condiciones de empleo entre industrias suelen verse afectadas por el nivel de especialización educativa en el estudio, por lo que decide incluir un acápite especial. En este evaluará si es que, para los profesionales de cada sector, el promedio de tiempo de estudio es mayor a 13 años.

d) Otro aspecto fundamental del estudio es la percepción de la satisfacción laboral de las personas encuestadas. En este sentido, se le pide evaluar si existen diferencias significativas en la proporción de trabajadores satisfechos con su empleo actual, de acuerdo con el sexo de la persona.

e) Por último, existen múltiples estudios que relacionan el nivel de ingresos que recibe un trabajador y la satisfacción laboral que percibe en su empleo. En el presente estudio se desea evaluar si —efectivamente— existen diferencias significativas en el salario entre las personas satisfechas con su empleo actual y aquellas que mostraron una opinión negativa o neutral sobre su satisfacción. Utilizar un nivel de significancia de 0.10 para esta prueba.

Solucionario

Ejercicio I: Prueba de dos muestras independientes

Se requiere adecuar la base de datos, para lo cual se ordenan los datos de forma vertical de la siguiente manera.

Tabla 7-3: Notas promedio de Estadística por sexo

HOMBRES	MUJERES
11	16
11	14
9	15
15	17
11	15
13	16
13	

En Excel, se utiliza la herramienta de análisis de datos y se selecciona "prueba T para dos muestras"; mientras que en SPSS se usarían los siguientes comandos.

En SPSS:

- **Prueba T para muestras independientes:**
  Analizar > Comparar medias > Prueba T para muestras independientes.

Del análisis en Excel se tienen los siguientes resultados. A un 95 % de confianza, sí existen diferencias significativas en las notas promedio de Estadística de los hombres y mujeres.

Tabla 7-4: Usando prueba T para hombres y mujeres

	HOMBRES	MUJERES
Media	**11.9**	**15.5**
Varianza	3.8	1.1
Observaciones	7	6
Varianza agrupada	2.58	
Diferencia hipotética de las medias	0.00	
Grados de libertad	11.00	
Estadístico t	-4.08	
P(T<=t) una cola	0.00	
Valor crítico de t (una cola)	1.80	
P(T<=t) dos colas	**0.0018**	
Valor crítico de t (dos colas)	**2.20**	

Ejercicio II: Prueba de dos muestras relacionadas

Se adecua la base de datos de forma vertical como se muestra a continuación.

Tabla 7-5: Notas promedio de Estadística vs. curso afín

Estadística	Curso afín
11	16
16	11
11	14
14	17
9	9
15	18
15	16
11	15
13	16
17	18
15	17
16	15
13	15

Se utiliza la herramienta de análisis de datos y se selecciona "prueba T para dos muestras emparejadas"; mientras que en SPSS se usarían los siguientes comandos.

En SPSS:

- **Prueba T para muestras dependientes:**
  Analizar > Comparar medias > Prueba T para muestras relacionadas.

Tabla 7-6: Usando prueba T para Estadística y curso afín

	Estadística	Curso afín
Media	13.54	15.15
Varianza	5.94	6.81
Observaciones	13	13
Coeficiente de correlación de Pearson	0.484	
Diferencia hipotética de las medias	0	
Grados de libertad	12	
Estadístico t	-2.2689	
P(T<=t) una cola	0.0213	
Valor crítico de t (una cola)	1.7823	
P(T<=t) dos colas	**0.0425**	
Valor crítico de t (dos colas)	**2.1788**	

Del análisis en Excel se tienen los siguientes resultados. A un 95 % de confianza, sí existen diferencias significativas en las notas promedio de Estadística y el curso afín para cada uno de los estudiantes en el último ciclo.

Caso aplicado: Factores de empleo. Una mirada a los niveles de ingreso y satisfacción laboral en dos distritos de Lima Metropolitana

Para este ejercicio, emplee el archivo SPSS "SOL_7".

a) De acuerdo con la Encuesta Nacional de Hogares (ENAHO), los distritos elegidos para el análisis son muy parecidos entre sí. Por esto, la muestra seleccionada debe representar dicha homogeneidad. ¿Existen diferencias significativas en los promedios de edad y años de estudio, de acuerdo con el distrito de residencia?

Para este ejercicio, se pide analizar las variables: edad en años y años de educación, por lo que primero ambas pasarán por prueba de normalidad para luego realizar la prueba T según corresponda.

En SPSS:

- **Prueba de normalidad:**
  Analizar > Estadísticos descriptivos > Explorar > Ingresar variables > Gráficos > *Check* en gráficos de normalidad con pruebas > Continuar > Aceptar.
- **Prueba T para muestras independientes:**
  Analizar > Comparar medias > Prueba T para muestras independientes.

Distrito de residencia

Pruebas de normalidad

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Distrito de residencia	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Edad en años	Distrito 1	.099	45	.200*	.947	45	.040
	Distrito 2	.111	44	.200*	.946	44	.038
Años de educación	Distrito 1	.093	45	.200*	.965	45	.194
	Distrito 2	.121	44	.112	.965	44	.201

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Prueba T

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
									95% de intervalo de confianza de la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	Inferior	Superior
Edad en años	Se asumen varianzas iguales	.659	.419	-1.216	87	.227	-4.461	3.670	-11.756	2.834
	No se asumen varianzas iguales			-1.214	86.100	.228	-4.461	3.673	-11.764	2.841
Años de educación	Se asumen varianzas iguales	1.133	.290	1.313	87	.193	.938	.715	-.482	2.359
	No se asumen varianzas iguales			1.315	85.735	.192	.938	.714	-.480	2.357

Prueba de medias de dos muestras independientes (dos colas):

Hipótesis:

H0: u-edad (distrito1) = u-edad (distrito2) = u-años-educ (distrito2)

H0: u-años-educ (distrito1)

H1: u-edad (distrito1) ≠ u-edad (distrito2) ≠ u-años-educ (distrito2)

H1: u-años-educ (distrito1)

Resultados:

P-value(edad): 0.227 > 0.050 → No se rechaza H0

P-value(años-educ): 0.193 > 0.050 → No se rechaza H0

Interpretación:

Los distritos no presentan diferencias significativas en relación con el promedio de edad y años de educación de los encuestados.

b) Según el Instituto Nacional de Estadística e Informática (INEI), "al año 2016, las mujeres ganan en promedio 29.2 % menos que sus pares masculinos", siendo una de las principales razones de esto la disparidad en los años de estudio entre ambos. ¿La muestra es representativa de esta brecha en años de estudio y salario?

Para este ejercicio, se pide analizar las variables: años de educación, salario laboral e ln_ingresos, por lo que estas primero pasarán por prueba de normalidad para luego realizar la prueba T según corresponda.

En SPSS:

- **Prueba de normalidad:**
  Analizar > Estadísticos descriptivos > Explorar > Ingresar variables > Gráficos > *Check* en gráficos de normalidad con pruebas > Continuar > Aceptar.
- **Prueba T para muestras independientes:**
  Analizar > Comparar medias > Prueba T para muestras independientes.

Género

Pruebas de normalidad

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Género	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Años de educación	Hombre	.099	45	.200*	.949	45	.048
	Mujer	.112	44	.198	.980	44	.619
Salario laboral	Hombre	.227	45	.000	.671	45	.000
	Mujer	.226	44	.000	.649	44	.000
ln_ingresos	Hombre	.078	45	.200*	.979	45	.561
	Mujer	.072	44	.200*	.957	44	.097

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Prueba T

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
									95% de intervalo de confianza de la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	Inferior	Superior
Años de educación	Se asumen varianzas iguales	2.047	.156	-.008	87	.994	-.006	.722	-1.440	1.429
	No se asumen varianzas iguales			-.008	84.803	.994	-.006	.720	-1.438	1.427
ln_ingresos	Se asumen varianzas iguales	1.026	.314	.570	87	.570	.09730	.17081	-.24220	.43681
	No se asumen varianzas iguales			.571	85.834	.570	.09730	.17054	-.24173	.43634

Prueba de medias de dos muestras independientes (dos colas)

Hipótesis:

H0: u-años-educ (hombre) = u-años-educ (mujer) H0: u-ln-ingreso (hombre) = u-ln-ingreso (mujer)

H1: u-años-educ (hombre) ≠ u-años-educ (mujer) H1: u-ln-ingreso (hombre) ≠ u-ln-ingreso (mujer)

Resultados:

P-value(años-educ): 0.994 > 0.050 → No se rechaza H0

P-value(ln-ingreso): 0.570 > 0.050 → No se rechaza H0

Interpretación:

No se presentan diferencias significativas en relación con los años de educación y el nivel de ingresos, de acuerdo con el género del encuestado.

c) Las condiciones de empleo entre industrias suelen verse afectadas por el nivel de especialización educativa en el estudio, por lo que decide incluir un acápite especial. En este evaluará si es que, para los profesionales de cada sector, el promedio de tiempo de estudio es mayor a 13 años.

Para este ejercicio, se pide analizar las siguientes variables: años de educación y sector laboral, por lo que primero ambas pasarán por prueba de normalidad para luego realizar la prueba T según corresponda.

En SPSS:

- **Prueba de normalidad:**
  Analizar > Estadísticos descriptivos > Explorar > Ingresar variables > Gráficos > *Check* en gráficos de normalidad con pruebas > Continuar > Aceptar.
- **Prueba T para una muestra:**
  Analizar > Comparar medias > Prueba T para 1 muestra.

Sector Laboral

Pruebas de normalidad

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Sector Laboral	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Años de educación	Servicios	.121	24	.200*	.927	24	.083
	Agroalimentario	.115	37	.200*	.952	37	.113
	Industrial	.125	14	.200*	.949	14	.548
	Formación	.165	14	.200*	.948	14	.523

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Prueba T

Prueba para una muestra

		Valor de prueba = 13					
						90% de intervalo de confianza de la diferencia	
Sector Laboral		t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
Servicios	Años de educación	3.226	23	.004	2.250	1.05	3.45
Agroalimentario	Años de educación	1.862	36	.071	.946	.09	1.80
Industrial	Años de educación	.370	13	.717	.357	-1.35	2.06
Formación	Años de educación	1.194	13	.254	1.214	-.59	3.02

Prueba de medias de una muestra (una cola a la derecha)

Hipótesis:

H0: u-años-educ (servicios) ≤ 13H0: u-años-educ (agroalimentario) ≤ 13

H1: u-años-educ (servicios) > 13H1: u-años-educ (agroalimentario) > 13

H0: u-años-educ (industrial) ≤ 13 H0: u-años-educ (formación) ≤ 13

H1: u-años-educ (industrial) > 13 H1: u-años-educ (formación) > 13

Resultados:

P-value(servicios): 0.004 / 2 < 0.050 → Se rechaza H0

P-value(agroalimentario): 0.071 / 2 < 0.050 → Se rechaza H0

P-value(industrial): 0.717 / 2 > 0.050 → No se rechaza H0

P-value(formación): 0.254 / 2 > 0.050 → No se rechaza H0

Interpretación:

Solamente los trabajadores de los sectores de servicios y agroalimentario cuentan con un nivel de especialización promedio mayor a los 13 años.

d) Otro aspecto fundamental del estudio es la percepción de la satisfacción laboral de las personas encuestadas. En este sentido, se le pide evaluar si existen diferencias significativas en la proporción de trabajadores satisfechos con su empleo actual, de acuerdo con el sexo de la persona.

Para este ejercicio, se pide analizar la variable satisfacción laboral de toda la muestra, para lo cual primero se realiza un análisis de las frecuencias por categoría, para así luego realizar la prueba T según corresponda.

En SPSS:

- **Tablas cruzadas:**
  Analizar > Estadísticos descriptivos > Tablas cruzadas > Inserte las variables a analizar (fila y columna) > Aceptar.
- **Prueba T para muestras independientes:**
  Analizar > Comparar medias > Prueba T para muestras independientes.

Tabla cruzada Satisfacción laboral*satisfecho

Recuento

		satisfecho		
		.00	1.00	Total
Satisfacción laboral	Muy insatisfecho	24	0	24
	Algo insatisfecho	16	0	16
	Neutral	21	0	21
	Algo satisfecho	0	18	18
	Muy satisfecho	0	10	10
Total		61	28	89

Prueba T

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
									95% de intervalo de confianza de la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	Inferior	Superior
satisfecho	Se asumen varianzas iguales	6.500	.013	1.296	87	.199	.12778	.09863	-.06826	.32381
	No se asumen varianzas iguales			1.297	86.306	.198	.12778	.09850	-.06803	.32358

Prueba de proporciones dos muestras independientes (dos colas)

Hipótesis:

H0: π-satisfaccion (hombre) = π-satisfaccion (mujer)

H1: π-satisfaccion (hombre) ≠ π-satisfaccion (mujer)

Resultados:

p-value(satisfaccion): 0.198 > 0.050 → No se rechaza H0

Interpretación:

No se presentan diferencias significativas en relación con la satisfacción laboral, de acuerdo con el género del encuestado.

e) Por último, existen múltiples estudios que relacionan el nivel de ingresos que recibe un trabajador y la satisfacción laboral que percibe en su empleo. En el presente estudio se desea evaluar si —efectivamente— existen diferencias significativas en el salario entre las personas satisfechas con su empleo actual y aquellas que mostraron una opinión negativa o neutral sobre su satisfacción. Utilizar un nivel de significancia de 0.10 para esta prueba.

Para este ejercicio, se pide analizar las siguientes variables: salario laboral e ln_ingresos, por lo que estas primero pasarán por prueba de normalidad para luego realizar la prueba T según corresponda.

En SPSS:

- **Prueba de normalidad:**
  Analizar > Estadísticos descriptivos > Explorar > Ingresar variables > Gráficos > *Check* en gráficos de normalidad con pruebas > Continuar > Aceptar.
- **Prueba T para muestras independientes:**
  Analizar > Comparar medias > Prueba T para muestras independientes.

satisfecho

Pruebas de normalidad

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	satisfecho	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Salario laboral	.00	.243	61	.000	.590	61	.000
	1.00	.234	28	.000	.763	28	.000
ln_ingresos	.00	.082	61	.200*	.960	61	.046
	1.00	.092	28	.200*	.978	28	.805

*. Esto es un límite inferior de la significación verdadera.

a. Corrección de significación de Lilliefors

Prueba T

Prueba de muestras independientes

		Prueba de Levene de igualdad de varianzas		prueba t para la igualdad de medias						
									90% de intervalo de confianza de la diferencia	
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	Inferior	Superior
ln_ingresos	Se asumen varianzas iguales	.052	.820	-1.839	87	.069	-.33250	.18077	-.63305	-.03196
	No se asumen varianzas iguales			-1.789	49.151	.080	-.33250	.18584	-.64405	-.02095

Prueba de medias de dos muestras independientes (dos colas)

Hipótesis:

H0: u-ln-ingresos (no satisfechos) = u-ln-ingresos (satisfechos)

H1: u-ln-ingresos (no satisfechos) ≠ u-ln-ingresos (satisfechos)

Resultados:

p-value(ln-ingresos): 0.069 < 0.100 → Se rechaza H0

Interpretación:

Existe una diferencia significativa en el nivel de ingresos entre los trabajadores satisfechos con su empleo actual y los que no.

Lecturas recomendadas

- Anderson, D., Sweeney, D., y Williams, T. (2008). *Estadística para administración y economía* (10.ª ed., caps. 9 y 10). Cengage Learning Editores.
- Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4.ª ed., caps. 6 y 9). SAGE Publications.
- Levin, R., y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed. revisada, caps. 8 y 9). México: Pearson Educación de México; Prentice Hall.

8. Análisis de varianza (ANOVA) de un factor[12]

El objetivo de este capítulo es identificar variables independientes importantes dentro de un estudio estadístico con el final de determinar cómo interactúan entre sí y afectan la respuesta. El modelo que se usará es la distribución de probabilidad F.

Palabras clave:

- ANOVA de un factor

[12] Los ejercicios presentados a continuación asumen el cumplimiento de los supuestos de la metodología ANOVA: independencia entre grupos, homocedasticidad y distribución normal. Por esa razón, al aplicarse en datos reales, los supuestos señalados deben verificarse previamente antes de la aplicación de la metodología.

- Varianza
- Distribución de probabilidad F
- Grados de libertad

Ejercicio I: Solicitudes de baja de 25 grupos posterior a la evaluación de *marketing*

Para este ejercicio, emplee el archivo en Excel "BD_8".

Se quiere evaluar la eficacia de una campaña de *mailing* publicitario de una tienda de *retail* a partir de la tasa mensual de solicitudes de baja del servicio de *mailing* de esta compañía. Para ello, se seleccionan al azar 25 conjuntos de 1000 clientes cada uno y estos, a su vez, son distribuidos aleatoriamente en 5 grupos.

Al primero se le envía publicidad 5 veces a la semana, al segundo 3 veces a la semana, al tercero 1 vez por semana, al cuarto 2 veces al mes, y al quinto 1 vez al mes. Los resultados de la tasa mensual de las solicitudes de baja de los 25 grupos al finalizar la evaluación de *marketing* los podrá encontrar en la base de datos proporcionada para este ejercicio.

Determine el planteamiento de las hipótesis nula y alternativa para la investigación y analice si se encuentran diferencias significativas entre los resultados de las solicitudes de baja del servicio para cada grupo a un nivel de confianza del 95 %.

Ejercicio II: Evaluación de estrategias de penetración de mercado

Para este ejercicio, emplee el archivo en Excel "BD_8".

Una empresa desea evaluar sus estrategias de penetración de mercado. Para ello, cuenta con los volúmenes de venta mensual en dólares asociados a los tres diferentes medios por los que publicita sus productos (radio, televisión y redes sociales).

A un nivel de confianza del 97 %, ¿se encuentran diferencias estadísticamente significativas en los volúmenes de venta asociados a las distintas estrategias de venta?

Caso aplicado: Examen de Certificación en la empresa MIC

Para este ejercicio, emplee el archivo en SPSS "BD_8".

El director ejecutivo de la empresa "MIC" desea capacitar a su personal profesional en nuevos enfoques de gestión empresarial debido a que deben rendir un examen de certificación, el cual les permita obtener los estándares necesarios solicitados por el grupo de inversionistas mayoritarios. Para esto, envía a parte de sus empleados a dos

universidades: el primer grupo es enviado a la Universidad DF y el otro grupo los manda a capacitarse a la Universidad EG. El director desea comparar los resultados del examen de certificación luego de ser capacitados ambos grupos en sus respectivas universidades, ya que esto le permitirá decidir si una universidad es mejor que la otra para los fines de capacitación del resto de sus empleados rumbo al examen de certificación solicitado. A usted, como analista de la empresa, se le solicita realizar un análisis de los resultados obtenidos en el examen de certificación y verificar si existe una diferencia significativa en los promedios obtenidos por las personas capacitadas en las universidades escogidas por el director.

Un aspecto importante para el director de la empresa es el estado de salud de sus profesionales, por lo que contrata a un doctor investigador con el objeto de evaluar e implementar una dieta a 16 trabajadores, los cuales tienen un alto índice de faltas por temas de enfermedades cardiovasculares y sanguíneas. El doctor plantea una dieta con el fin de evaluar su efecto sobre el peso y los niveles de triglicéridos de sus pacientes, por lo que aplicará esta dieta a los 16 trabajadores durante 6 meses. Se le solicita a usted, como especialista, evaluar los resultados de los indicadores antes y después del tratamiento para medir el efecto (significativo o no) que tiene la dieta sobre los trabajadores. ¿A qué conclusión llega con los resultados obtenidos?

Solucionario

Ejercicio I: Solicitudes de baja de 25 grupos posterior a la evaluación de *marketing*

Para este ejercicio, emplee el archivo en Excel "SOL_8".

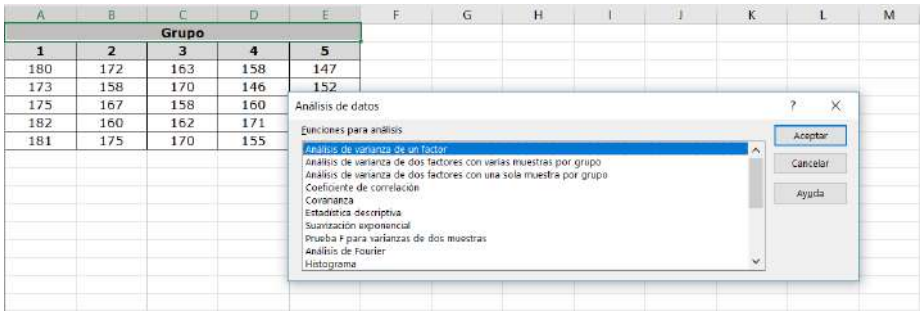
(Continúa del enunciado del Ejercicio 1 del Capítulo 8)

Para responder a lo solicitado, se debe emplear el análisis de varianza (ANOVA) de un factor, ya que la finalidad es contrastar estadísticamente si el factor "frecuencia de publicidad" influye en las solicitudes de baja. Para esto, en primer lugar, es necesario plantear la hipótesis nula y alternativa para este análisis:

H_0: Las medias del número de bajas de los cinco grupos son iguales, por lo que el factor no tiene un efecto significativo sobre el número de bajas.
H_1: Las medias del número de bajas de los cinco grupos no son iguales, por lo que el factor tiene un efecto significativo sobre el número de bajas.

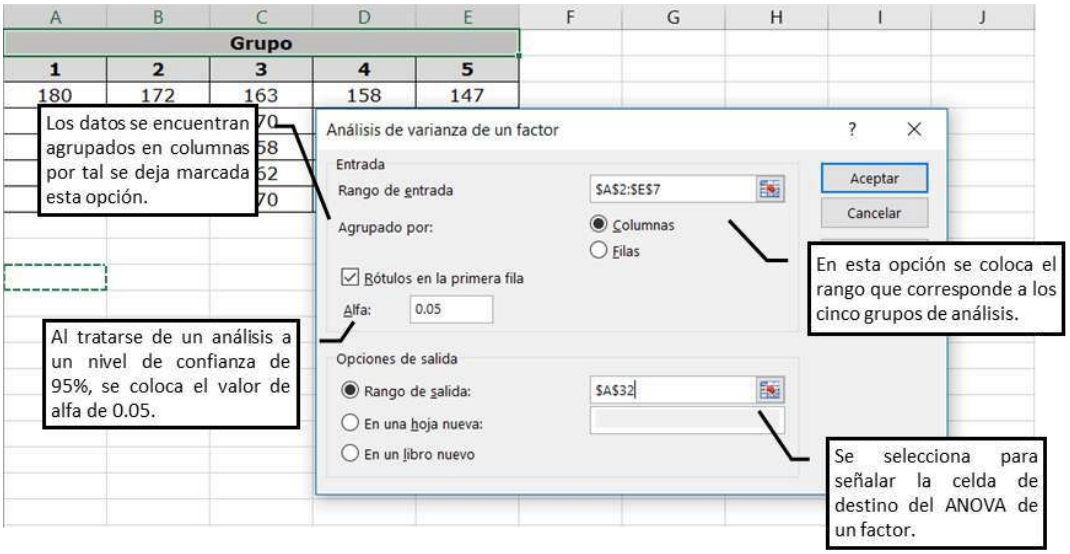
A continuación, se aplicará la herramienta "Análisis de datos" como se ha visto en unidades anteriores, pero en este caso se seleccionará la opción "Análisis de varianza de un factor" y se dará clic en "Aceptar".

Figura 8-1: Análisis de datos - Análisis de varianza de un factor



Posteriormente, se debe seleccionar el "Rango de entrada", que son las columnas que corresponden a los cinco grupos. Luego se indica el alfa (nivel de significancia), que en este caso es 0.05. Asimismo, se indica el "Rango de salida", como se presenta a continuación:

Figura 8-2: Análisis de varianza de un factor en Excel



Se obtiene el siguiente resultado:

Figura 8-3: Interpretación de ANOVA en Excel

Análisis de varianza de un factor

RESUMEN

Grupos	*Cuenta*	*Suma*	*Promedio*	*Varianza*
1	5	891	178.20	15.70
2	5	832	166.40	54.30
3	5	823	164.60	27.80
4	5	790	158.00	81.50
5	5	757	151.40	44.30

ANÁLISIS DE VARIANZA

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
Entre grupos	2010.64	4	502.66	11.24	0.00	2.87
Dentro de los grupos	894.4	20	44.72			
Total	2905.04	24				

Como puede verse, para este caso el $F_{estadístico}$ es mayor que el $F_{crítico}$; por tal razón se rechaza la hipótesis nula que indica que no existe un efecto de la frecuencia de la publicidad en las bajas del servicio.

Ejercicio II: Evaluación de estrategias de penetración de mercado

Para este ejercicio, emplee el archivo Excel "SOL_8".

(Continúa del enunciado del Ejercicio 2 del Capítulo 8)

Al tratarse de un ANOVA de un factor, se plantearán las hipótesis de investigación para iniciar con el análisis.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

H_0: Las tres medias son iguales; cualquier diferencia en los volúmenes de venta es producto del azar.

$$H_1: \mu_1 \neq \mu_2 \neq \mu_3$$

H_1: Las tres medias o alguna de ellas son distintas, por lo que la publicidad por radio, televisión y redes sociales tiene un efecto diferenciador en los volúmenes de venta.

Una vez planteadas las hipótesis de investigación, se ejecuta la herramienta "Análisis de datos" y se selecciona la opción ANOVA de un factor. Se toma el arreglo de datos proporcionado en la base de datos para este ejercicio y, a diferencia del ejercicio anterior, se selecciona la opción "Agrupado por" en filas, además el alfa (nivel de significancia) a un nivel de confianza de 97 % será de 0.03. Se obtienen los siguientes resultados:

Figura 8-4: Interpretación de ANOVA en Excel

Análisis de varianza de un factor

RESUMEN

Grupos	*Cuenta*	*Suma*	*Promedio*	*Varianza*
Radio	12	1388	115.67	3675.52
Televisión	12	1385	115.42	3588.63
Redes sociales	12	1154	96.17	2796.15

ANÁLISIS DE VARIANZA

Origen de las variaciones	*Suma de cuadrados*	*Grados de libertad*	*Promedio de los cuadrados*	*F*	*Probabilidad*	*Valor crítico para F*
Entre grupos	3003.50	2	1501.75	0.45	0.64	3.91
Dentro de los grupos	110663.25	33	3353.43			
Total	113666.75	35				

Es así que, para este caso, el $F_{estadístico}$ es menor que el $F_{crítico}$ por tal no existe evidencia estadística para rechazar la hipótesis nula. Esto quiere decir que no hay diferencias en publicitar en cualquiera de los tres medios.

Caso aplicado: Examen de certificación en la empresa MIC

Para este ejercicio, emplee el archivo SPSS "SOL_8".

(Continúa del enunciado del caso aplicado del Capítulo 8)

Solucionario parte a):

Paso 1: Planteamiento de hipótesis

Hipótesis del investigador:

El promedio de las calificaciones en el examen de certificación de los empleados en la Universidad DF es mayor que el de los empleados en la Universidad EG.

H0= No existe una diferencia significativa entre la media de calificaciones del grupo DF y la media de calificaciones del grupo EG.

H1= Existe una diferencia significativa entre la media de calificaciones del grupo DF y la media de calificaciones del grupo EG.

Paso 2: Determinar el "alfa"

Alfa= 5 %

Paso 3: Elección de la prueba de hipótesis

		PRUEBAS NO PARAMÉTRICAS			PRUEBAS PARAMÉTRICAS
Variable Fija \ Variable Aleatoria		NOMINAL DICOTÓMICA	NOMINAL POLITÓMICA	ORDINAL	NUMÉRICA
Estudio Transversal Muestras Independientes	Un grupo	X^2 Bondad de Ajuste Binomial	X^2 Bondad de Ajuste	X^2 Bondad de Ajuste	T de Student (una muestra)
	Dos grupos	X^2 Bondad de Ajuste Corrección de Yates Test exacto de Fisher	X^2 de Homogeneidad	U Mann-Withney	T de Student (muestras Independientes)
	Más de dos grupos	X^2 Bondad de Ajuste	X^2 Bondad de Ajuste	H Kruskal-Wallis	ANOVA con un factor INTERsujetos
Estudio Longitudinal Muestras Relacionadas	Dos medidas	Mc Nemar	Q de Cochran	Wilcoxon	T de Student (muestras Relacionadas)
	Más de dos Medidas	Q de Cochran	Q de Cochran	Friedman	ANOVA para medidas repetidas (INTRAsujetos)

Estudio transversal, ya que se están analizando 2 grupos en un mismo momento y la variable aleatoria es una variable numérica (calificación).

Paso 4: Calcular el P-valor

Antes de calcular el valor t student, se tienen que corroborar dos supuestos:

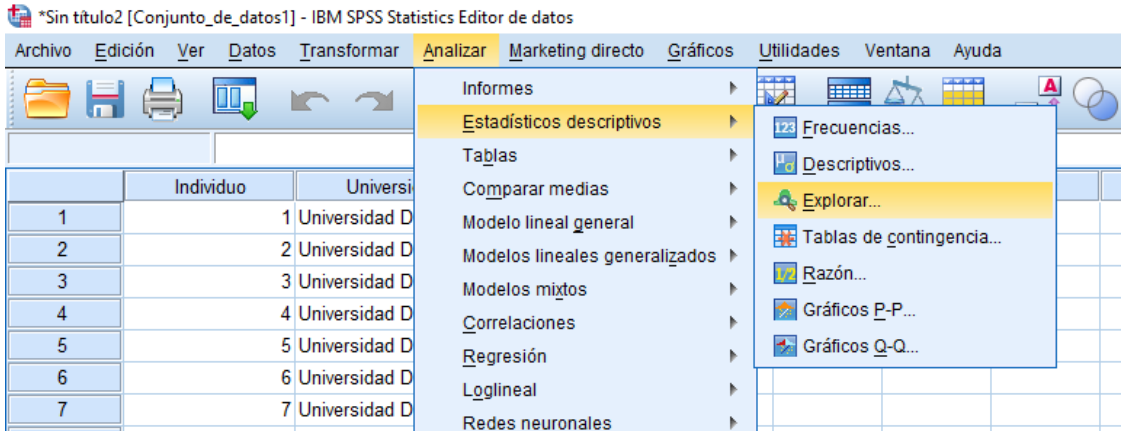
Normalidad: Se debe corroborar que la variable aleatoria en ambos grupos se distribuye normalmente. Para ello se utiliza la prueba de Kolmogorov-Smirnov (cuando las muestras son grandes n>30) o la prueba **Shapiro-Wilk** cuando el tamaño de la muestra es menor a 30 (n<30), el criterio para determinar si la variable aleatoria se distribuye normalmente es:

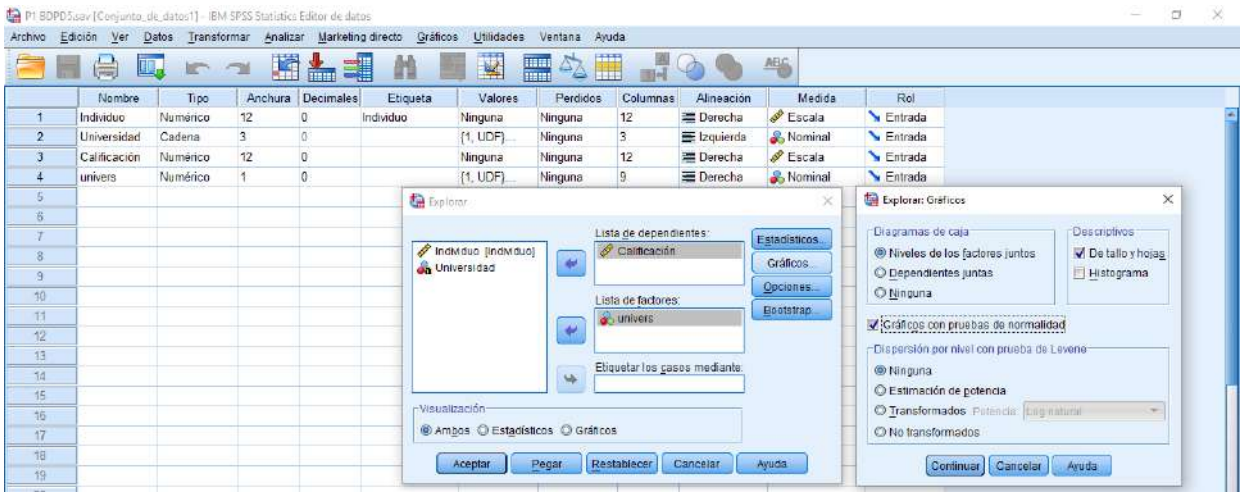
a. Si el p-valor =>alfa entonces se acepta H0=Los datos provienen de una distribución normal
b. Si el p-valor <alfa entonces se acepta H1=Los datos NO provienen de una distribución normal

Normalidad de Calificaciones

P-Valor (UEG) = 0.257 > 0.05
p-Valor (UDF) = 0.156 > 0.05

Conclusión: La variable calificación en ambos grupos se comporta normalmente

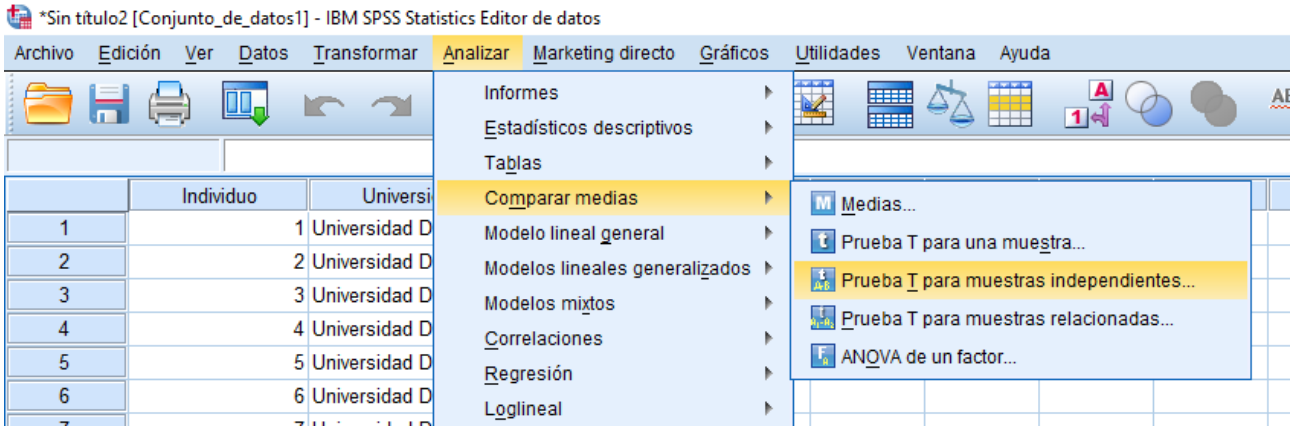




Verificar en opciones el nivel de confianza (95 %)

<u>Igualdad de varianza:</u> Se usa la prueba de Levene

P-valor => 0.05 entonces se acepta H0 = Las varianzas son iguales
P-valor < 0.05 entonces se acepta H1= Existe diferencia significativa entre las varianzas



Una vez identificada la variable para contrastar ("Calificación"), se definen los grupos, en este caso "1" y "2".

Prueba de muestras independientes

		Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	95 % Intervalo de confianza para la diferencia	
									Inferior	Superior
Calificación	Se han asumido varianzas iguales	,152	,699	,443	28	,661	2,200	4,962	-7,964	12,364

No se han asumido varianzas iguales			,443	27,475	,661	2,200	4,962	-7,973	12,373

P-valor=0.699 >0.05. Por lo tanto, se acepta H0, la cual indica que las varianzas son iguales.

Paso 5: Decisión estadística

El criterio para decidir es:

Si la probabilidad obtenida p-valor =<alfa entonces se rechaza H0

Si la probabilidad obtenida p-valor > alfa entonces se acepta H0

CALCULAR P-valor:

P-valor= 0.661 > 0.05

Por lo tanto, no existe una diferencia significativa entre la media de calificaciones del grupo DF y la media de calificaciones del grupo EG.

T student 2 muestras relacionadas

Un aspecto importante para el director de la empresa es el estado de salud de sus profesionales, por lo que contrata a un doctor investigador con el objeto de evaluar e implementar una dieta a 16 trabajadores, los cuales tienen un alto índice de faltas por temas de enfermedades cardiovasculares y sanguíneos. El doctor plantea una dieta con el fin de evaluar su efecto sobre el peso y los niveles de triglicéridos de sus pacientes, por lo que aplicará esta dieta a los 16 trabajadores durante 6 meses. Se le solicita a usted, como especialista, evaluar los resultados de los indicadores antes y después del tratamiento para medir el efecto (significativo o no) que tiene la dieta sobre los trabajadores. ¿A qué conclusión llega con los resultados obtenidos?

Hipótesis del investigador:

Existirá una diferencia significativa entre las medidas del peso antes de someterse al plan de dieta (pretest) y las medidas después de someterse a la dieta (posttest).

Paso 1: Planteamiento de hipótesis

H0= No hay diferencia significativa en las medidas del peso antes y después del tratamiento.

H1= Hay una diferencia significativa en las medidas del peso antes y después del tratamiento.

Paso 2: Definir el "alfa" (porcentaje de error que se está dispuesto a tomar para la prueba)

Paso 3: Elección de la prueba de hipótesis

Se utiliza en los estudios de tipo longitudinal, cuando se realizan medidas en dos momentos distintos (prueba antes y después), analizar el mismo grupo, analizar una variable numérica en el mismo grupo en dos momentos distintos. La variable aleatoria es el "Peso", la cual es una variable numérica, por lo que la prueba escogida es "T student para muestras relacionadas".

Paso 4: Calcular el P-valor

Antes de calcular el valor t student, se tienen que corroborar dos supuestos:

Primero se verifica que la variable peso, que es la variable de comparación, se comporta normalmente.

Normalidad: Se debe corroborar que la variable aleatoria en ambos grupos se distribuye normalmente. Para ello, se utiliza la prueba de Kolmogorov-Smirnov cuando las muestras son grandes (n>30) o la prueba **Shapiro-Wilk** cuando el tamaño de la muestra es menor a 30 (n<30). El criterio para determinar si la variable aleatoria se distribuye normalmente es:

a. Si el p-valor =>alfa entonces se acepta H0=Los datos provienen de una distribución normal
b. Si el p-valor <alfa entonces se acepta H1=Los datos NO provienen de una distribución normal

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
peso 0	,156	16	,200*	,938	16	,320
peso 4	,153	16	,200*	,938	16	,327

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

P-valor (peso antes)=0.320 > 0.05 entonces se acepta la H0, los datos provienen de una distribución formal.

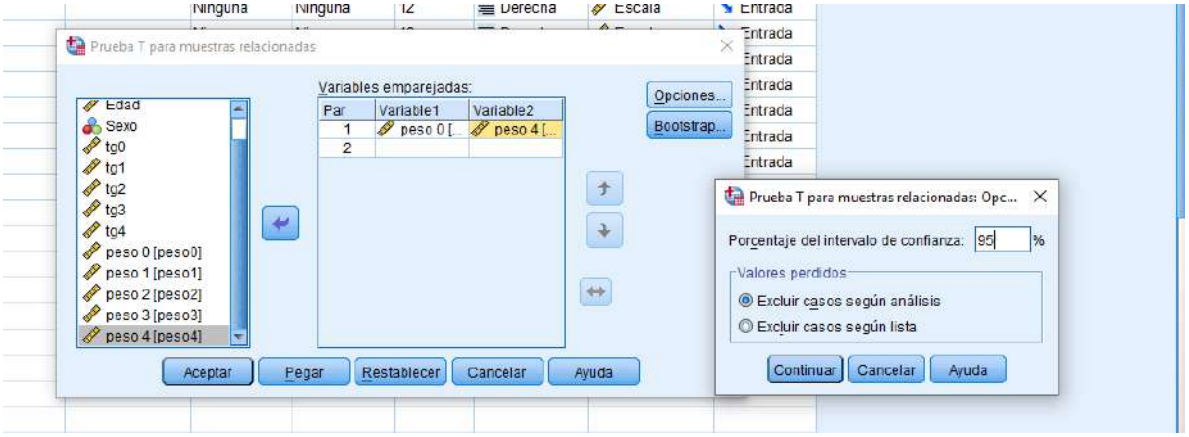
P-valor (peso después)= 0.327 > 0.05 entonces se acepta la H0, los datos provienen de una distribución normal

Los datos de la variable peso se comportan normalmente.

Paso 5: Decisión Estadística

Si la probabilidad obtenida p-valor =< alfa entonces se rechaza H0
Si la probabilidad obtenida p-valor > alfa entonces se acepta H0

P-valor = 0,000 < 0.05, por lo tanto se rechaza H0



Prueba de muestras relacionadas

	Diferencias relacionadas					t	gl	Sig. (bilateral)
	Media	Desviación típ.	Error típ. de la media	95 % Intervalo de confianza para la diferencia				
				Inferior	Superior			
Par 1 peso 0 - peso 4	6,563	2,476	,619	5,243	7,882	10,603	15	,000

Conclusión:

Hay una diferencia significativa en las medidas del peso antes y después del tratamiento. Por esa razón, se concluye que el tratamiento (la dieta) SÍ tiene efectos significativos sobre el peso de los pacientes.

Los pacientes en promedio bajaron su peso de 198.38 lbs a 191.81 lbs.

Lecturas recomendadas

- Levin, R., y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed. revisada; cap. 11, 11.4). México: Pearson Educación de México; Prentice Hall.
- Lind, D. A., Marchal, W. G., y Wathen, S. A. (2008). *Estadística aplicada a los negocios y la economía* (13.ª ed., cap. 12). México: McGraw-Hill Interamericana.

9. Estadísticos de asociación: Chi-cuadrado y Tau de Kendall

Esta sección inicia con la exploración de la relación entre dos variables cualitativas, ya sean estas nominales u ordinales. Dada la naturaleza de estas variables y la dificultad de realizar operaciones matemáticas con sus valores, se recurre al análisis de sus tablas de contingencia. Las pruebas que se utilizarán son Chi-cuadrado y Tau de Kendall (Tau-b o Tau-c).

Palabras clave:

- Tabla de contingencia (tabla cruzada)
- Chi-cuadrado
- Tau de Kendall
- Estadísticos de asociación

Ejercicio I: Relación entre variables cualitativas

Anualmente, el Ministerio de Educación (MINEDU) lleva a cabo un censo de todas las instituciones educativas a nivel nacional. Con la información recogida en ese censo, elabore y publique el "Padrón de instituciones educativas". Parte de los datos de ese padrón son presentados en el archivo "Padron_web.sav" y la siguiente tabla describe las variables presentes. Complete los tipos.

Tabla 9-1: Variables del Padrón de IE

VARIABLE	DESCRIPCIÓN	TIPO
NOMBRE	Nombre del colegio	
C_NIVEL	Nivel, en código numérico	
NIVEL	Nivel (inicial, primaria, secundaria)	
C_CATEGORIA	Categoría, en código numérico	
CATEGORIA	Categoría, según se tengan o no docentes para todos los grados. - Polidocente completo: aulas diferentes para todos los grados - Unidocente o Polidocente Multigrado: alumnos de más de un grado comparten aula	
C_GESTION	Tipo de gestión, en código numérico	
GESTION	Tipo de gestión del colegio. - Pública de gestión directa: gestionado por el sector público - Pública de gestión privada: colegio público con gestión privada - Privada: colegio privado	
ZONA	Zona (rural o urbana) en la que se encuentra el colegio	
TALUMNO	Total de estudiantes	
TSECCION	Total de secciones	

Realice las pruebas necesarias para respaldar o rechazar las siguientes dos afirmaciones:

- Los colegios con mayor nivel de gestión privada están más concentrados en los últimos años de la educación básica.
- La categoría de colegio está distribuida uniformemente entre las zonas rural y urbana.

Ejercicio II: Relación entre variables cuantitativas

En el contexto del ejercicio anterior, use las variables cuantitativas para lo siguiente:

- Elaborar un diagrama de dispersión entre variables TALUMNO y TSECCION para el nivel secundario. ¿Observa una relación lineal?
- Realizar el análisis de regresión lineal utilizando a TALUMNO y TSECCION para los niveles secundaria, primaria e inicial por separado.

Caso aplicado: Tiendas de conveniencia

Ha pasado medio año desde que usted ingresó al programa *trainee* de una empresa con presencia en varios países, que tiene, entre otras unidades de negocios, una cadena de tiendas de conveniencia (TdC). Gracias a los buenos resultados en su evaluación semestral, se le ha asignado para el segundo medio año un puesto en la Gerencia

Comercial de esa unidad de negocios en Europa (la posibilidad de rotar posiciones entre oficinas de diferentes países fue uno de los criterios que usted tomó en cuenta cuando decidió postular al programa).

Una vez en Europa (al menos por los siguientes seis meses), recibe un correo electrónico de bienvenida de la gerenta comercial, su nueva jefa. El mensaje, aparte de las indicaciones administrativas, menciona que tiene un importante tema que hablar con usted en su primer día de trabajo.

La mañana de su primer día de trabajo en Europa, lee en la sección negocios del periódico local que un grupo económico importante, basado en Estados Unidos, que principalmente posee 2 tipos de negocios *retail* (una cadena de TdC y una cadena de venta al menudeo de productos orgánicos) está interesado en ingresar al mercado europeo.

Ya en la reunión con su jefa, ella le menciona que el tema sobre el que tienen que hablar es sobre el interés del grupo económico americano en ingresar al mercado europeo. Usted afirma lo mencionado por su jefa; sin embargo, ella añade: "Lo que no han mencionado en la nota es que ellos han iniciado el registro de una marca en varios países de la Unión Europea. Además, han estado interesados en alquilar o comprar varios espacios que bien podrían usarse para implementar puntos *retail*. Lo sabemos porque algunos de estos puntos también nos interesaban". Usted consulta si saben para cuál de los dos negocios *retail* que posee ese grupo económico es la marca que han estado registrando y su jefa le responde que no y añade lo siguiente: "Tenemos que estar preparados; si es para su cadena de comida orgánica tendríamos más tiempo, porque por su actitud en otras regiones, una vez que inician en un mercado con una de estas cadenas, no introducen la otra hasta estar bien asentados. Hemos armado un comité de gerencia *ad hoc* para explorar las opciones de respuesta. El comité tendrá su primera reunión en dos días y nosotros, como Gerencia Comercial, somos parte de él y seremos de los primeros en presentar información relevante". Finalmente, le menciona que tiene excelentes referencias (de su anterior jefe) sobre los análisis cuantitativos que realizó en Perú, y por eso ha decidido encargarle los estudios iniciales. Lo que al comité le gustaría saber es lo siguiente:

- ¿Para cuál de las cadenas han iniciado con el registro de marcas en varios países?
- ¿El negocio de TdC del grupo económico americano se orienta al mismo segmento que el nuestro?

Dado el límite de tiempo, usted empieza su primer día laboral en Europa por averiguar con qué información cuenta la compañía y qué información pública podría ser relevante para el caso. Pasa la mañana entrevistando a personas y reuniendo toda la información posible. Al inicio de la tarde, ya ha logrado reunir la siguiente información:

- Una tesis de bachillerato de una universidad alemana sobre la situación de la regulación para la importación de productos orgánicos en cada país de la Unión Europea. Con este documento, pudo catalogar a cada país miembro de la Unión Europea como MUY REGULADO o POCO REGULADO. La tesis es de dos años atrás.
- Un estudio encargado por su empresa a Euromonitor (empresa internacional de investigación de mercados) un año atrás. Fue elaborado con el propósito de evaluar la situación general para el negocio de TdC en Europa de cara a una posible expansión. En este documento se clasificó a los países como CONVENIENTES o NO CONVENIENTES para la introducción de una cadena de este tipo. El análisis tomó en cuenta criterios como número de jugadores locales, perfiles de los consumidores, tamaños de los mercados, etc.
- Un informe interno para una ciudad en Estados Unidos en el que se menciona someramente a este competidor. Este informe contiene una división de la ciudad en 40 zonas; la clasificación de cada una de estas zonas fue llevada a cabo según el nivel socioeconómico (ALTO, MEDIO o BAJO) y la presencia de cadenas de tiendas, entre las que está el competidor en cuestión. Con los datos del informe usted logra clasificar a las mismas 40 zonas según el grado de presencia del competidor (si tiene presencia ALTA o BAJA según el número de tiendas por zona).
- La lista de países en los que el grupo económico en cuestión ha registrado la nueva marca.
- Gracias a una entrevista al jefe de segmentación de mercados, sabe que las TdC de su empresa se han caracterizado por mantener un formato pequeño y orientarse a niveles socioeconómicos altos, en los que se puede tener márgenes de ganancia mayores.

Si bien la información tiene entre uno y dos años de antigüedad, la situación no debe haber cambiado mucho (apunta este supuesto para comentarlo al exponer su trabajo). Construye dos bases de datos (archivos "Caso_Regulacion por país.sav" y "Caso_Locaciones USA.sav") y decide iniciar el análisis explorando las posibilidades de que la nueva marca de ese grupo económico sea para TdC o una cadena de productos orgánicos. Para esto:

1. sigue el siguiente razonamiento: "El ingreso será con una cadena de productos orgánicos. Entonces, los países de interés (en los que registró la marca) deben ser aquellos con poca regulación";
2. "En cambio, si el ingreso será con una cadena de TdC, los países de interés deberían guardar alguna relación con los resultados del estudio de Euromonitor".

Para contestar la pregunta sobre los segmentos objetivo de su empresa y la competencia, usted quiere determinar si la presencia alta o baja de tiendas de la

competencia en esa ciudad norteamericana tiene alguna relación con el nivel socioeconómico de las zonas.

Ya llegó el final del día y, antes de retirarse, envía los resultados encontrados a su jefa. Al día siguiente, temprano, se reúne con ella. La gerenta comercial le dice que está bastante contenta con ese análisis inicial y que comparte sus conclusiones. "Como opción de respuesta rápida tenemos las campañas de publicidad para reforzar nuestro valor de marca, pero en el comité hay gerentes que siempre han estado en contra de ese tipo de gastos. Por eso, quiero que determines si el impacto de la publicidad que hemos contratado ha sido económicamente positivo o no, y qué tanto lo ha sido o no. Yo estoy convencida de que ese impacto ha sido positivo".

Nuevamente, inicia su trabajo recolectando información, teniendo en mente que en esta oportunidad también debe preocuparse por cuantificar la relación entre publicidad y ventas. Luego de reunir la información disponible, ha elaborado un nuevo cuadro (archivo "Caso_Datos ventas.sav") con la siguiente información:

Tabla 9-2: Nuevas variables

Variable	Descripción
Semana	Número de semana del año, de los últimos dos años
Ventas	Venta promedio por semana por tienda (€)
TV	*Gross rating point* (GRP) para TV: cada punto es un 1 % del público objetivo alcanzado una vez. El valor indica el GRP en TV que tuvo la publicidad de su empresa esa semana
Radio	GRP para radio: cada punto es un 1 % del público objetivo alcanzado una vez. El valor indica el GRP en Radio que tuvo la publicidad de su empresa esa semana

Empieza por pensar en qué tipo de relaciones esperaría encontrar entre las variables. Luego:

3. ¿Qué información saca de una exploración de la relación entre variables? ¿Es lo que esperaba?

Después, continúa su trabajo:

4. cuantificando la dependencia de Ventas en TV y Radio, por separado.

Ahora piensa que sería importante tomar en cuenta otros factores relevantes para el incremento o disminución de las ventas. ¡A buscar más información!

Solucionario

Ejercicio I: Relación entre variables cualitativas

Para este ejercicio, emplee el archivo SPSS "SOL_9.1".

Cuadro de variables y sus tipos. Según estos tipos, se elegirán las pruebas para apoyar o rebatir las dos afirmaciones solicitadas.

Tabla 9-3: Tipo de variables del padrón de IE

VARIABLE	DESCRIPCIÓN	TIPO
NOMBRE	Nombre del colegio	Cualitativa nominal
C_NIVEL	Nivel, en código numérico	Cualitativa ordinal
NIVEL	Nivel (inicial, primaria, secundaria)	Cualitativa ordinal
C_CATEGORIA	Categoría, en código numérico	Cualitativa ordinal
CATEGORIA	Categoría, según se tengan o no docentes para todos los grados. - Polidocente completo: aulas diferentes para todos los grados - Unidocente o Polidocente Multigrado: alumnos de más de un grado comparten aula	Cualitativa nominal
C_GESTION	Tipo de gestión, en código numérico	Cualitativa ordinal
GESTION	Tipo de gestión del colegio. - Pública de gestión directa: gestionado por el sector público - Pública de gestión privada: colegio público con gestión privada - Privada: colegio privado	Cualitativa ordinal
ZONA	Zona (rural o urbana) en la que se encuentra el colegio	Cualitativa nominal
TALUMNO	Total de estudiantes	Cuantitativa
TSECCION	Total de secciones	Cuantitativa

Para la afirmación: "Los colegios con mayor nivel de gestión privada están más concentrados en los últimos años de la educación básica", debe explorarse la relación entre las variables NIVEL y GESTION. Ambas variables son cualitativas ordinales, por lo que se usarán las aplicaciones de la tabla de contingencia. Las pruebas que se usarán serán las Tau de Kendall. Para estas, se tiene que:

$H_0: No\ existe\ relación\ entre\ las\ variables$

$H_1: Existe\ relación\ entre\ las\ variables$

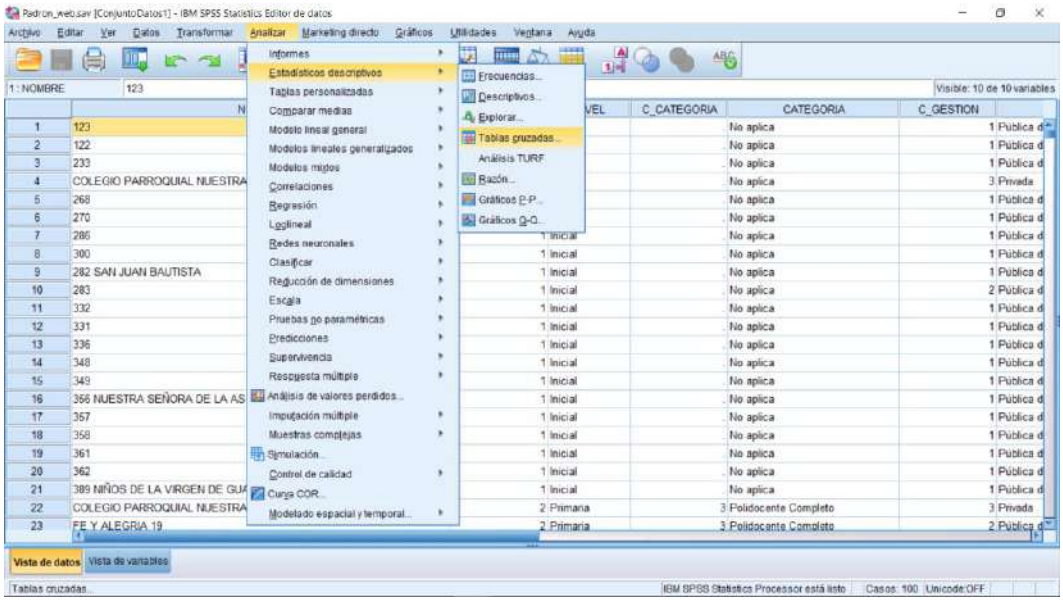
Para implementar estas pruebas, se hace uso del *software* SPSS. El archivo con la base de datos Padrón contiene las variables antes descritas adecuadamente codificadas.

Además, dado que se implementará una prueba para variables ordinales, en lugar de utilizar NIVEL y GESTION, se hará uso de las variables C_NIVEL y C_GESTION, pues ya se tienen los niveles con valores numéricos.

A continuación, se describe paso a paso el proceso para llevar a cabo esa prueba.

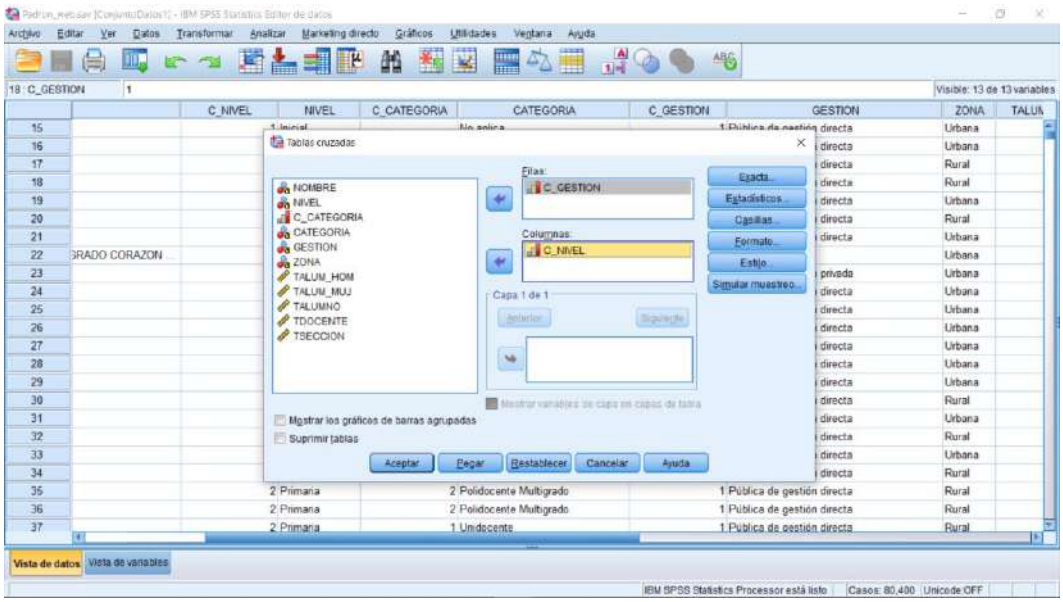
Paso 1: seleccionar análisis de tablas cruzadas.

Figura 9-1: Tablas cruzadas en SPSS



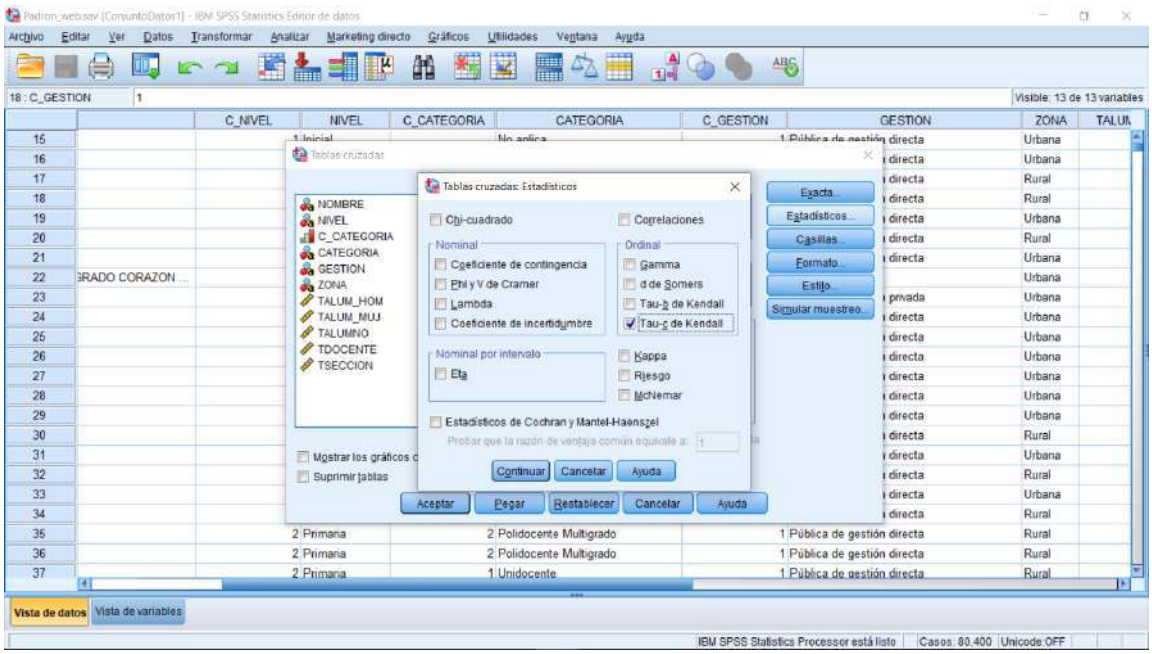
Paso 2: elegir las variables para la fila y columna de la tabla de contingencia.

Figura 9-2: Elección de variables para tablas cruzadas en SPSS



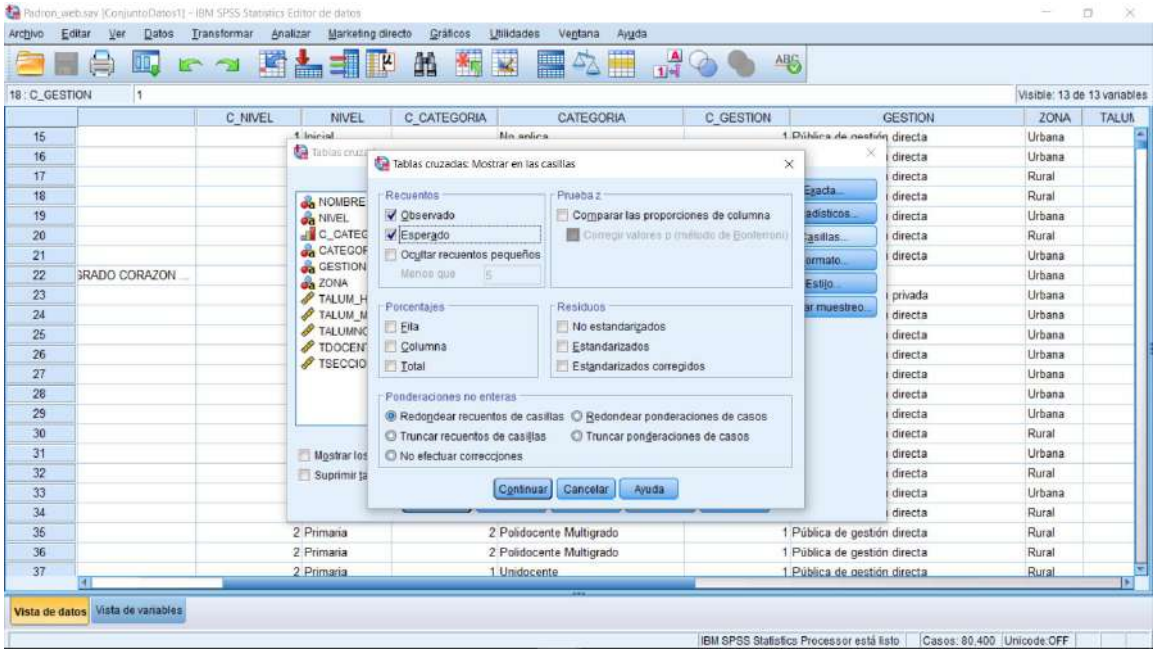
Paso 3: activar las pruebas que se quieren realizar, en este caso Tau-b y Tau-c de Kendall.

Figura 9-3: Tau de Kendall en SPSS



Paso 4: activar la opción "esperado" para visualizar datos esperados en cada casilla.

Figura 9-4: Activar valores esperados en SPSS



Luego de esos pasos, en la ventana de resultados se tendrán las siguientes tablas:

Tabla 9-4: Resumen de procesamiento de casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
C_GESTION * C_NIVEL	87721	100,0 %	0	0,0 %	87721	100,0 %

Tabla 9-5: Tabla cruzada C_GESTION*C_NIVEL

			C_NIVEL			Total
			1	2	3	
C_GESTION	1	Recuento	24339	29282	9064	62685
		Recuento esperado	25068,7	27293,2	10323,0	62685,0
	2	Recuento	250	330	307	887
		Recuento esperado	354,7	386,2	146,1	887,0
	3	Recuento	10492	8582	5075	24149
		Recuento esperado	9657,6	10514,6	3976,9	24149,0
Total		Recuento	35081	38194	14446	87721
		Recuento esperado	35081,0	38194,0	14446,0	87721,0

Tabla 9-6: Medidas simétricas

		Valor	Error estandarizado asintótico[a]	T aproximada[b]	Significación aproximada
Ordinal por ordinal	Tau-b de Kendall	,004	,003	1,313	,189
	Tau-c de Kendall	,003	,003	1,313	,189
N de casos válidos		87721			

a. No se presupone la hipótesis nula.

b. Utilización del error estándar asintótico que presupone la hipótesis nula.

Para evaluar la relación de dos variables ordinales, pueden utilizarse las pruebas Tau-b y Tau-c de Kendall. La primera de estas no es muy recomendable para casos en los que la tabla de contingencia no sea "cuadrada" (igual número de niveles para ambas variables).

El valor de los resultados en ambas pruebas es ligeramente diferente porque los métodos de cálculo también lo son. Al obtenerse un valor de 0,189 > 0.05 para el estadístico del nivel de significancia, se acepta la hipótesis nula que es "no hay relación estadísticamente significativa entre la Gestión (en términos de qué tan privada o pública es) y el Nivel (inicial, primaria o secundaria)". Por esta razón, la afirmación hecha es FALSA.

Para la segunda afirmación, "La categoría de colegio está distribuida uniformemente entre las zonas rural y urbana", se estudia la relación de las variables cualitativas nominales CATEGORIA y ZONA. Para esto, se utiliza la prueba Chi-cuadrado en la que:

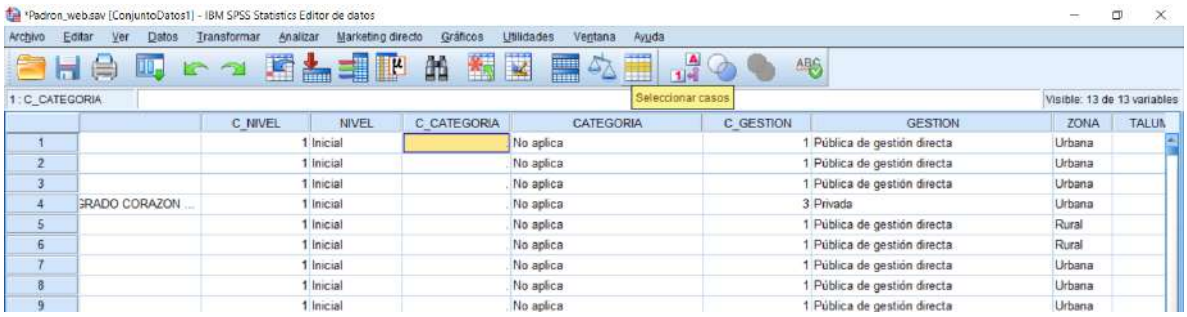
H_0: *No existe relación entre las variables*

H_1: *Existe relación entre las variables*

Cabe resaltar que este análisis se realizará solo a las observaciones de colegios del nivel primaria (puesto que los otros niveles no tienen un valor para CATEGORIA). Esto puede lograrse filtrando solo los valores de C_CATEGORIA que no sean "no observados" (marcados como "." en la base de datos). A continuación, se presenta el procedimiento paso a paso:

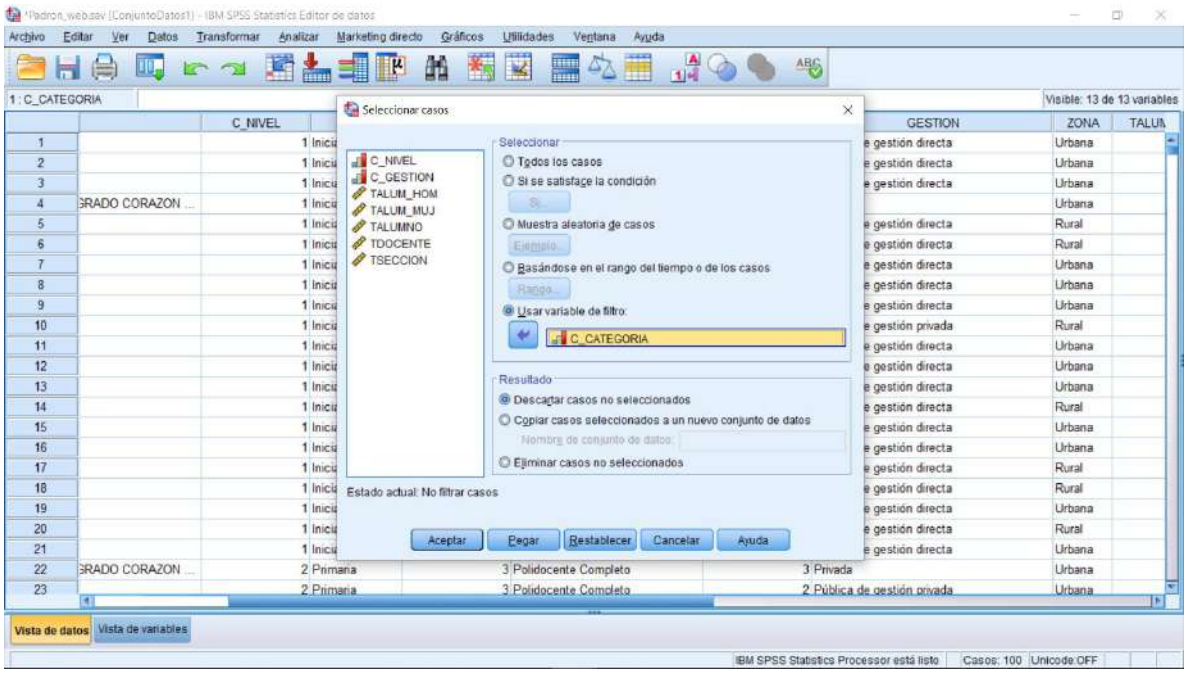
Paso 1: abrir la ventana "seleccionar casos".

Figura 9-5: Selección de casos en SPSS



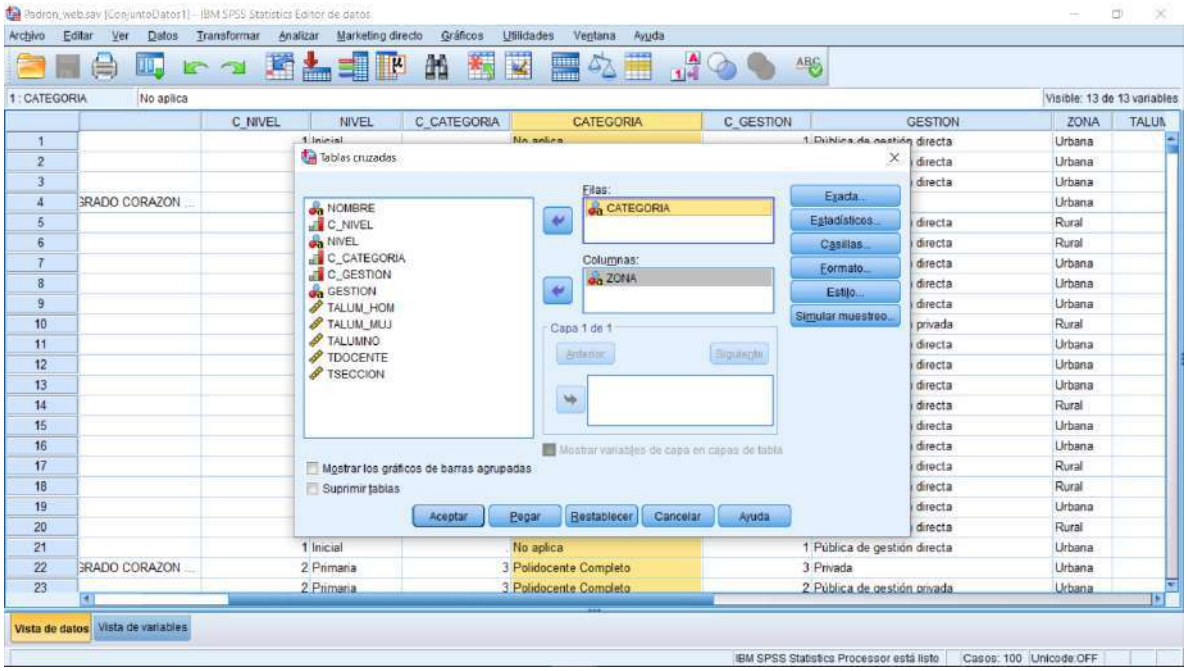
Paso 2: elegir "elegir variable de filtro" para solo tomar observaciones con valores válidos de C_CATEGORIA.

Figura 9-6: Elegir variable filtro en SPSS



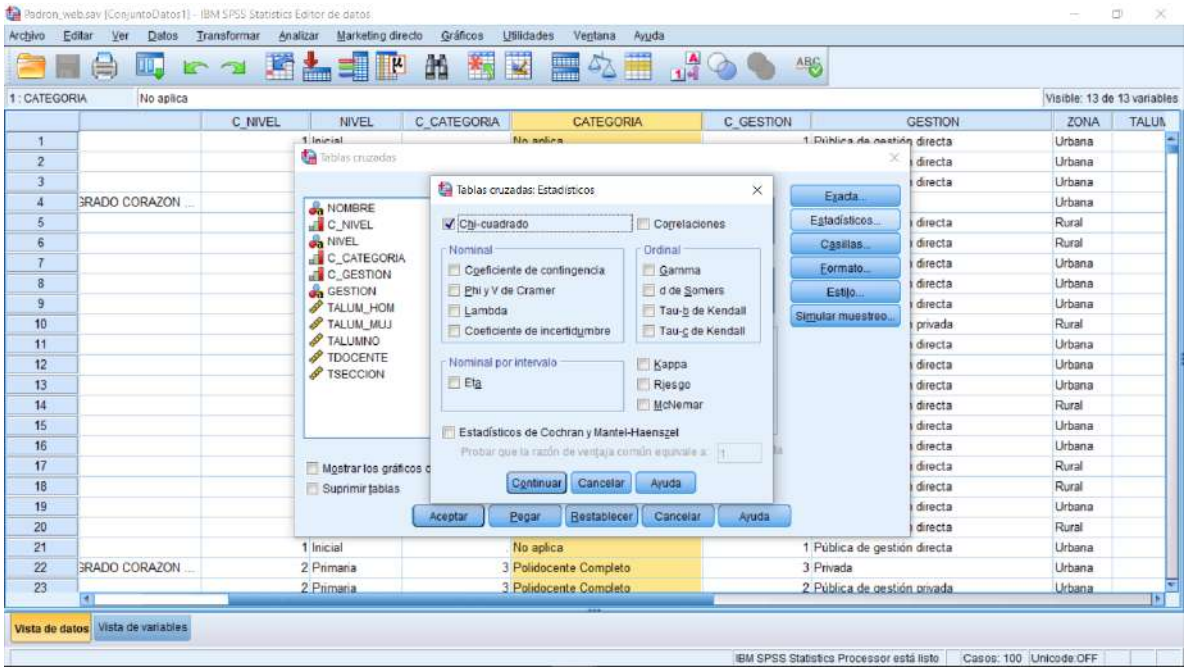
Paso 3: elegir el análisis de tablas cruzadas y luego las variables CATEGORIA y ZONA.

Figura 9-7: Tablas cruzadas en SPSS



Paso 4: activar el análisis Chi-cuadrado.

Figura 9-8: Chi cuadrado en SPSS



Las siguientes tablas de resultados estarán disponibles en el Visor de SPSS:

Tabla 9-7: Resumen de procesamiento de casos

	Casos					
	Válidos		Perdidos		Total	
	N	Porcentaje	N	Porcentaje	N	Porcentaje
CATEGORIA * ZONA	38194	100,0 %	0	0,0 %	38194	100,0 %

Tabla 9-8: Tabla cruzada CATEGORIA*ZONA

			ZONA		
			Rural	Urbana	Total
CATEGORIA	Polidocente Completo	Recuento	1623	12234	13857
		Recuento esperado	8082,2	5774,8	13857,0
	Unidocente o Polidocente	Recuento	20654	3683	24337
		Recuento esperado	14194,8	10142,2	24337,0
Total		Recuento	22277	15917	38194
		Recuento esperado	22277,0	15917,0	38194,0

Tabla 9-9: Pruebas de Chi-cuadrado

	Valor	gl	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
Chi-cuadrado de Pearson	19439,795[a]	1	,000		
Corrección de continuidad[b]	19436,785	1	,000		
Razón de verosimilitud	21187,832	1	,000		
Prueba exacta de Fisher				,000	,000
N de casos válidos	38194				

a. 0 casillas (0,0 %) han esperado un recuento menor que 5. El recuento mínimo esperado es 5774,78.

b. Solo se ha calculado para una tabla 2x2

Se observa que el análisis es estadísticamente significativo (<0,05). Esto indica que se rechaza la hipótesis nula (no hay relación entre ZONA y CATEGORIA) y se acepta la alternativa (sí hay relación). Comparando los valores de los Recuentos y los Recuentos Esperados de la tabla de contingencia, puede notarse que los colegios Unidocentes o Polidocentes Multigrado están concentrados en la zona rural. Entonces, la afirmación es FALSA.

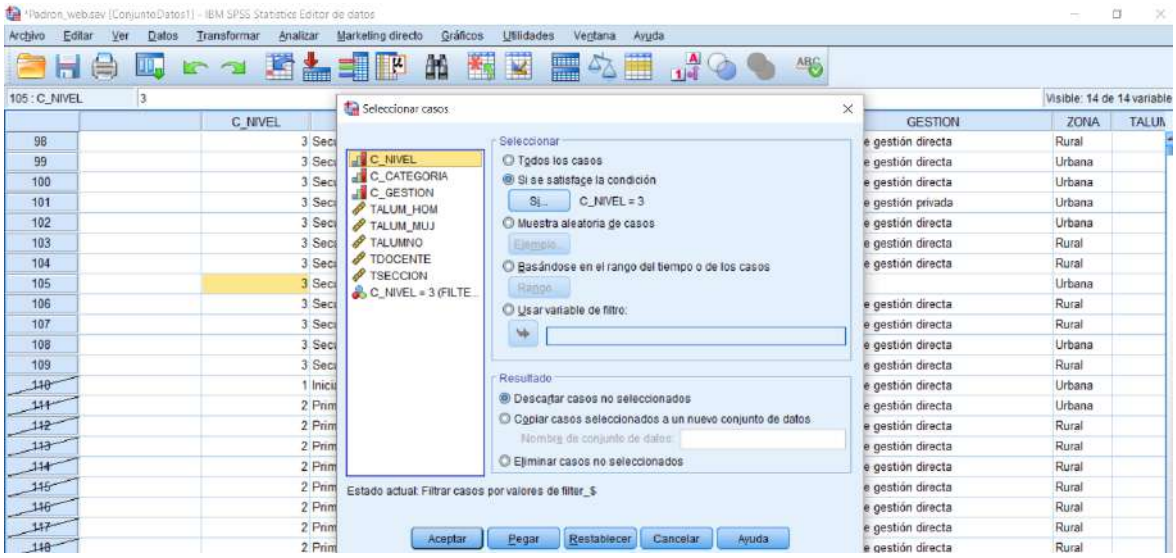
Ejercicio II: Relación entre variables cuantitativas

Para este ejercicio, emplee el archivo SPSS "SOL_9.1".

Un diagrama de dispersión ayuda a identificar, visualmente, la naturaleza de la relación entre variables (lineal, cuadrática, directa, inversa, etc.) si hubiera. Como para este caso se utilizará solo las observaciones de colegios del nivel secundario, se vuelve a activar la ventana "Seleccionar casos" y se sigue los siguientes pasos:

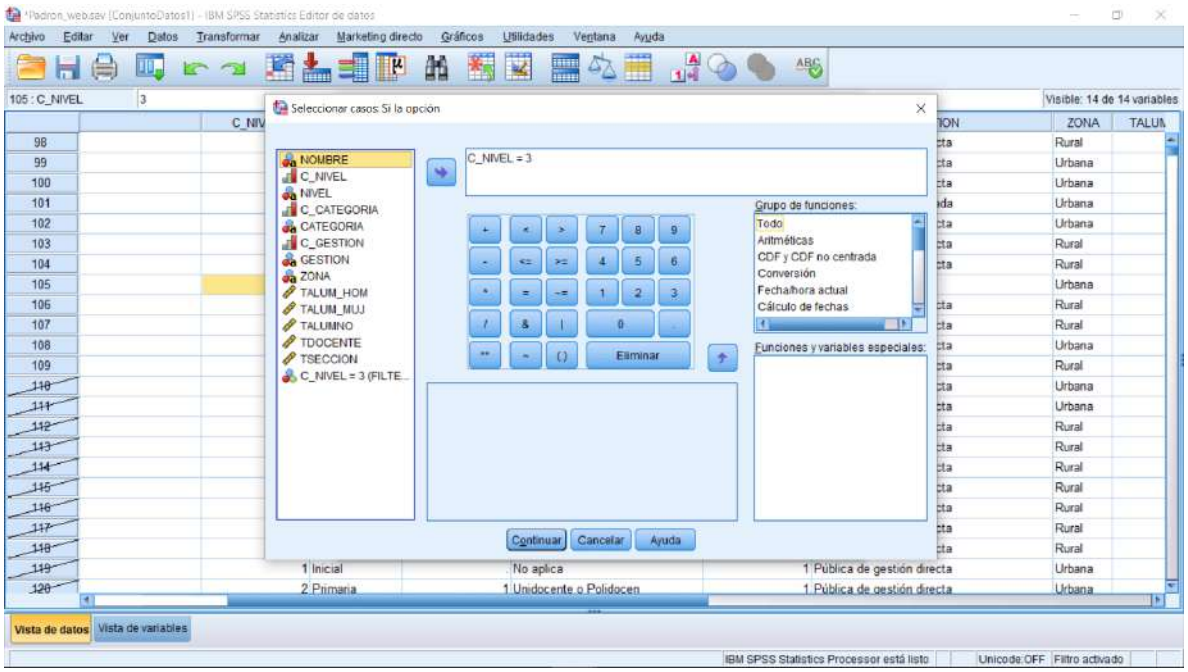
Paso 1: utilizar la opción "Si se satisface la condición".

Figura 9-9: Selección de casos en SPSS



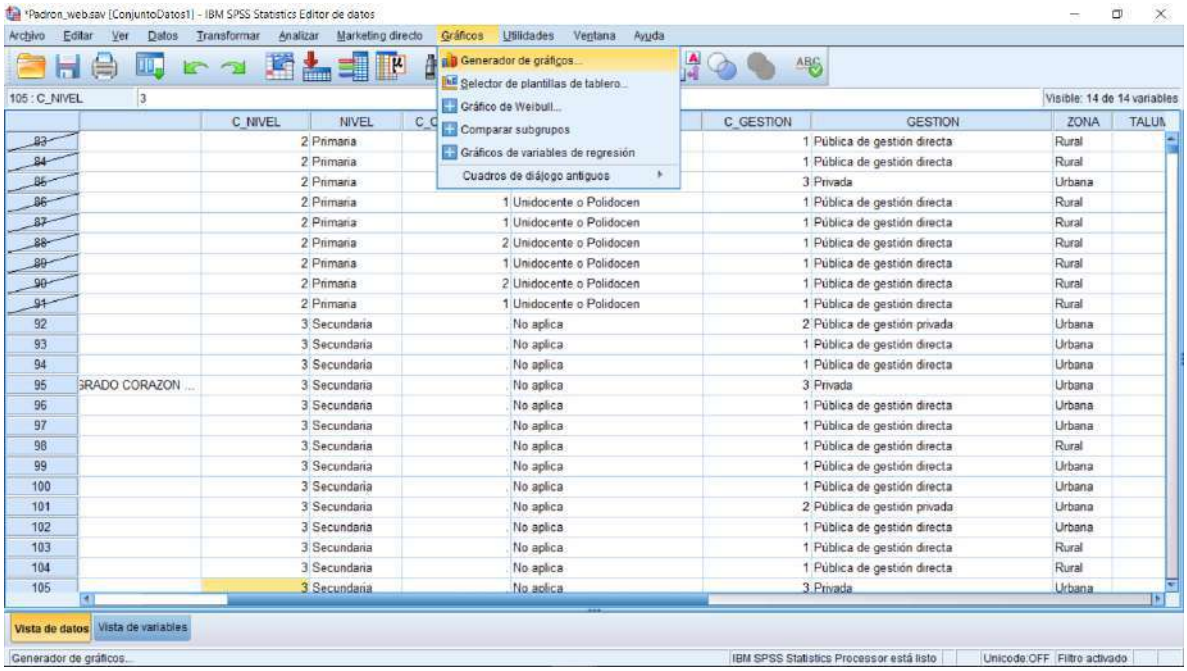
Paso 2: colocar la condición C_NIVEL = 3 (solo nivel secundario).

Figura 9-10: "Si se satisface la condición" en selección de datos



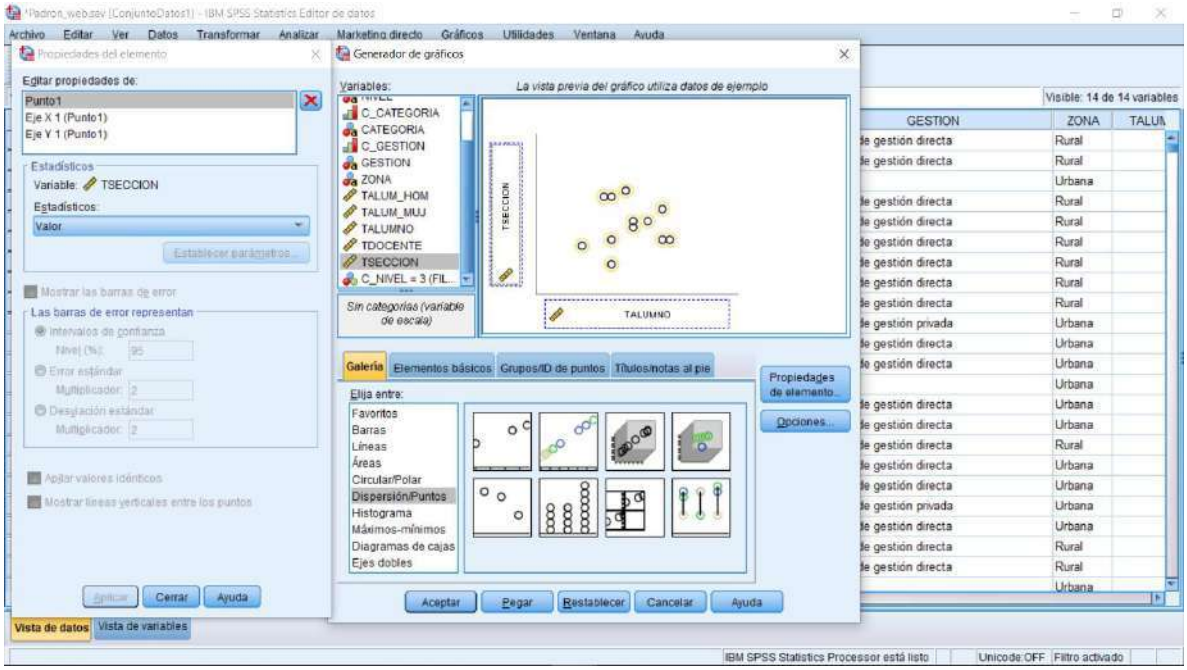
Paso 3: para el gráfico de dispersión, ingresar al "Generador de gráficos".

Figura 9-11: Generador de gráficos en SPSS



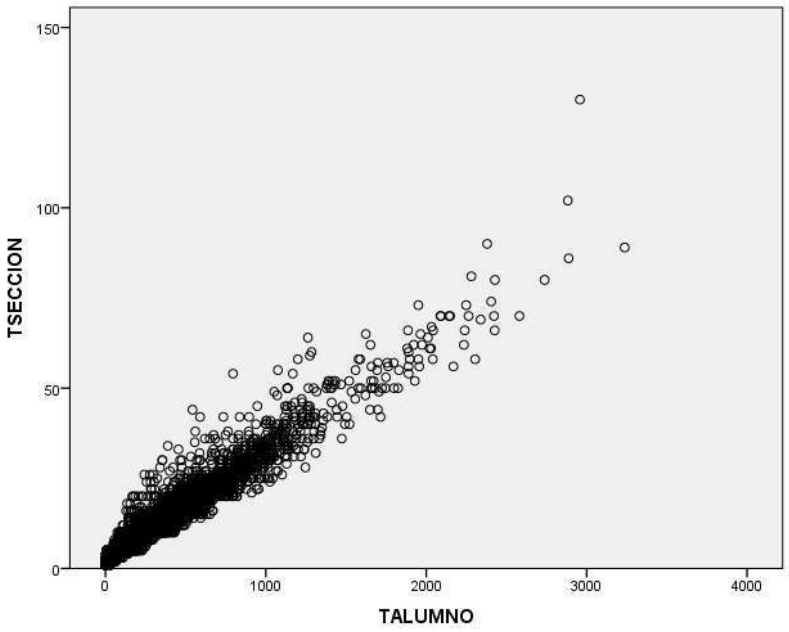
Paso 4: seleccionar el gráfico de dispersión adecuado y arrastrar las variables a los ejes correspondientes, como se muestra.

Figura 9-12: Gráfico de dispersión en SPSS



El resultado es el siguiente:

Figura 9-13: Gráfico de dispersión de estudiantes y secciones en el colegio

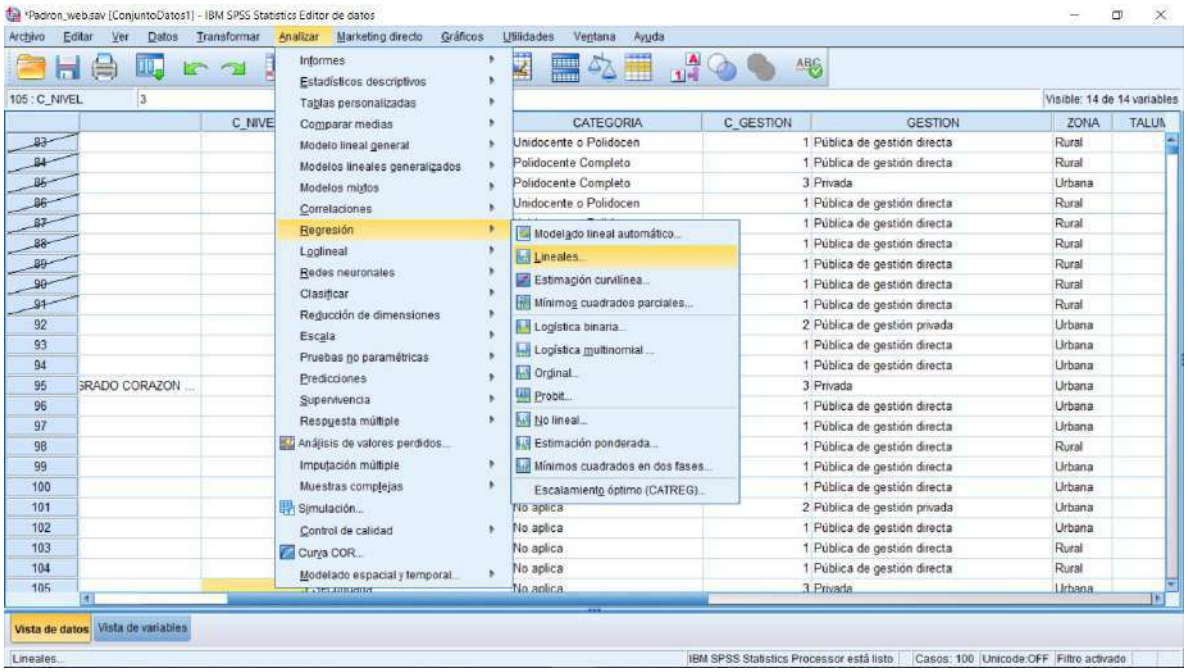


Se observa claramente una relación lineal positiva entre las dos variables (número de estudiantes y número de secciones en el colegio).

Para el análisis de regresión, según el nivel por separado, se activarán y desactivarán los niveles de manera similar a la ya mostrada. Se detalla el proceso para el nivel secundario.

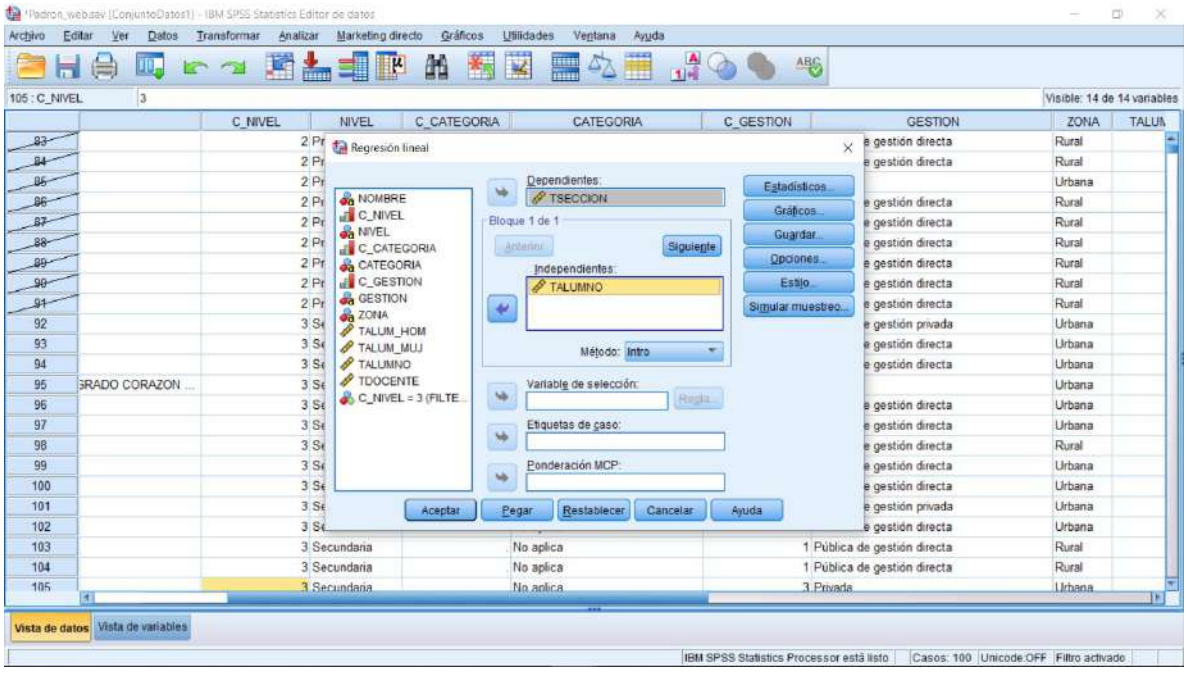
Paso 1: una vez que active solo las observaciones para secundaria, abra la ventana de regresiones.

Figura 9-14: Regresión lineal en SPSS



Paso 2: seleccionar las variables dependiente e independiente adecuadas

Figura 9-15: Selección de variables para una regresión en SPSS



Los resultados en el Visor de SPSS son los siguientes:

Tabla 9-10: Resumen del modelo, TALUMNO

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,966[a]	,934	,934	1,956

a. Predictores: (Constante), TALUMNO

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	2,939	,020		147,037	,000
	TALUMNO	,030	,000	,966	450,941	,000

a. Variable dependiente: TSECCION

El análisis para los niveles inicial y primaria es similar. A continuación, los resultados.

Nivel primaria:

Tabla 9-11: Resumen del modelo, primaria

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,951[a]	,904	,904	1,443

a. Predictores: (Constante), TALUMNO

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	4,268	,009		500,770	,000
	TALUMNO	,028	,000	,951	601,081	,000

a. Variable dependiente: TSECCION

Nivel inicial:

Tabla 9-12: Resumen del modelo, inicial

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,906[a]	,821	,821	,831

a. Predictores: (Constante), TALUMNO

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	2,046	,006		370,104	,000
	TALUMNO	,030	,000	,906	401,117	,000

a. Variable dependiente: TSECCION

En resumen:

Tabla 9-13: Significancias

NIVEL	COEFICIENTE	SIGNIFICANCIA	R CUADRADO
INICIAL	0,030	0,000	0,821
PRIMARIA	0,028	0,000	0,904
SECUNDARIA	0,030	0,000	0,934

Los tres análisis son significativos (se rechaza la hipótesis nula, que es que el coeficiente es igual a cero) y tienen un alto porcentaje de la varianza de la variable dependiente explicado por la variable independiente (R cuadrado o R cuadrado ajustado). El coeficiente en cada caso se interpreta como "el número de aulas que se incrementan al incrementar en uno el número de alumnos".

Caso aplicado: Tiendas de conveniencia

Para este ejercicio, emplee el archivo SPSS "SOL_9.2".

Para la solución del caso se utilizan las pruebas ya descritas en las soluciones de los ejercicios I y II por lo que solo se mostraran las tablas y gráficos de resultado y no el proceso paso a paso en SPSS.

Pregunta 1: Se analiza la relación entre las variables REGULACIÓN y REGISTRO DE NUEVA MARCA. Ambas son variables ordinales, por lo que se utilizará la prueba Chi-cuadrado. Se obtiene:

Tabla 9-14: Tabla cruzada REGULACIÓN*REGISTRO DE NUEVA MARCA

			REGISTRO DE NUEVA MARCA		
			NO	SÍ	Total
REGULACIÓN	MUY REGULADO	Recuento	3	4	7
		Recuento esperado	4,0	3,0	7,0
	POCO REGULADO	Recuento	13	8	21
		Recuento esperado	12,0	9,0	21,0
Total		Recuento	16	12	28
		Recuento esperado	16,0	12,0	28,0

Tabla 9-15: Pruebas de Chi-cuadrado

	Valor	gl	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
Chi-cuadrado de Pearson	,778[a]	1	,378		
Corrección de continuidad[b]	,194	1	,659		
Razón de verosimilitud	,772	1	,380		
Prueba exacta de Fisher				,418	,328
N de casos válidos	28				

a. 2 casillas (50.0 %) han esperado un recuento menor que 5. El recuento mínimo esperado es 3.00.

b. Solo se ha calculado para una tabla 2x2.

De acuerdo con esos resultados, no se rechaza la hipótesis nula que es "no hay relación entre las variables". Se concluye que hay evidencia estadísticamente significativa para afirmar que no hay relación entre los países en los que registraron la nueva marca y aquellos en los que la regulación es baja.

Pregunta 2: Se analiza la relación entre variables REGISTRO DE NUEVA MARCA y ESTUDIO DE MERCADO. Al igual que en el caso anterior, ambas variables son nominales, por lo que se utiliza nuevamente la prueba Chi-cuadrado, obteniéndose:

Tabla 9-16: Tabla cruzada ESTUDIO DE MERCADO*REGISTRO DE NUEVA MARCA

			REGISTRO DE NUEVA MARCA		
			NO	SÍ	Total
ESTUDIO DE MERCADO	CONVENIENTE	Recuento	3	12	15
		Recuento esperado	8,6	6,4	15,0
	NO CONVENIENTE	Recuento	13	0	13
		Recuento esperado	7,4	5,6	13,0
Total		Recuento	16	12	28
		Recuento esperado	16,0	12,0	28,0

Tabla 9-17: Pruebas de Chi-cuadrado

	Valor	gl	Significación asintótica (bilateral)	Significación exacta (bilateral)	Significación exacta (unilateral)
Chi-cuadrado de Pearson	18,200[a]	1	,000		
Corrección de continuidad[b]	15,080	1	,000		
Razón de verosimilitud	23,231	1	,000		
Prueba exacta de Fisher				,000	,000
N de casos válidos	28				

a. 0 casillas (.0 %) han esperado un recuento menor que 5. El recuento mínimo esperado es 5.57.

b. Solo se ha calculado para una tabla 2x2.

El resultado de esa prueba sí indica que se rechaza la hipótesis nula y se acepta la alternativa (sí hay relación entre las variables). Además, la tabla de contingencia muestra una concentración de las observaciones en las casillas Conveniente vs. Sí y No conveniente vs. No. Esto indica que hay evidencia estadísticamente significativa que muestra que los países en los que han registrado la nueva marca son aquellos en los que el estudio mostró que es conveniente implementar las TdC. Ambos análisis señalan que hay evidencia estadísticamente significativa que apoya la presunción de que estarían interesados en ingresar al mercado con una cadena de TdC.

Pregunta 3: Se utiliza el archivo "Caso_Locaciones USA.sav" y se estudia la relación entre las variables NSE_rec y PdC que representan la codificación de las variables NIVEL SOCIOECONOMICO y PRESENCIA DE COMPETIDOR, respectivamente. Ambas son ordinales y no tienen el mismo número de niveles (no se trata de una tabla de contingencia cuadrada), por lo que se hará uso de la prueba Tau-c de Kendall. Los resultados son:

Tabla 9-18: Tabla cruzada NSE_rec*PdC

			PdC		
			1,00	2,00	Total
NSE_rec	1,00	Recuento	9	2	11
		Recuento esperado	5,5	5,5	11,0
	2,00	Recuento	9	10	19
		Recuento esperado	9,5	9,5	19,0
	3,00	Recuento	2	8	10
		Recuento esperado	5,0	5,0	10,0
Total		Recuento	20	20	40
		Recuento esperado	20,0	20,0	40,0

Tabla 9-19: Medidas simétricas

		Valor	Error estandarizado asintótico[a]	T aproximada[b]	Significación aproximada
Ordinal por ordinal	Tau-c de Kendall	,480	,141	3,413	,001
N de casos válidos		40			

a. No se presupone la hipótesis nula.

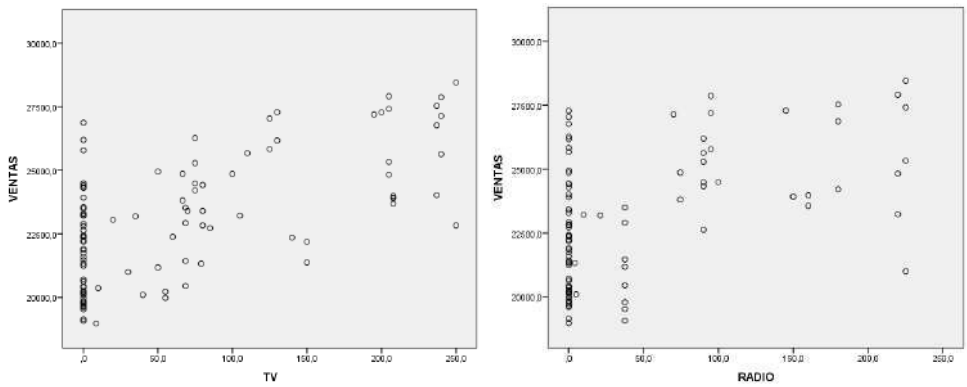
b. Utilización del error estándar asintótico que presupone la hipótesis nula.

La significancia muestra que debe rechazarse la hipótesis nula y aceptarse la alternativa (sí hay relación entre las variables). Además, al ser el valor positivo señala que la relación entre las variables es positiva, es decir "a mayor nivel socioeconómico, más puntos de venta del competidor".

Ese resultado es un indicio de que las cadenas de TdC del competidor están orientadas a niveles socioeconómicos altos.

Pregunta 4: Se utiliza el archivo "Caso_Datos ventas.sav". El siguiente es el gráfico de dispersión de las variables VENTAS vs. TV y VENTAS vs. RADIO:

Figura 9-16: Dispersión de VENTAS vs. TV y VENTAS vs. RADIO



En ambos casos se aprecia una ligera relación positiva. Los coeficientes de correlación ayudarán a complementar ese análisis inicial:

Tabla 9-20: Correlaciones

		VENTAS	TV	RADIO
VENTAS	Correlación de Pearson	1	,632**	,505**
	Sig. (bilateral)		,000	,000
	N	101	101	97
TV	Correlación de Pearson	,632**	1	,361**
	Sig. (bilateral)	,000		,000
	N	101	101	97
RADIO	Correlación de Pearson	,505**	,361**	1
	Sig. (bilateral)	,000	,000	
	N	97	97	97

**. La correlación es significativa en el nivel 0,01 (bilateral).

Los valores de los coeficientes corroboran lo señalado al interpretar los gráficos. La relación entre las ventas y la publicidad, en Radio o TV, es positiva. Esto apoya lo señalado por la gerenta.

Pregunta 5: Para cuantificar la relación encontrada, se recurre al análisis de regresión lineal. Se estimarán dos modelos: para ambos, la variable dependiente será VENTAS mientras que, en un caso, la independiente será TV y en el otro, RADIO.

Ventas vs. TV:

Tabla 9-21: Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,632[a]	,399	,393	1908,6866

a. Predictores: (Constante), TV

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	21803,378	245,352		88,866	,000
	TV	18,617	2,294	,632	8,114	,000

b. Variable dependiente: VENTAS

Ventas vs. RADIO:

Tabla 9-22: Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	,505[a]	,255	,247	2169,5496

a. Predictores: (Constante), RADIO

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error estándar	Beta		
1	(Constante)	22251,917	261,172		85,200	,000
	RADIO	17,825	3,129	,505	5,696	,000

a. Variable dependiente: VENTAS

En resumen:

Tabla 9-23: Significancias

DEPENDIENTE	COEFICIENTE	SIGNIFICANCIA	R CUADRADO
TV	18,617	0,000	0,399
RADIO	17,825	0,000	0,255

Ambos análisis son significativos (se rechaza la hipótesis nula, que es que el coeficiente es igual a cero) pero tienen un bajo porcentaje de la varianza de la variable dependiente explicado por la variable independiente (R cuadrado bajo). El coeficiente en cada caso se interpreta como "incremento en ventas (€) por cada GRP, en TV o Radio según sea el caso".

Lecturas recomendadas

- Levin, R., y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed. revisada, cap. 12.1). México: Pearson Educación de México; Prentice Hall.
- Gujarati, D. (2004). *Econometría básica* (4.ª ed., introducción). McGraw-Hill Companies.

10. Introducción a la econometría, el modelo de regresión lineal general

La introducción a conceptos econométricos nos permitirá acceder a herramientas matemáticas y estadísticas que tienen como objetivo buscar, definir y cuantificar relaciones entre distintas variables que pertenecen a un modelo econométrico. En este capítulo se introducirá el uso del modelo de regresión lineal general.

Palabras clave:

- Econometría
- Regresión
- Modelo de regresión lineal general
- Relación entre variables

Ejercicio I: Estimación de ventas

Para este ejercicio, emplee el archivo en Excel "BD_10".

El gerente de ventas de una compañía de refacciones para automóviles tiene como objetivo predecir sus ventas anuales por región. Para ello, plantea un modelo lineal que tiene como variable dependiente sus ventas (en millones de soles) y como variables explicativas el número de distribuidoras que posee la compañía y el número de automóviles que se encuentran registrados en cada región.

Estime el modelo planteado y responda las siguientes preguntas: ¿cuánto es la variación esperada de las ventas si aumentara el número de distribuidoras en una unidad? ¿Cuánto es la variación esperada de las ventas si disminuye el número de automóviles registrados en mil unidades? Además, discuta la significancia individual y global del modelo a un nivel de significancia del 5 %.

Ejercicio II: Estimación de la demanda del pollo

Para este ejercicio, emplee el archivo en Excel "BD_10".

Un investigador desea analizar la demanda del pollo en función del precio del pollo y de productos sustitos como la carne y el pescado. Para ello, plantea el siguiente modelo que considera como variables explicativas la evolución de los precios promedios mensuales al consumidor final desde el año 2010 al 2015 en Lima Metropolitana publicados por el INEI:

$$Q_i = \beta_1 + \beta_2 \times PC_i + \beta_3 \times PD_i + \beta_4 PP_i + \mu_i$$

Donde Q representa la demanda pollo en miles de kilos, PC el precio de la carne, PD el precio del pescado y PP el precio del pollo. Los precios están en soles por kilo.

Estime el modelo e interprete los coeficientes estimados y la bondad de ajuste del modelo. Asimismo, responda a la siguiente pregunta: ¿qué variación se produce en la demanda del pollo si el precio del pescado se reduce en S/0.90, manteniéndose los precios del resto de los productos constantes?

Caso aplicado: Determinantes del sueldo

Para este ejercicio, emplee el archivo en Excel "BD_10".

La consultora MIC S.A.C. cuenta con una amplia experiencia en investigaciones sociales y económicas. Por ello, ha sido contratada para elaborar un estudio sobre los factores que determinan el sueldo de los recién egresados de la carrera de Gestión. Dicho estudio ha sido encargado al equipo de Investigaciones Sociales, del cual usted forma parte.

Luego de analizar la población objetivo, el equipo estimó que el tamaño de muestra óptimo para realizar el estudio es de 114 egresados en los últimos dos años de la carrera de Gestión.

Usted cuenta con sólidos conocimientos en estimación de modelos econométricos. Por ello, el jefe del equipo le encargó lo siguiente:

a) Definir las posibles variables que podrían explicar el sueldo de los egresados de la carrera de Gestión, y justificar su respuesta.

Luego de la revisión del listado de variables explicativas que proporcionó, se consideran las siguientes variables para realizar el estudio: sueldo anterior, edad, género, promedio de notas y prácticas preprofesionales.

b) Establecer *a priori* el signo esperado de la relación entre las variables explicativas del modelo con la variable dependiente. Asimismo, ¿espera que la variable sea incluida en el modelo final?

c) Plantear el modelo matemático.

Al revisar las variables que finalmente habían sido seleccionadas, usted tiene sospechas que para el presente estudio la edad no es relevante, dado que la edad debería ser similar entre los recién egresados, por lo que espera que esta variable no sea determinante en el sueldo.

d) ¿Cómo podría verificar si su sospecha es cierta? Justifique.

Usted está interesado en identificar si existe una brecha salarial entre hombres y mujeres. Por ello, empieza a revisar diversas investigaciones al respecto, entre las cuales encontró una investigación sobre la brecha salarial donde una de las principales conclusiones es que esta se había reducido de forma sostenida en los últimos años, principalmente cuando se analiza a personas con estudios superiores. Ante ello, usted le comenta de la investigación encontrada a su jefe, ante lo cual le solicita:

e) revisar la interpretación realizada en la pregunta b y verificar si es necesario cambiar el análisis.

Al revisar los datos de la encuesta en la base de datos proporcionada para esta unidad, verificó que se cuenta con diversas variables cualitativas y cuantitativas. Por ello, le solicitó:

f) determinar cuántas variables cualitativas contiene la encuesta, y cuántas variables *dummy* se tienen que crear para poder estimar la regresión.

Asimismo, se le solicita:

g) plantear el modelo econométrico que considere solo las variables cuantitativas, calcular los coeficientes y realizar el análisis estadístico correspondiente;
h) plantear el modelo con todas las variables (es decir, tanto cualitativas como cuantitativas) y realizar el análisis estadístico correspondiente;

i) ¿cuál es el modelo econométrico planteado finalmente?

Al observar los resultados, valida que efectivamente todas las variables son significativas individualmente, que el modelo cuenta con significancia conjunta y que presenta un adecuado R^2. Sin embargo, tiene dudas sobre si es el mejor modelo.

Para eliminar dicha duda, realiza modelos alternativos utilizando el *software* Eviews y, entre cuatro posibles modelos, se le solicita:

j) indicar cuál de los modelos que se presentan a continuación es el mejor, y justificarlo.

Figura 10-1: Modelos

Equation: EQ01 Workfile: BD::Bd\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: SUELDO
Method: Least Squares
Date: 02/16/18 Time: 13:35
Sample: 1 114
Included observations: 114

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-2800.948	410.7584	-6.818967	0.0000
SUELDO_PASADO	0.535755	0.078193	6.851706	0.0000
PROMEDIO_DE_NOTAS_DE_LA_	292.4807	31.26924	9.353623	0.0000
RP	179.2236	76.10000	2.355107	0.0203

R-squared	0.924495	Mean dependent var	2730.281
Adjusted R-squared	0.922436	S.D. dependent var	743.2098
S.E. of regression	206.9867	Akaike info criterion	13.53764
Sum squared resid	4712783.	Schwarz criterion	13.63365
Log likelihood	-767.6457	Hannan-Quinn criter.	13.57661
F-statistic	448.9519	Durbin-Watson stat	0.738335
Prob(F-statistic)	0.000000		

Equation: EQ03 Workfile: BD::Bd\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: SUELDO
Method: Least Squares
Date: 02/16/18 Time: 13:36
Sample: 1 114
Included observations: 114

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	941.6071	123.8976	7.599884	0.0000
SUELDO_PASADO	0.932721	0.087599	10.64758	0.0000
RP	521.7984	88.97811	5.864346	0.0000

R-squared	0.864441	Mean dependent var	2730.281
Adjusted R-squared	0.861998	S.D. dependent var	743.2098
S.E. of regression	276.0919	Akaike info criterion	14.10531
Sum squared resid	8461171.	Schwarz criterion	14.17731
Log likelihood	-801.0026	Hannan-Quinn criter.	14.13453
F-statistic	353.9151	Durbin-Watson stat	0.955492
Prob(F-statistic)	0.000000		

Equation: EQ02 Workfile: BD::Bd\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: SUELDO
Method: Least Squares
Date: 02/16/18 Time: 13:35
Sample: 1 114
Included observations: 114

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	451.8856	104.2804	4.333371	0.0000
SUELDO_PASADO	1.346604	0.059122	22.77667	0.0000

R-squared	0.822441	Mean dependent var	2730.281
Adjusted R-squared	0.820856	S.D. dependent var	743.2098
S.E. of regression	314.5668	Akaike info criterion	14.35766
Sum squared resid	11082651	Schwarz criterion	14.40566
Log likelihood	-816.3865	Hannan-Quinn criter.	14.37714
F-statistic	518.7767	Durbin-Watson stat	1.454719
Prob(F-statistic)	0.000000		

Equation: EQ04 Workfile: BD::Bd\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: SUELDO
Method: Least Squares
Date: 02/16/18 Time: 13:36
Sample: 1 114
Included observations: 114

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2211.735	47.38856	46.67234	0.0000
RP	1285.091	74.60141	17.22609	0.0000

R-squared	0.725986	Mean dependent var	2730.281
Adjusted R-squared	0.723539	S.D. dependent var	743.2098
S.E. of regression	390.7761	Akaike info criterion	14.79153
Sum squared resid	17103068	Schwarz criterion	14.83954
Log likelihood	-841.1175	Hannan-Quinn criter.	14.81102
F-statistic	296.7382	Durbin-Watson stat	0.091143
Prob(F-statistic)	0.000000		

Una vez que se ha definido el mejor modelo posible, se le solicita:

k) interpretar los coeficientes.

Solucionario

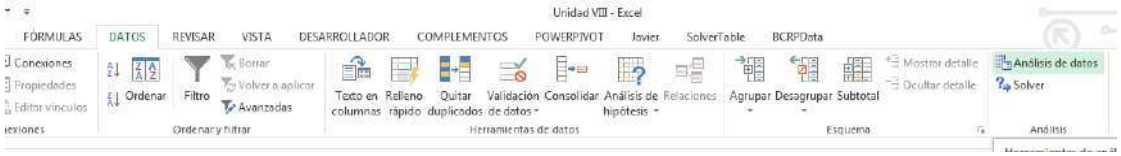
Ejercicio I: Estimación de ventas

Para este ejercicio, emplee el archivo en Excel "SOL_10".

(Continúa del enunciado del Ejercicio 1 del Capítulo 10)

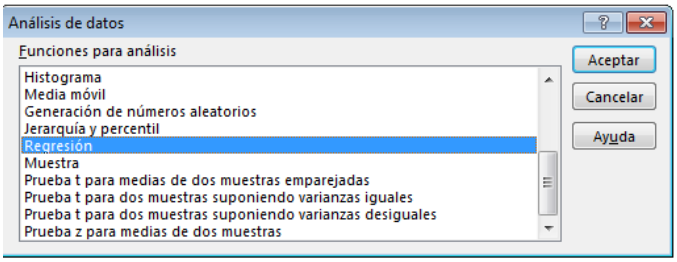
Para estimar el modelo se debe de ir a la pestaña Datos y seleccionar "Análisis de datos":

Figura 10-2: Análisis de datos en Excel



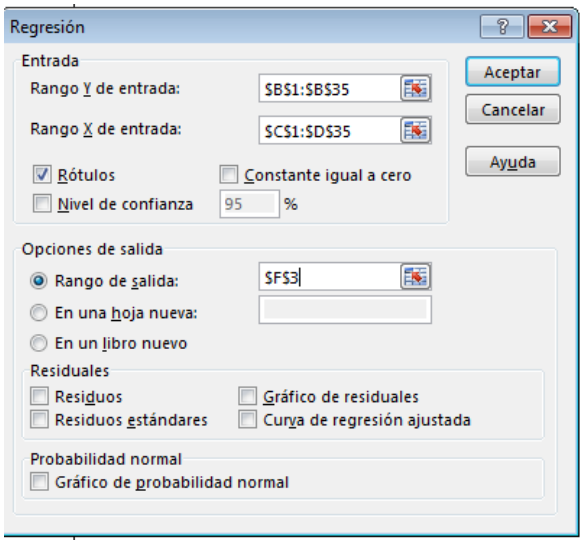
Luego se debe seleccionar "Regresión".

Figura 10-3: Regresión en Excel



Se debe seleccionar a la variable "Ventas" como variable dependiente mientras que a las variables "Número de distribuidoras" y "Número de automóviles registrados" como variables independientes:

Figura 10-4: Regresión en Excel



No olvidar de activar la casilla "Rótulos", en caso de que se haya seleccionado a las variables con sus respectivos nombres como en este caso. El resultado obtenido es el siguiente:

Figura 10-5: Estadísticos de regresión en Excel

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.507482313
Coeficiente de determinación R^2	0.257538298
R^2 ajustado	0.209637543
Error típico	11.82387398
Observaciones	34

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	2	1503.311717	751.6558583	5.3764977	0.009895731
Residuos	31	4333.923872	139.8039959		
Total	33	5837.235588			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
Intercepción	7.728847615	6.170868021	1.252473329	0.2197641	-4.85672069	20.31441592	-4.85672069	20.31441592
Número de Distribuidoras	0.004365965	0.002868961	1.521793095	0.1381971	-0.001485319	0.010217249	-0.00148532	0.010217249
Número de Automóviles Registrados	0.945294559	0.332497843	2.843009601	0.0078393	0.267160738	1.62342838	0.267160738	1.62342838

Del cual se obtiene la siguiente ecuación de regresión:

$$Ventas = 7.7288 + 0.0046 \times (Distribuidores) + 0.9453 \times (Automoviles) + \mu$$

A partir del modelo estimado, se espera que las ventas aumenten en 0.0046 millones de soles por cada unidad que se aumente en el número de distribuidoras. Además, se espera que las ventas aumenten en 0.9453 millones soles por cada automóvil registrado adicional.

La hipótesis para evaluar la significancia individual es la siguiente para cada parámetro del modelo:

$$H_o: \beta = 0$$
$$H_1: \beta \neq 0$$

Para verificar el rechazo de la hipótesis de no significancia individual (hipótesis nula), se debe verificar que la probabilidad asociada a cada parámetro estimado sea menor al nivel de significancia (por lo general es 0.05).

El parámetro estimado de la variable "Número de distribuidoras" tiene una probabilidad de 0.1382. Entonces, no cuenta con significancia individual, es decir, no se puede rechazar H_o. Por esa razón, se debería sacar del modelo.

La variable "Número de automóviles registrados" tiene una probabilidad inferior a 0.05, por lo que sí es significativa individualmente; de manera que se rechaza la H_o.

En lo que respecta a la significancia conjunta del modelo, se tiene la siguiente hipótesis:

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_1: Al\ menos\ uno\ de\ los\ coeficientes\ es\ diferente\ de\ cero$$

A partir del modelo estimado, se verifica el rechazo de la hipótesis de no significancia conjunta (hipótesis nula) siendo que la probabilidad asociada al estadístico F del modelo estimado tiene el valor de 0.0098, el cual es menor al nivel de significancia, que en este caso es 0.05.

Ejercicio II: Estimación de la demanda del pollo

Para este ejercicio, emplee el archivo en Excel "SOL_10".

(Continúa del enunciado del Ejercicio II del Capítulo 10)

Para responder esta pregunta se debe de realizar la regresión correspondiente con los datos indicados. El modelo a estimar es el siguiente:

$$Q = \beta_0 + \beta_1 \times (IPC) + \beta_2 \times (PC) + \beta_3 \times (PD) + \beta_3 \times (PP) + \mu$$

Para realizar la regresión, se ejecuta la herramienta "Análisis de datos" tal como se hizo en el ejercicio anterior. Para configurar el análisis, se debe seleccionar a la variable "Demanda de Pollo" como variable dependiente mientras que las variables "Precio de la carne", "Precio del pescado" y "Precio del pollo" se seleccionarán como variables independientes. De esta manera, se obtiene el siguiente resultado:

Figura 10-6: Estadísticos de regresión en Excel

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.871485459
Coeficiente de determinación R^2	0.759486904
R^2 ajustado	0.748876033
Error típico	15.90670707
Observaciones	72

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	3	54331.41817	18110.47272	71.576296	5.29501E-21
Residuos	68	17205.58642	253.0233297		
Total	71	71537.0046			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
Intercepción	-534.1370454	44.3147736	-12.05325001	1.533E-18	-622.5657993	-445.7082914	-622.5657993	-445.7082914
PP	-8.737448205	5.089181362	-1.716867131	0.0905555	-18.8927515	1.41785509	-18.8927515	1.41785509
PD	-3.811481305	3.550616062	-1.073470417	0.2868548	-10.89662534	3.273662734	-10.89662534	3.273662734
PC	55.14685621	6.38831982	8.632450748	1.56E-12	42.39916249	67.89454994	42.39916249	67.89454994

Reemplazando los valores obtenemos el siguiente modelo estimado:

$$Q = -534.13 + 55.15 \times (PC) - 3.81 \times (PD) - 8.74 \times (PP) + \mu$$

Interpretación de coeficientes[13]:

PC: Precio de la carne

El coeficiente de PC es 55.15, de lo que se desprende que por cada unidad adicional en el PC, se espera que la Demanda de pollo (Q) aumente en 55.15 unidades.

[13] Con fines didácticos, en este ejercicio no se ha hecho antes la evaluación de significancia de parámetros del modelo. Se recomienda revisar este aspecto en el caso aplicado presentado a continuación.

PD: Precio de pescado

El coeficiente de PD es -3.81, de lo que se desprende que por cada unidad adicional en el PD, se espera que la demanda de pollo (Q) disminuya en 3.81 unidades.

PP: Precio de pollo

El coeficiente de PP es -8.74, de lo que se desprende que por cada unidad adicional en el PP se espera que la demanda de pollo (Q) disminuya en 8.74 unidades.

Sobre la bondad de ajuste del modelo (R^2), alcanza un valor de 0.7594, es decir, las variaciones de la demanda del pollo son explicadas por el precio de la carne, precio de pescado y precio de pollo en un 75.94 %.

Con respecto a la variación de la demanda del pollo, si el precio del pescado (PD) se reduce en S/0.90, se puede estimar con dos métodos:

- Método simplificado

  En el primer enunciado se indica que el precio del pescado se reduce en 0.90 y se desea saber qué efecto se produce en la demanda del pollo. Lo que se debe hacer es trabajar con el modelo en variaciones y considerar que la variación en PD es -0.90 y asumir que el resto de variables se mantienen constantes, esto es:

$$\Delta Q = -3.81(-0.90)$$
$$\Delta Q = 3.429$$

- Método no simplificado

  Se debe escoger una situación inicial, sobre la cual se va a disminuir el precio del pescado en 0.90 y luego se analizará mediante una diferencia cuál es el efecto sobre la demanda del pollo. Por simplicidad, supondremos que el PC es 20, el de PD 10 y el de PP valen 15, reemplazamos:

$$Q = -534.13 + 55.15 \times (20) - 3.81 \times (10) - 8.74 \times (15)$$
$$Q = 399.67 \dots (1)$$

  Para analizar el efecto de la reducción, realizamos el mismo cálculo, pero disminuyendo el PD en 0.90.

$$Q = -534.13 + 55.15 \times (20) - 3.81 \times (9.10) - 8.74 \times (15)$$
$$Q = 403.09 \dots (2)$$

  Restamos $(2) - (1)$ y obtenemos:

$$\Delta Q = 3.429$$

Caso aplicado: Determinantes del sueldo

Para este ejercicio, emplee el archivo en Excel "SOL_10".

(Continúa del enunciado del caso aplicado del Capítulo 10)

a) A continuación, se brindan algunas posibles variables. Luego de ellas, usted puede completar con otras que se han omitido, incluyendo una pequeña justificación.

Tabla 10-1: Variables

Nombre de variable	Justificación
Manejo de Excel	El uso de herramientas informáticos es considerado para contratar personal en una empresa.
Conocimiento de idiomas	El conocimiento de idiomas, al igual que el manejo de herramientas informáticas, es valorado por las empresas.
Prácticas preprofesionales	La experiencia previa también es un factor importante debido a que se espera que esta aporte mayor valor agregado.

Luego de la revisión del listado de variables explicativas que proporcionó, se consideran las siguientes variables para realizar el estudio: sueldo anterior, edad, género, promedio de notas y prácticas preprofesionales.

b) Lista de variables y signo esperado:

Tabla 10-2: Signo esperado de las variables

Nombre de variable	Signo esperado	¿Incluir en modelo?
Sueldo anterior	Signo positivo, relación directa	SÍ
Edad	Signo positivo, relación directa	SÍ
Género	No se puede determinar *a priori*, posiblemente no influya	SÍ
Promedio de notas	Signo positivo, relación directa	SÍ
Prácticas preprofesionales	Signo positivo, relación directa	SÍ

c) Modelo planteado:

$$Sueldo = \beta_1 + \beta_2(Sueldo\ anterior) + \beta_3(Edad) + \beta_4(Genero) + \beta_5(Promedio\ de\ Notas) + \beta_6(Practicas)$$

Al revisar las variables que finalmente habían sido seleccionadas, usted tiene sospechas que para el presente estudio la edad no es relevante, dado que esta debería ser similar entre los recién egresados, por lo que espera que esta variable no sea determinante en el sueldo.

d) En la base de datos se observa que la variable edad va desde 22 hasta 26 años. Para verificar la sospecha, se puede realizar una pequeña regresión entre ambas variables:

$$Sueldo = \beta_1 + \beta_2(Edad)$$

Para realizar la regresión, se debe emplear la herramienta "Análisis de datos". Para este caso, se debe seleccionar el rango de Y, que en este caso es "Sueldo" y el rango de X, que en este caso es "Edad". Si incluimos el nombre de las variables, debemos seleccionar "Rótulos" y, finalmente, debemos indicar el rango de salida. Se obtiene el siguiente resultado:

Figura 10-7: Estadísticos de regresión

Resumen

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,068592422
Coeficiente de determinación R^2	0,00470492
R^2 ajustado	-0,004181643
Error típico	744,7620731
Observaciones	114

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	293665,9155	293665,9155	0,529442059	0,468359891
Residuos	112	62123101,1	554670,5456		
Total	113	62416767,02			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	3615,13003	1218,073153	2,9679088	0,003667083	1201,674216	6028,585844	1201,674216	6028,585844
Edad	-37,30503824	51,26940417	-0,727627692	0,468359891	-138,8887906	64,27871414	-138,8887906	64,27871414

Se observa que el R^2 es casi cero y, además, la variable "Edad" no cuenta con significancia individual dado que tiene probabilidad asociada a la hipótesis nula de no significancia individual es de 0.4683. Es así que sus sospechas fueron validadas estadísticamente, de manera que la variable edad no se debería incluir en el modelo.

e) En "b" se había concluido lo siguiente:

Tabla 10-3: Variable edad

Nombre de variable	Interpretación	¿Incluir en modelo?
Edad	Signo positivo, relación directa	SÍ

Sin embargo, según la investigación encontrada, la edad no es un determinante del sueldo.

f) Se observa que hay dos variables cualitativas, según el siguiente detalle:

- Género: se tiene dos categorías; se debe crear 1 *dummy*.
- Realizo prácticas: se tiene dos categorías; se debe crear 1 *dummy*

Recordar que el número de variables *dummy* es igual al número de categorías de la variable cualitativa menos uno.

g) Modelo:

$$Sueldo = \beta_1 + \beta_2(Sueldo\ anterior) + \beta_3(Edad) + \mu$$

Para esta pregunta, se aplicará la herramienta "Análisis de datos" y se ejecutará el análisis de regresión como se ha venido haciendo en las preguntas anteriores. Seguidamente, se seleccionará el rango de Y, que en este caso es la variable "Sueldo" y el rango de X, que en este caso está comprendido por las variables "Edad", "Sueldo pasado" y "Promedio de notas". Los resultados obtenidos son los siguientes:

Figura 10-8: Estadísticos de regresión

Resumen

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0,960027939
Coeficiente de determinación R^2	0,921653644
R^2 ajustado	0,919516925
Error típico	210,8451881
Observaciones	114

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	3	57526640,75	19175546,92	431,3406332	1,20013E-60
Residuos	110	4890126,268	44455,69335		
Total	113	62416767,02			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	-3014,300921	460,609467	-6,54415755	1,96244E-09	-3927,120772	-2101,481071	-3927,120772	-2101,481071
Edad	-17,05576889	14,64585805	-1,164545555	0,246722413	-46,0804219	11,96888412	-46,0804219	11,96888412
Sueldo Pasado	0,585122005	0,075810082	7,71826112	5,79035E-12	0,434884217	0,735359792	0,434884217	0,735359792
Promedio de notas de la Carrera	331,4590772	28,0853181	11,80186303	2,89768E-21	275,8005689	387,1175856	275,8005689	387,1175856

Se observa que el R^2 alcanza el valor de 0.92, de lo que se desprende que las variaciones de las variables independientes explican a las dependientes en 92 %, es decir, el modelo explica bastante bien las variaciones del sueldo.

En lo que respecta la significancia individual, se plantean las siguientes hipótesis nula y alternativa:

$$H_o: \beta = 0$$
$$H_1: \beta \neq 0$$

Se verifican estas hipótesis con el valor de las probabilidades de cada variable; estas deben ser menores al nivel de significancia (por lo general es 0.05) para rechazar la hipótesis nula.

La variable edad tiene una probabilidad de 0.24, por lo que no cuenta con significancia individual, es decir, no se puede rechazar H_o. Por esa razón, no debería ser empleada dentro del modelo. Asimismo, las variables "Sueldo Pasado" y "Promedio de notas de la carrera" tienen una probabilidad inferior a 0.05, por lo que sí son significativas individualmente, es decir, en estos casos se rechaza la H_o.

En lo que respecta a significancia conjunta, se tienen las siguientes hipótesis:

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_1: Al\ menos\ uno\ de\ los\ coeficientes\ es\ diferente\ de\ cero$$

Se verifica con el valor de la probabilidad del modelo, el cual debe ser menor al nivel de significancia (por lo general es 0.05). El valor de $Fcrítico$ es inferior a 0.05, por lo que el modelo cuenta con significancia conjunta.

h) El modelo a estimar sería el siguiente:

$$Sueldo = \beta_1 + \beta_2(Sueldo\ anterior) + \beta_3(Edad) + \beta_4(Promedio\ de\ notas) + \beta_5(G) + \beta_6(Rp) + \mu$$

Se debe iniciar con la codificación de las variables *dummy* para poder incluirlas en el modelo, para lo cual seguiremos el siguiente criterio de codificación:

- Variable "Género" (G): 0 cuando la persona sea mujer y 1 cuando la persona sea hombre.
- Variable "Realizó prácticas" (Rp): 0 cuando la persona no realizó prácticas y 1 cuando la persona si realizó prácticas.

La codificación en el Excel se realiza de la siguiente manera:

1. Insertamos columnas al lado izquierdo de la variable "Género" (ello para tener a todas las variables independientes juntas.
2. En la celda "F2" ingresamos el condicional para codificar la variable "Género" según lo descrito líneas arriba:

$$= SI(H2 = "Mujer", 0,1)$$

3. Repetir la fórmula para cada uno de los casos de la base de datos.

La codificación de la variable "Realizó prácticas" se codifica de manera similar:

4. En la celda "G2" ingresamos el condicional para codificar la variable "Realizó prácticas" según lo descrito líneas arriba:

$$= SI(I2 = "No", 0,1)$$

5. Repetir la fórmula para cada uno de los casos de la base de datos.

Con las variables codificadas se realiza la regresión como se ha venido haciendo anteriormente. En este caso se selecciona el rango de Y que, en este caso, es la variable "Sueldo" y el rango de X que, en este caso, está comprendido por las variables "Edad", "Sueldo pasado", "Promedio de notas", "Género" y "Realizó prácticas". El resultado se presenta a continuación:

Figura 10-9: Estadísticos de regresión

Resumen

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.962033819
Coeficiente de determinación R^2	0.925509068
R^2 ajustado	0.922060414
Error típico	207.4867943
Observaciones	114

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	5	57767283.88	11553456.78	268.3681809	3.41947E-59
Residuos	108	4649483.138	43050.76979		
Total	113	62416767.02			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
Intercepción	-2493.130202	512.4198846	-4.86540487	3.91372E-06	-3508.835309	-1477.425094	-3508.835309	-1477.425094
Edad	-14.14798129	14.71846423	-0.961240321	0.338578526	-43.32253081	15.02656823	-43.32253081	15.02656823
Sueldo Pasado	0.529956086	0.079219839	6.689689029	1.02647E-09	0.372928623	0.68698355	0.372928623	0.68698355
Promedio de notas de la Carrera	295.6581354	31.58218719	9.361547178	1.31888E-15	233.0567614	358.2595093	233.0567614	358.2595093
G	-21.2955904	39.93713476	-0.533227797	0.594971516	-100.4579199	57.86673912	-100.4579199	57.86673912
Rp	176.625792	76.36038897	2.313055163	0.022614195	25.26625377	327.9853302	25.26625377	327.9853302

Se observa que el R^2 alcanza el valor de 0.92 de lo que se desprende que las variaciones de las variables independientes explican a la dependiente en 92 %, es decir, el modelo explica bastante bien las variaciones del sueldo.

En lo que respecta a la significancia individual, se plantean las siguientes hipótesis nula y alternativa:

$$H_o: \beta = 0$$
$$H_1: \beta \neq 0$$

Se verifican estas hipótesis con el valor de las probabilidades de cada variable. Estas deben ser menores al nivel de significancia (por lo general es 0.05) para rechazar la hipótesis nula.

La variable edad tiene una probabilidad de 0.33, por lo que no cuenta con significancia individual, es decir, no se puede rechazar H_o. De la misma manera, la variable "G" tiene una probabilidad de 0.59, por lo que no cuenta con significancia individual, es decir, no se puede rechazar H_o. Esto significa que ambas variables no deberían formar parte del modelo.

Las demás variables tienen una probabilidad inferior a 0.05, es decir, en todas ellas se rechaza la H_o, por lo que sí son significativas individualmente.

En lo que respecta a significancia conjunta, se tienen las siguientes hipótesis:

$$H_0: \beta_1 = \beta_2 = 0$$
$$H_1: Al\ menos\ uno\ de\ los\ coeficientes\ es\ diferente\ de\ cero$$

Se verifica con el valor de la probabilidad, el cual debe ser menor al nivel de significancia (por lo general es 0.05) que la probabilidad asociada al estadístico F es inferior a 0.05, por lo que el modelo cuenta con significancia conjunta.

Por último, realizamos la regresión sin incluir las variables "Edad" y "Género" ante lo cual se obtiene lo siguiente:

Figura 10-10: Estadísticos de regresión

Resumen

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.961506588
Coeficiente de determinación R^2	0.924494919
R^2 ajustado	0.922435689
Error típico	206.986672
Observaciones	114

ANÁLISIS DE VARIANZA

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	3	57703983.95	19234661.32	448.9518649	1.57593E-61
Residuos	110	4712783.063	42843.48239		
Total	113	62416767.02			

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
Intercepción	-2800.948233	410.7584195	-6.818967306	5.17152E-10	-3614.975008	-1986.921458	-3614.975008	-1986.921458
Sueldo Pasado	0.535754982	0.078192928	6.851706355	4.40579E-10	0.380794952	0.690715012	0.380794952	0.690715012
Promedio de notas de la Carrera	292.4807016	31.26924233	9.353622916	1.18546E-15	230.5124028	354.4490005	230.5124028	354.4490005
Rp	179.2236108	76.10000176	2.355106526	0.020289869	28.41126931	330.0359522	28.41126931	330.0359522

Ahora se observa que todas las variables tienen significancia individual y, a su vez, que el modelo cuenta con significancia conjunta.

i) El modelo econométrico planteado es el siguiente:

$$Sueldo = \beta_1 + \beta_2(Sueldo\ anterior) + \beta_3(Promedio\ de\ notas) + \beta_6(Rp) + \mu$$

Modelo estimado:

$$Sueldo = -2800.94 + 0.5357 * (Sueldo\ anterior) + 292.48 * (Promedio\ de\ notas) + 179.22(Rp)$$

j) Si utilizamos un *software* econométrico como el Eviews, tendremos mayor información para poder seleccionar el mejor modelo, como por ejemplo los criterios de Akaike, Schwarz y Hannan-Quin, cuyo criterio de selección es elegir el menor entre los modelos seleccionados, a diferencia del R2 ajustado el cual debe de ser el mayor.

Tabla 10-4: Criterios de Akaike, Schwarz y Hannan-Quin en Eviews

Criterio	Modelo 01	Modelo 02	Modelo 03	Modelo 04
Akaike	13.53	14.35	14.10	14.79
Schwarz	13.63	14.40	14.17	14.83
Hannan-Quin	13.57	14.37	14.13	14.81
R^2 ajustado	0.92	0.82	0.86	0.72

De la anterior tabla se desprende que el mejor modelo es el Modelo 01, el cual es el mismo que se calculó en la pregunta "i".

k) Interpretación de coeficientes:

Sueldo pasado: 0.5357

La relación entre el sueldo pasado y el sueldo actual es directa, siendo que su coeficiente es 0.5357. Es decir, por cada unidad adicional en el sueldo pasado, se espera que el sueldo actual aumente en 0.5357 unidades.

Promedio de notas de la carrera: 292.48

La relación entre el promedio de notas de la carrera y el sueldo actual es directa, siendo que su coeficiente es 292.48. Es decir, por cada unidad adicional en el promedio de notas de la carrera, se espera que el sueldo actual aumente en 292.48 unidades.

Realizó prácticas: 179.22
La interpretación de una variable va depender de la codificación, que en este caso fue:

- 0 cuando la persona no realizó prácticas;
- 1 cuando la persona si realizó prácticas.

Si la persona realizó prácticas, se espera que tenga un sueldo promedio mayor en 179.22 unidades respecto al sueldo promedio de una persona que no realizó prácticas.

Lecturas recomendadas

- Anderson, D., Sweeney, D., y Williams, T. (2008). *Estadística para administración y economía* (10.ª ed.; cap. 14, 14.1-14.7; cap. 15, 15.1-15.6). Cengage Learning Editores.
- Levin, R., y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed. revisada; cap. 12, 12.2; cap. 13, 13.1-13.4). México: Pearson Educación de México; Prentice Hall.

11. Regresión con variables dicotómicas y análisis residual de la estimación

El presente apartado tiene como objetivo complejizar el modelo de regresión lineal a través de la introducción de variables cualitativas como exógenas. Además, se repasarán los principales supuestos que debe cumplir toda regresión lineal para garantizar que el modelo se ha llevado a cabo de manera adecuada. Los objetivos son los siguientes:

- Entender el concepto de "variable ficticia" o "*dummy*": cómo se convierte una variable cualitativa y cómo se usa dentro de un modelo de regresión lineal.
- Interpretar adecuadamente los coeficientes de una variable *dummy*.
- Interiorizar los requisitos que se deben cumplir para que una regresión lineal sea confiable.
- Aprender las distintas pruebas que existen para comprobar que se cumplen los supuestos de una regresión.

Palabras clave:

- Regresión lineal múltiple
- Variables ficticias, variable *dummy*
- Supuestos de una regresión, normalidad, linealidad, homocedasticidad, no colinealidad, independencia.

Ejercicio I: Comparación de notas finales utilizando variables dicotómicas

Luego de evaluar diversos exámenes de aptitud para aplicar a una beca de maestría, una universidad obtuvo la siguiente función de regresión múltiple para predecir las notas:

$$Nota\ Final = 5.2 - 0.09(Edad) + 2.33(Tipo\ de\ universidad) - 1.78(Sexo) + 0.73(Nota\ Promedio\ de\ egresado) - 1.5(Norte) - 1.2\ (Sur) + e$$

En donde "Nota final" es el puntaje obtenido en escala vigesimal en el examen de admisión a la maestría, "Edad" es el número de años cumplidos del postulante, "Tipo de universidad" es una variable *dummy* cuyo valor es 1 si la universidad de procedencia es privada y 0 si es pública, "Sexo" es 1 si es hombre y 0 si es mujer, "Nota promedio de egresado" es el promedio ponderado o nota final que obtuvo el postulante al egresar del pregrado, y "Procedencia" es una variable nominal de tres categorías para señalar si el postulante viene de la capital, del norte o del sur del país (se asumió

"Capital" como referencia y, por ser tres categorías, se usó dos variables *dummy*: "Norte" y "Sur").

Encuentre la diferencia esperada de puntajes entre un hombre de 27 años egresado de una universidad pública del norte con promedio de egreso 19, con una mujer de 21 años egresada de una universidad privada de la capital con promedio de egreso 13; y la diferencia esperada entre dos mujeres postulantes con las mismas características a excepción de que una es del norte y la otra del sur.

Ejercicio II: Análisis de espuriedad y colinealidad

El Ministerio de Producción (PRODUCE) realizó una investigación sobre los factores que pueden predecir el ingreso en el sector agrícola. Para ello, consideró como variables explicativas el "dominio geográfico" (costa, sierra y selva), el "nivel educativo" (sin estudios, con educación pero sin secundaria completa, con secundaria completa y con estudios superiores), el "sexo" de la persona, su "edad" y el tiempo que lleva en su actual empleo.

Una vez que tuvieron los resultados, uno de los consultores sugirió revisar el criterio de espuriedad con el fin de descartar que se haya asumido una falsa relación entre las variables independientes y la dependiente, así como el criterio de no colinealidad. Así, con el programa SPSS, se obtuvo la siguiente tabla de correlaciones y estadísticas de colinealidad:

Tabla 11-1: Correlaciones y estadísticas de colinealidad

	Correlaciones			Estadísticas de colinealidad	
	Orden cero	Parcial	Parte	Tolerancia	VIF
(Constante)					
Sierra	-0.081	-0.144	-0.129	0.711	1.407
Selva	-0.004	-0.099	-0.088	0.764	1.308
Con educación pero sin secundaria completa	-0.191	0.090	0.081	0.300	3.334
Secundaria completa	-0.021	0.132	0.119	0.381	2.625
Estudios superiores	0.348	0.304	0.283	0.351	2.851
Hombre	-0.101	-0.152	-0.137	0.973	9.028
Edad	0.176	0.065	0.058	0.676	1.480
Tiempo en empleo actual (meses)	0.230	0.141	0.126	0.676	10.479

Señale si se cumplen o no los dos criterios mencionados.

Caso aplicado: Efectos de la publicidad en las ventas[14]

Para este ejercicio, emplee el archivo en Excel "BD_11".

La empresa ABC se dedica a la venta de detergente para lavavajillas, y durante el último reporte financiero anual ha evidenciado un retroceso en sus utilidades netas explicado por el incremento significativo en los gastos de publicidad en comparación al incremento de las ventas. Por esa razón, el directorio de la compañía ha solicitado a la Gerencia Financiera que evalúe reducir los gastos en publicidad.

Preocupada por el tema, la Gerencia Financiera solicita la siguiente información al área de *Business Intelligence*: ventas en unidades durante los últimos 100 meses, precio de venta del producto durante el mismo periodo, precio de venta del principal competidor, el monto de gasto en publicidad, y el medio predominante de la publicidad en cada mes (radio, TV o medios impresos).

Adicionalmente, con el fin de considerar el impacto de variables macroeconómicas en su investigación, la Gerencia Financiera decide considerar a la demanda interna del país y al producto bruto interno registrados durante los últimos 100 meses.

Con la información disponible, la Gerencia Financiera plantea un modelo econométrico lineal multivariado donde la variable dependiente son las ventas, y las variables explicativas o independientes son todas las variables solicitadas al área de *Business Intelligence*, así como las variables macroeconómicas anteriormente señaladas.

El objetivo es evidenciar que el impacto del gasto de publicidad en las ventas es positivo, y que los resultados negativos en las ventas en el año del reporte financiero se han debido a causas diferentes que han ralentizado las ventas.

No obstante, antes de evaluar los resultados, se requiere que no haya multicolinealidad en el modelo, siendo que este problema generaría que los parámetros del modelo no se logren identificar, y que se evidencie un problema econométrico denominado "inflación de varianza".

---

[14] Caso adaptado del documento "Estimación de los efectos de la publicidad en las ventas. Un análisis empírico en España y Alemania", de Joaquín Sanchez Herrera, Teresa Pintado Blanco, María Avello Iturriagagoitia y Carmen Abril Barrie (2011). El documento puede ser visualizado en http://adresearch.esic.edu/files/2011/03/Art%C3%ADculo4.pdf

Asimismo, un econometrista del área de Estudios Económicos de la compañía recomienda que para validar la significancia individual de las variables explicativas y del modelo, se debe verificar la ausencia de autocorrelación y la presencia de homocedasticidad en los errores del modelo.

En base a lo señalado, siendo usted miembro de la Gerencia Financiera, se le pide elaborar un reporte econométrico que sirva de sustento al informe para la alta dirección. En el reporte econométrico deberá responder las siguientes preguntas:

a) ¿Se evidencia multicolinealidad en el modelo? ¿Cómo se corrige el problema de multicolinealidad?

b) ¿Se puede señalar que los errores del modelo tienen distribución normal?

c) ¿Se evidencia presencia de autocorrelación o heterocedasticidad en el modelo?

d) ¿Es bueno el ajuste del modelo?

Finalmente, teniendo en cuenta el modelo estimado, responda las siguientes preguntas:

e) Interprete los parámetros estimados del modelo.

f) ¿Recomendaría reducir los gastos de publicidad?

g) ¿Qué recomendaría para incrementar las ventas de la compañía?

Caso aplicado: Del cole a tu primera chamba[15]

Para este ejercicio, emplee el archivo en Excel "BD_11".

La ONG "Alternative World" tiene como misión generar programas que ayuden a combatir la pobreza; dado que el fenómeno tiene múltiples manifestaciones, en esta ocasión ha decidido enfocarse en los bajos niveles de empleabilidad de los más jóvenes en sectores marginales urbanos. El grupo de expertos contratados por la ONG propone "Del cole a tu primera chamba", proyecto que pretende impulsar el desarrollo de habilidades psicosociales y capacidades técnicas en jóvenes de educación secundaria de instituciones educativas de dos distritos del cono sur de Lima Metropolitana: Villa El Salvador y Villa María del Triunfo. Se identificó que uno de los problemas más importantes en dichas zonas es la dificultad que tiene la población juvenil egresada de colegios públicos para ingresar al mercado laboral.

[15] El presente caso es la adaptación de una consultoría real hecha por Marie Fazzari, Diego Mendoza, Christian Trujillo, Norma Vergara y Katherine Zegarra (de la Facultad de Ciencias Sociales) para una ONG que trabaja temas de infancia y pobreza.

Bajo los enfoques de desarrollo humano y el de capacidades, se centró el foco del estudio sobre el factor humano y la apuesta por promover que los jóvenes participen activamente en la construcción de sus propios futuros e identificar los problemas estructurales que dificultan su desarrollo personal, escolar y que limitan sus oportunidades laborales.

En primer lugar, se planteó un estudio de línea base para observar el comportamiento de los salarios de los egresados de los colegios en ambos distritos. Para ello se tomó en cuenta los siguientes resultados:

Tabla 11-2: Resultados

Coeficiente de correlación múltiple	0.73706424
Coeficiente de determinación R^2	0.5432637
R^2 ajustado	0.53565143
Error típico	242.985204
Observaciones	367

ANÁLISIS DE VARIANZA	Grados de libertad	Suma de cuadrados
Regresión	6	25281760.6

La teoría de capital humano señala que la integración en el mercado laboral no depende exclusivamente de la capacidad de un individuo de desempeñar tareas específicas, sino que el valor de su trabajo se eleva en tanto sea capaz de poder resolver situaciones adicionales, no convencionales o retadoras; dicha habilidad se encuentra asociada a los años de educación, los que generalmente proveen de conocimiento, pero también de una serie de estrategias sociales y psicológicas. Así, la investigación esperaba que la nota final promedio, como indicador aproximado de haber adquirido estas habilidades tuviera un rol fundamental en la explicación del salario del egresado de secundaria.

Sin embargo, dada la estructura particular de mercado a la que se insertan los jóvenes, se creyó conveniente tomar en cuenta la situación de formalidad o informalidad en la que se encuentran trabajando, los años que han estado fuera del colegio, el sexo de los egresados y si tuvieron experiencias previas de trabajo mientras estaban en el colegio[16]. Adicionalmente, los investigadores creyeron que podría existir diferencia entre los egresados de 4 colegios.

[16] Variable que decidió incluirse después de tener entrevistas con los docentes y que mencionaran como situación recurrente. La experiencia previa puede fácilmente categorizarse dentro del conocimiento práctico y de cómo funciona un oficio

El marco muestral conformado por las 8172 nóminas de los egresados en los últimos 5 años de 4 colegios fue usado para determinar una muestra de 367 alumnos (ver base de datos proporcionada para esta unidad).

a) Generar un modelo que permita predecir el salario basado en las variables de importancia para el proyecto. Evaluar su significancia y el porcentaje de explicación de la varianza.
b) A partir de la interpretación de las variables, ofrecer una conclusión inicial para el proyecto.

Los investigadores quedaron bastante satisfechos con los resultados. Sin embargo, notaron que las condiciones de los dos colegios de Villa El Salvador ("A" y "B") eran notablemente distintas que las de los dos colegios de Villa María del Triunfo ("C" y "D"). Por ello, decidieron probar una nueva codificación de la variable "Colegio" considerando una variable *dummy* que diferencia la pertenencia de los colegios a Villa El Salvador y a Villa María del Triunfo.

c) Generar un nuevo modelo con la variable "Colegio" corregida. Evaluar la pertinencia del modelo y reconstruir la función.
d) Corroborar que el modelo cumpla con los criterios de normalidad de residuos.

Solucionario

Ejercicio I: Comparación de notas finales utilizando variables dicotómicas

(Continúa del enunciado del Ejercicio I del Capítulo 11)

La primera parte de lo planteado en este ejercicio ofrece dos casos concretos y totalmente distintos, por lo que se debe remplazar las variables con los datos ofrecidos:

$$NF1 = 5.2 - 0.09 \times (27) + 2.33 \times (0) - 1.78 \times (1) + 0.73 \times (19) - 1.5 \times (1) - 1.2 \times (0)$$
$$NF1 = 13.36$$

$$NF2 = 5.2 - 0.09 \times (21) + 2.33 \times (1) - 1.78 \times (0) + 0.73 \times (13) - 1.5 \times (0) - 1.2 \times (0)$$
$$NF2 = 15.13$$

Se observa, entonces, que la diferencia de puntajes es de 1.77.

Para la segunda parte, se señala que son dos mujeres de igual edad y con la misma nota promedio, por ello solo se tomará en cuenta las variables que cambian para encontrar la diferencia de puntajes. Es importante hacer notar que no es necesario saber cuánto de nota sacó cada una, ya que como las dos sacaron la misma nota no es necesario a la hora de hacer la comparación. Entonces:

$$M1 = -1.5 \times (1) - 1.2 \times (0)$$

$$M1 = -1.5$$

$$M2 = -1.5 \times (0) - 1.2 \times (1)$$
$$M2 = -1.2$$

La mujer 1 (del norte) tiene 1.5 puntos; en promedio, menos que alguien de la capital. Por otro lado, la mujer 2 (del sur) tiene 1.2 puntos menos en promedio. Eso quiere decir que, entre ellas, la diferencia es de 0.3. Por lo tanto, la mujer 1(M1) tiene 0.3 puntos menos que la mujer 2 (M2).

Ejercicio II: Análisis de espuriedad y colinealidad

(Continúa del enunciado del Ejercicio II del Capítulo 11)

Los análisis de espuriedad y colinealidad se aplican a variables cuantitativas continuas. En ese sentido, se analizarán los estadísticos relacionados a las variables de este tipo: edad y tiempo en empleo actual.

Para saber si existe espuriedad, es decir, falsa asociación, las correlaciones encontradas no deben tener un orden descendente. En otras palabras, los números no deben experimentar dos caídas consecutivas. Para el criterio de no colinealidad se observa el factor de inflación de la varianza (VIF, por sus siglas en inglés). Si va de 1 a 5, no hay multicolinealidad; de 5 a 10, existen problemas leves de multicolinealidad; mayor a 10, implica un grave problema de multicolinealidad.

A continuación, se muestra una tabla de resumen con todas las variables en cuestión:

Tabla 11-3: Resumen de variables

	Correlaciones				Estadísticas de colinealidad		
	Orden cero	Parcial	Parte	Espuriedad	Tolerancia	VIF	Multicolinealidad
Edad	0.176	0.065	0.058	Sí	0.676	1.480	Sí, grave
Tiempo en empleo actual (meses)	0.230	0.141	0.126	No	0.676	10.479	No

Caso aplicado: Efectos de la publicidad en las ventas

Para este ejercicio, emplee el archivo SPSS "SOL_11".

a) ¿Se evidencia multicolinealidad en el modelo? ¿Cómo se corrige el problema de multicolinealidad?

Un análisis preliminar de la presencia de multicolinealidad es la matriz de correlaciones entre las variables explicativas. Al respecto, en la matriz

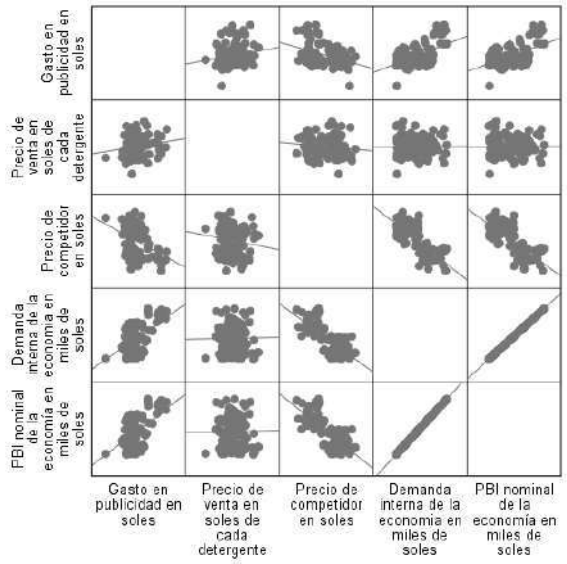
consideramos a las variables cuantitativas no dicotómicas, siendo que las variables TV y Radio toman solo los valores de 0 y 1.

Al respecto, utilizamos la matriz de diagramas de dispersión con los siguientes comandos en SPSS.

En SPSS:

- **Prueba:**
  Gráficos > Dispersión puntos > Dispersión matricial.

A partir de los diagramas de dispersión, observamos cierta relación lineal entre las variables explicativas, destacando la correlación entre el PBI y la demanda interna.



Un análisis formal de la presencia de multicolinealidad es el uso de números índices y de las estadísticas de colinealidad: "Tolerancia" y "Factor de inflación de la varianza" (VIF).

[11-1] Fórmula para Tolerancia

$$Tolerancia=1-Rj2$$

[11-2] Fórmula para factor de inflación de la varianza

$$FIV=1/1-Rj2$$

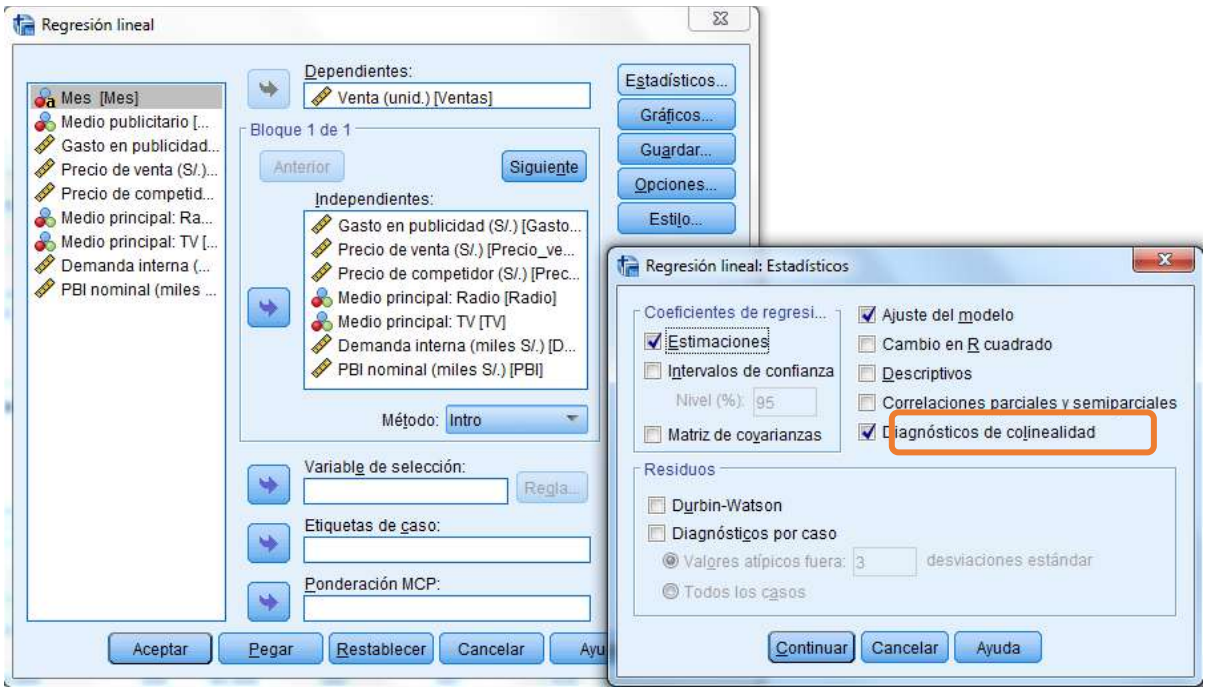
[11-3] Fórmula para índice de condición

$$\text{Índice de condición}=\sqrt{\frac{\lambda max}{\lambda i}}$$

Entonces, para empezar con el análisis de este ejercicio se usarán los siguientes comandos en SPSS.

En SPSS:

- **Multicolinealidad:**
  Analizar > Regresión lineal > Ingreso de variables dependiente e independientes > "Estadísticos" > *Check* en diagnósticos de colinealidad.



Coeficientes[a]

Modelo	Coeficientes no estandarizados	Coeficientes estandarizados	t	Sig.	Estadísticas de colinealidad

		B	Desv. Error	Beta			Tolerancia	VIF
1	(Constante)	2207.199	181.401		12.168	.000		
	Gasto en publicidad (S/)	10.023	.032	.502	311.180	.000	.646	1.547
	Precio de venta (S/)	-2222.315	17.057	-.176	-130.284	.000	.922	1.085
	Precio de competidor (S/)	1797.595	3.614	1.042	497.351	.000	.383	2.611
	Medio principal: Radio	796.073	18.064	.068	44.069	.000	.697	1.435
	Medio principal: TV	1000.624	11.404	.121	87.742	.000	.879	1.137
	PBI nominal (miles S/)	.000	.000	.022	8.775	.000	.263	3.808

a. Variable dependiente: Venta (unid.)

De la siguiente tabla, se tiene que SPSS identifica el problema de multicolinealidad severa que genera la variable "Demanda interna", por lo que la excluye directamente del modelo. Esto no ocurre siempre y puede requerir exclusión manual para así corregir el problema de multicolinealidad. El resto de variables no genera problemas de multicolinealidad (VIF < 5).

Variables excluidas[a]

Modelo		En beta	t	Sig.	Correlación parcial	Estadísticas de colinealidad		
						Tolerancia	VIF	Tolerancia mínima
1	Demanda interna (miles S/)	-.225[b]	-1.090	.279	-.113	3.942E-5	25365.212	3.942E-5

a. Variable dependiente: Venta (unid.)

b. Predictores en el modelo: (Constante), PBI nominal (miles S/), Precio de venta (S/), Medio principal: TV, Medio principal: Radio, Gasto en publicidad (S/), Precio de competidor (S/)

b) ¿Se puede señalar que los errores del modelo tienen distribución normal?

Para este ejercicio se requiere la creación de variables: residuos y pronósticos. Para esto usaremos los siguientes comandos en SPSS.

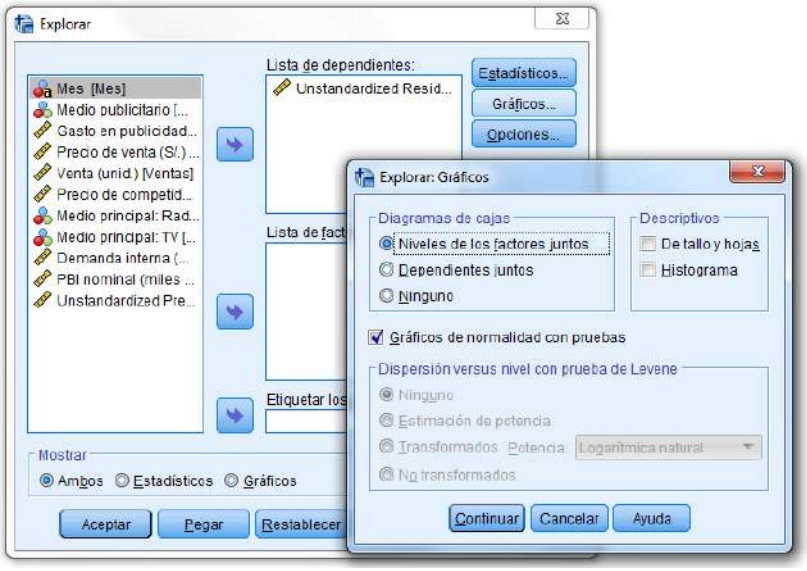
En SPSS:

- **Distribución normal de errores:**
  Analizar > Regresión lineal > Ingreso de variables dependiente e independientes > "Guardar" > *Check* en "No estandarizados" tanto para valores pronosticados como para residuos.

Adicionalmente, para evaluar la normalidad de los residuos, se recurrirá a las pruebas de SPSS mediante los siguientes comandos.

En SPSS:

- **Distribución normal de errores:**
  Analizar > Estadísticos descriptivos > Explorar > Seleccionamos a Residuos estandarizados > *Check* en gráficos de normalidad con pruebas.



Los residuos de un modelo se componen por la diferencia entre el valor real de la variable dependiente y el valor pronosticado por el modelo para la variable dependiente. En un modelo adecuado, los residuos deben tener una distribución normal. Para esta base se utiliza la prueba K-S (n > 30) para verificar normalidad, en donde:

H0: Normalidad

H1: No normalidad

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Unstandardized residual	.084	100	.078	.979	100	.112

a. Corrección de significación de Lilliefors

De la tabla anterior se tiene que utilizando la prueba de Kolgomorov-Smirnov (n>30), no se puede rechazar la distribución normal de los errores del modelo al 5 % de nivel de significancia.

c) ¿Se evidencia presencia de autocorrelación o heterocedasticidad en el modelo?

La homocedasticidad implica que la variabilidad de los errores se mantenga constante a lo largo del valor estimado por el modelo.

Prueba de hipótesis:

H0: Homocedasticidad

H1: Heterocedasticidad

[11-3] Pruebas para heterocedasticidad

White: Todo tipo de heterocedasticidad

Breusch-Pagan: Heterocedasticidad lineal

En SPSS:

- **Heterocedasticidad:**
  Analizar > Modelo lineal general > Univariado > Ingreso de variables dependiente y covariables > "Opciones" > *Check* en "Test de Breusch-Pagan" y "Prueba de White".

Prueba de White para heterocedasticidad[a,b,c]

Chi-cuadrado	gl	Sig.
29.264	24	.210

a. Variable dependiente: Venta (unid.).

b. Prueba la hipótesis nula de que la varianza de los errores no depende de los valores de las variables independientes.

c. Diseño : Intersección + Gasto_publicidad + Precio_venta + Precio_competidor + PBI + Radio + TV + Gasto_publicidad * Gasto_publicidad + Gasto_publicidad * Precio_venta + Gasto_publicidad * Precio_competidor + Gasto_publicidad * PBI + Gasto_publicidad * Radio + Gasto_publicidad * TV + Precio_venta * Precio_venta + Precio_venta * Precio_competidor + Precio_venta * PBI + Precio_venta * Radio + Precio_venta * TV + Precio_competidor * Precio_competidor + Precio_competidor * PBI + Precio_competidor * Radio + Precio_competidor * TV + PBI * PBI + PBI * Radio + PBI * TV + Radio * Radio + Radio * TV + TV * TV

Test de Breusch-Pagan para heterocedasticidad[a,b,c]

Chi-cuadrado	gl	Sig.
.327	1	.568

a. Variable dependiente: Venta (unid.)

b. Prueba la hipótesis nula de que la varianza de los errores no depende de los valores de las variables independientes.

c. Valores pronosticados a partir del diseño: Intersección + Gasto_publicidad + Precio_venta + Precio_competidor + PBI + Radio + TV

Ahora para evaluar la autocorrelación de residuos (Durbin-Watson), se requiere ordenar la base de datos de acuerdo con la variable temporal mes. Para esto se usan los siguientes comandos en SPSS.

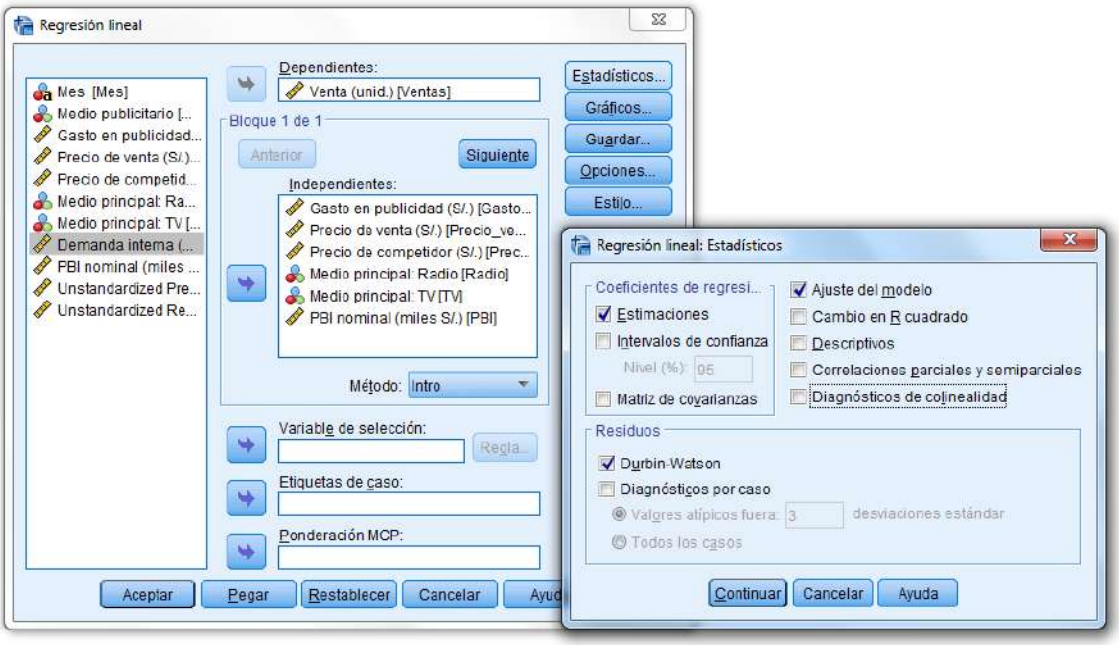
En SPSS:

- **Ordenar la base:**
  Vista de datos > Clic derecho > Ordenar de forma ascendente

Luego, para realizar la prueba de Durbin-Watson, se usan los siguientes comandos en SPSS.

En SPSS:

- **Prueba:**
  Analizar > Regresión lineal > Ingreso de variables dependiente e independientes > "Estadísticos" > *Check* en Durbin-Watson > Continuar.



Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Durbin-Watson
1	1.000[a]	1.000	1.000	49.00958	1.664

a. Predictores: (Constante), PBI nominal (miles S/), Precio de venta (S/), Medio principal: TV, Medio principal: Radio, Gasto en publicidad (S/), Precio de competidor (S/)

b. Variable dependiente: Venta (unid.)

La autocorrelación de residuos solo puede analizarse cuando la base de datos es longitudinal. Lo que se busca evaluar es una posible estacionalidad en la distribución de los residuos.

[11-4] Durbin-Watson

Durbin-Watson: Valores aceptados entre 1.5 y 2.5

Cercano a 0 - autocorrelación positiva

Cercano a 2 - no existe autocorrelación

Cercano a 4 - autocorrelación negativa

d) ¿Es bueno el ajuste del modelo?

En SPSS:

- **Regresión Lineal:**
  Analizar > Regresión lineal.

R-cuadrado (Bondad de ajuste):

El modelo puede explicar aproximadamente el 100 % del comportamiento de la variable dependiente (ventas).

Resumen del modelo[b]

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Durbin-Watson

1	1.000[a]	1.000	1.000	49.00958	1.664

a. Predictores: (Constante), PBI nominal (miles S/), Precio de venta (S/), Medio principal: TV, Medio principal: Radio, Gasto en publicidad (S/), Precio de competidor (S/)

b. Variable dependiente: Venta (unid.)

Significancia Global:

El modelo funciona en su conjunto, permite predecir el comportamiento de la variable dependiente (p-value < alfa).

ANOVA[a]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	1428726429.362	6	238121071.560	99137.027	.000[b]
	Residuo	223380.309	93	2401.939		
	Total	1428949809.671	99			

a. Variable dependiente: Venta (unid.)

b. Predictores: (Constante), PBI nominal (miles S/), Precio de venta (S/), Medio principal: TV, Medio principal: Radio, Gasto en publicidad (S/), Precio de competidor (S/)

Significancia individual:

Todas las variables del modelo son significativas de forma individual, permiten evaluar el efecto que tiene cada variable independiente sobre la variable dependiente (ventas).

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Desv. Error	Beta	t	Sig.
1	(Constante)	2207.199	181.401		12.168	.000

	Gasto en publicidad (S/)	10.023	.032	.502	311.180	.000
	Precio de venta (S/)	-2222.315	17.057	-.176	-130.284	.000
	Precio de competidor (S/)	1797.595	3.614	1.042	497.351	.000
	Medio principal: Radio	796.073	18.064	.068	44.069	.000
	Medio principal: TV	1000.624	11.404	.121	87.742	.000
	PBI nominal (miles S/)	.000	.000	.022	8.775	.000

a. Variable dependiente: Venta (unid.)

La evaluación de la significancia global, la significancia individual y los coeficientes solo es confiable una vez que se han validado los supuestos anteriores. De lo contrario, puede haber distorsión en estos datos.

e) Interprete los parámetros estimados del modelo.

De la pregunta anterior se tiene:

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Desv. Error	Beta	t	Sig.
1	(Constante)	2207.199	181.401		12.168	.000
	Gasto en publicidad (S/)	10.023	.032	.502	311.180	.000
	Precio de venta (S/)	-2222.315	17.057	-.176	-130.284	.000
	Precio de competidor (S/)	1797.595	3.614	1.042	497.351	.000
	Medio principal: Radio	796.073	18.064	.068	44.069	.000
	Medio principal: TV	1000.624	11.404	.121	87.742	.000

PBI nominal (miles S/)	.000454	.000	.022	8.775	.000

a. Variable dependiente: Venta (unid.)

Modelo matemático:

Ventas = B0 + B1 (Gasto publicidad) + B2 (Precio venta) + B3 (Precio competidor) + B4 (Medio principal de publicidad) + B5 (PBI Nominal)

Modelo estimado:

Ventas (estimadas) = 2207.10 + 10.02 (Gasto publicidad) – 2222.31 (Precio venta) + 1797.60 (Precio competidor) + 796.07 (Medio principal: Radio) + 1000.62 (Medio principal: TV) + 0.000454 (PBI Nominal)

Coeficientes no estandarizados:

- Por cada sol adicional de gasto en publicidad, las ventas se incrementan en 10 unidades.
- Cuando el precio de venta se incrementa un sol, las ventas disminuyen en 2222 unidades.
- Cuando el precio del competidor se incrementa en un sol, las ventas se incrementan en 1798 unidades.
- Cuando la publicidad se hace por radio, las ventas son mayores en 796 unidades respecto a la modalidad de medios impresos.
- Cuando la publicidad se hace por TV, las ventas son mayores en 1000 unidades respecto a la modalidad de medios impresos.
- Cuando el PBI se incrementa en 100 millones de soles, las ventas se incrementan en 45 unidades, por lo que la actividad económica impacta positivamente las ventas.
- Si el valor de todas las variables cuantitativas fuese cero, y el medio principal de difusión fuese impreso (valor base), la cantidad de ventas sería 2207.20 (constante). Sin embargo, interpretar la constante (B0) puede llevar a situaciones incoherentes. Por ejemplo, la mayoría de las variables independientes del modelo nunca tomarían el valor CERO en la vida real.

f) ¿Recomendaría reducir los gastos de publicidad?

De la pregunta anterior:

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Desv. Error	Beta	t	Sig.
1	(Constante)	2207.199	181.401		12.168	.000
	Gasto en publicidad (S/)	10.023	.032	.502	311.180	.000
	Precio de venta (S/)	-2222.315	17.057	-.176	-130.284	.000
	Precio de competidor (S/)	1797.595	3.614	1.042	497.351	.000
	Medio principal: Radio	796.073	18.064	.068	44.069	.000
	Medio principal: TV	1000.624	11.404	.121	87.742	.000
	PBI nominal (miles S/)	.000454	.000	.022	8.775	.000

a. Variable dependiente: Venta (unid.)

Coeficientes estandarizados:

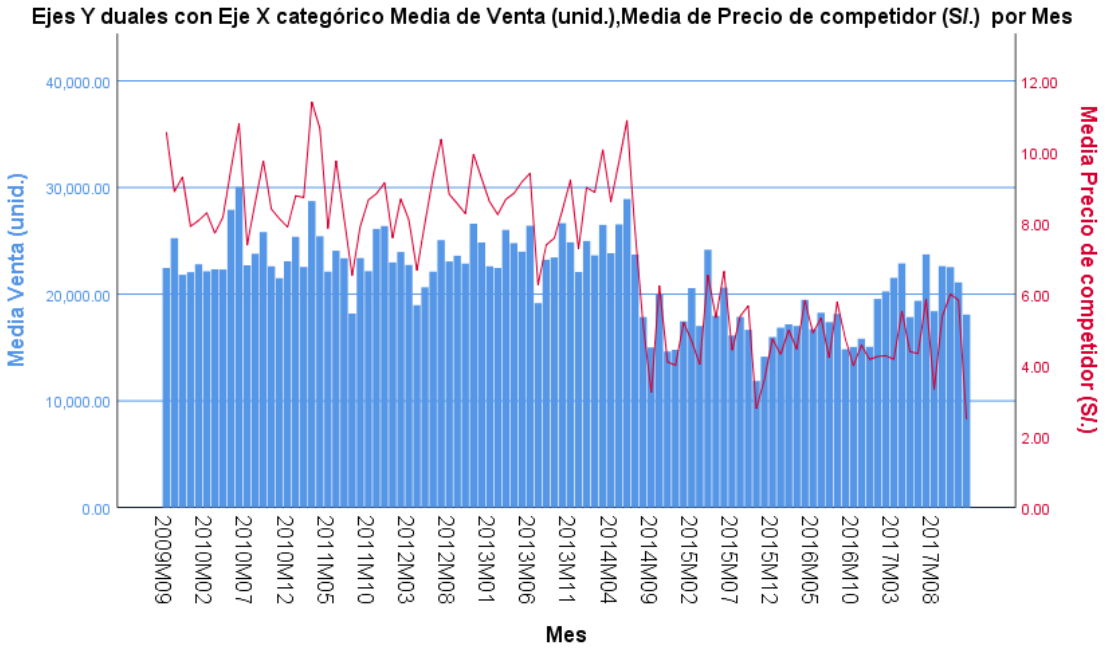
- El gasto en publicidad tiene un efecto positivo sobre las ventas (tiene el segundo efecto más importante dentro del modelo), por lo que recomendaría no recortar dicho gasto.
- El precio de la competencia es la variable más relevante sobre la dinámica de las ventas.

g) ¿Qué recomendaría para incrementar las ventas de la compañía?

La disminución de las ventas se comporta de forma muy relacionada con la disminución del precio de la competencia, por lo que cualquier variación en el resto de variables debe articularse con una política de nivelación de precios para poder competir.

En SPSS:

- **Ejes dobles:**
  Gráficos > Generador de gráficos > Ejes dobles.



Caso aplicado: Del cole a tu primera chamba[17]

Para este ejercicio, emplee el archivo Excel "SOL_11".

(Continúa del enunciado del caso aplicado del Capítulo 11)

a) Generar un modelo que permita predecir el salario basado en las variables de importancia para el proyecto. Evaluar su significancia y el porcentaje de explicación de la varianza.

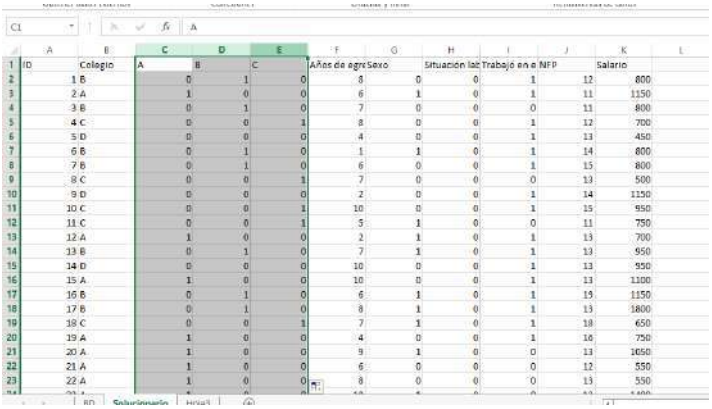
Solución:

Antes de iniciar con la regresión, se debe corroborar que todas las variables sean cuantitativas. Para el caso de las independientes que son de carácter cualitativo, se debe recodificar de tal forma que una variable quede como "0" y "1". En el presente caso, todas las variables cualitativas presentan esta recodificación con excepción de "Colegio", cuyas categorías son "A", "B", "C"

[17] El presente caso es la adaptación de una consultoría real hecha por Marie Fazzari, Diego Mendoza, Christian Trujillo, Norma Vergara y Katherine Zegarra (de la Facultad de Ciencias Sociales) para una ONG que trabaja temas de infancia y pobreza.

y "D", es decir, cuatro categorías que necesitan tres variables *dummy*. Tomando las primeras tres como referencia, la codificación debería quedar así:

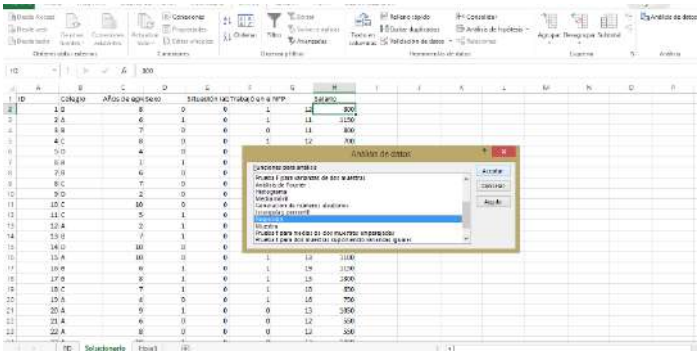
Figura 11-1: Las *dummy* A, B y C



ID	Colegio	A	B	C	Años de agre	Sexo		Situación la	Trabajó en e	NFP	Salario
1	B	0	1	0	8	0	0	1	12		800
2	A	1	0	0	6	1	0	1	11		1150
3	B	0	1	0	7	0	0	0	11		800
4	C	0	0	1	8	0	0	1	12		700
5	D	0	0	0	4	0	0	1	13		450
6	B	0	1	0	1	1	0	1	14		800
7	B	0	1	0	6	0	0	1	15		800
8	C	0	0	1	7	0	0	0	13		500
9	D	0	0	0	2	0	0	1	14		1150
10	C	0	0	1	10	0	0	1	15		950
11	C	0	0	1	5	1	0	0	11		750
12	A	1	0	0	2	1	0	1	13		700
13	B	0	1	0	7	1	0	1	13		950
14	D	0	0	0	10	0	0	1	13		950
15	A	1	0	0	10	0	0	1	13		1100
16	B	0	1	0	6	1	0	1	15		1150
17	B	0	1	0	8	1	0	1	13		1800
18	C	0	0	1	7	1	0	1	18		650
19	A	1	0	0	4	0	0	1	16		750
20	A	1	0	0	8	1	0	0	13		1650
21	A	1	0	0	6	0	0	0	12		550
22	A	1	0	0	8	0	0	0	13		550

Para generar una regresión lineal, se debe acceder a la pestaña "Datos"; luego, el botón "Análisis de datos" y, finalmente, seleccionar la opción "Regresión".

Figura 11-2: Regresión en Excel



Lo primero a ser analizado debe ser la tabla resumen. El coeficiente de determinación o R^2 determina el porcentaje de explicación de la varianza de la variable dependiente. En este caso, R^2 es 54.83 %. Adicional a ello, ya que es una regresión múltiple, si quisiéramos comparar este modelo con algún otro se debería usar R^2 ajustado (53.82 %). Finalmente, para evaluar la significancia global se debe observar el p-value de F, el cual sale como "valor crítico de F"[18]. Si el valor es mayor a 0.05, no se puede rechazar la hipótesis nula de no significancia conjunta, es decir, el modelo en conjunto no explica a la variable dependiente. En este caso, la probabilidad es menor a 5 %. Por ello, se rechaza

[18] En cualquier otro paquete estadístico, dirá *probabilidad* o, en inglés, p-value.

la H_0 y se acepta H_1. En otras palabras, el modelo sí es significativo conjuntamente para explicar a la variable dependiente.

Figura 11-3: Estadísticos de regresión

Resumen					
Estadísticas de la regresión					
Coeficiente de co	0.74044219				
Coeficiente de de	0.54825464				
R^2 ajustado	0.53815977				
Error típico	242.328029				
Observaciones	367				
ANÁLISIS DE VARIANZA					
	Grados de liberta	*na de cuadrado*	*io de los cua*	*F*	*Valor crítico de F*
Regresión	8	25514023.2	3189252.9	54.3102321	2.77823E-57
Residuos	358	21022788.8	58722.8737		
Total	366	46536812			

b) A partir de la interpretación de las variables, ofrecer una conclusión inicial para el proyecto.

Solución:

Primero, es necesario observar la significancia individual de todas las variables. Aquellas en rojo son las que tienen una probabilidad mayor a 0.05. Por ello, no se rechaza la H_0 de no significancia individual. Eso quiere decir que la variable "A" (una de las variables *dummy* utilizada para "Colegio"), "C" y "Situación laboral" no explican el salario. A partir de ello, se puede tener una primera conclusión que deberá ser contrastada en las entrevistas a profundidad: egresar de los colegios A y C no tiene alguna diferencia significativa.

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*nferior 95.0%*	*uperior 95.0%*
Intercepción	-59.9548187	94.5180464	-0.63432139	0.52627622	-245.835192	125.925555	-245.835192	125.925555
A	31.7739092	43.2844726	0.73407177	0.46338549	-53.34987638	116.897695	-53.3498764	116.897695
B	98.6114608	42.1000399	2.34231276	0.01971198	15.8169953	181.405926	15.8169953	181.405926
C	-10.3934454	42.1941835	-0.24632413	0.80557251	-93.37305488	72.5861641	-93.3730549	72.5861641
Años de egreso	62.3453426	4.3179746	14.4385617	1.465E-37	53.85355966	70.8371255	53.8535597	70.8371255
Sexo	236.577781	25.5619905	9.25506098	1.9983E-18	186.3072502	286.848311	186.30725	286.848311
Situación laboral	-25.9599123	31.2213409	-0.83147974	0.40625691	-87.36019223	35.4403675	-87.3601922	35.4403675
Trabajó en el cole	230.262288	26.1359603	8.81017133	5.4801E-17	178.8629811	281.661594	178.862981	281.661594
NFP	16.6230894	5.98703409	2.7765149	0.00578341	4.848913186	28.3972656	4.84891319	28.3972656

Asimismo, tener un trabajo formal o informal no afecta el posible salario de un egresado. Esto es un hallazgo especialmente importante dado que estos resultados podrían sugerir que no existen incentivos suficientes para optar por un trabajo formal, ya que no hay diferencia suficiente en los salarios.

Por otro lado, estudiar en el colegio "B" asegura, en promedio, 98.61 soles más que estudiar en cualquiera de los otros tres colegios. Por cada año más de egresado, el sueldo se incremente en 62.35 soles. Un punto más en la nota final, significa 16.62 soles más en el salario; es importante resaltar que, si bien se comprueba que mejores desempeños académicos se ven reflejados en mejores sueldos, no es realmente una suma importante.

Sin embargo, los hallazgos más importantes fueron que las mujeres ganan en promedio 236.58 soles mensuales menos que los hombres; un análisis más importante sobre los condicionantes que hacen que las mujeres tengan menores ingresos debe ser incluido en las entrevistas.

Otra conclusión es que en promedio aquellos que habían trabajado durante la etapa escolar ganaban 230.26 soles más que aquellos que no. Esta conclusión, si bien va contra el sentido común puede estar relacionada a que aquellos que han trabajado previamente cuentan experiencia en algún oficio o actividad técnica, además de haber conocido el mercado laboral y generado redes de confianza y laborales.

c) Generar un nuevo modelo con la variable "Colegio" corregida. Evaluar la pertinencia del modelo y reconstruir la función.

Solución:

Luego de recodificar "A" y "B" como 1, y "C" y "D" como 0, lo primero a ser analizado debe ser la tabla resumen. El coeficiente de determinación o R^2 determina el porcentaje de explicación de la varianza de la variable dependiente. En este caso, R^2 es 54.32 %. Adicional a ello, ya que es una regresión múltiple, si quisiéramos comparar este modelo con algún otro se debería usar R^2 ajustado (53.57 %). Finalmente, para evaluar la significancia global se debe observar el p-value de F, el cual se muestra como "valor crítico de F". Si el valor es mayor a 0.05, no se puede rechazar la hipótesis nula, es decir, el modelo no es significativo conjuntamente. En este caso, la probabilidad es menor a 5 %. Por ello, se rechaza la H_0 y se acepta H_1, es decir, el modelo sí es significativo conjuntamente.

Figura 11-4: Estadísticos de regresión

Estadísticas de la regresión					
Coeficiente (	0.73706424				
Coeficiente (	0.5432637				
R^2 ajustado	0.53565143				
Error típico	242.985204				
Observacion	367				
ANÁLISIS DE VARIANZA					
	Grados de liberta	*na de cuadrad*	*dio de los cua*	*F*	*alor crítico de*
Regresión	6	25281760.6	4213626.76	71.3668296	2.6951E-58
Residuos	360	21255051.4	59041.8095		
Total	366	46536812			

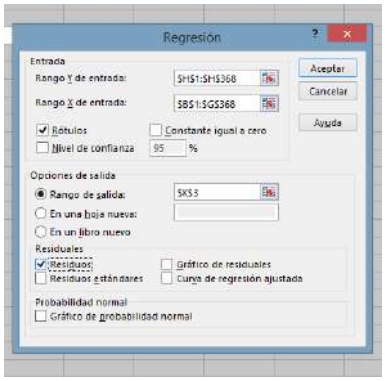
Observamos que "Colegio", "Años de egreso", "Sexo", "Trabajó en el colegio" y "Nota Final Promedio" fueron variables significativas individualmente. En esta ocasión se reconstruirá la ecuación a partir de los coeficientes que se muestran en la tabla:

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*nferior 95.0%*	*uperior 95.0%*
Intercepción	-80.2936781	89.9853999	-0.89229673	0.37282994	-257.256757	96.6694011	-257.256757	96.6694011
Colegio	75.7822429	25.9265645	2.92295738	0.00368639	24.7956975	126.768788	24.7956975	126.768788
Años de egre	62.59753	4.32101195	14.4867755	8.5829E-38	54.0999339	71.095126	54.0999339	71.095126
Sexo	241.01813	25.5322389	9.43975694	4.7986E-19	190.807055	291.229204	190.807055	291.229204
Situación lab	-30.398769	31.1612502	-0.97553111	0.32995193	-91.679719	30.882181	-91.679719	30.882181
Trabajó en el	228.592428	26.1852994	8.72980006	9.6868E-17	177.097061	280.087795	177.097061	280.087795
NFP	17.4670424	5.9879112	2.91705101	0.00375525	5.69136316	29.2427216	5.69136316	29.2427216

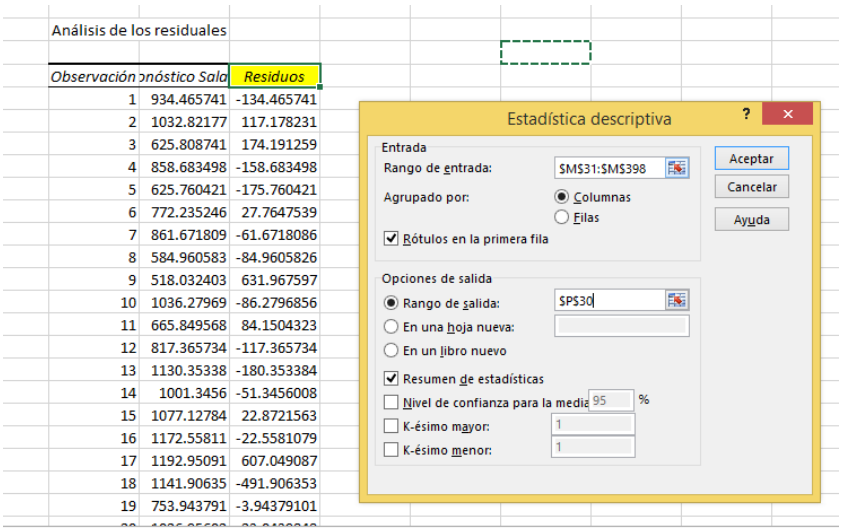
El modelo estimado, entonces, quedaría de la siguiente manera:

$$Salario = -80.2975.78 * (Colegio)62.6 \times (Años\ de\ egreso) + 241.02 \times (Sexo) + 228.59 \times (Trabajó\ en\ el\ colegio) + 17.48 \times (NFP)$$

d) Para hacer el análisis de normalidad de residuos, primero debemos conseguir los residuos del modelo. Para ello, simplemente, se le pide al programa los "Residuos" en la interfaz en la que normalmente se hace la regresión.



Al final de los cuadros que siempre se analizan, aparecerán tres nuevas columnas bajo el título "Análisis de los residuales". A la última columna, la de residuos, se le hará un análisis de estadística descriptiva simple.



Finalmente, se evalúa la asimetría y la curtosis de los residuos para saber si cumplen los criterios de normalidad. La curtosis debe ser muy cercana a tres para afirmar que la distribución es mesocúrtica[19]. En este caso, 0.5 es mucho menor a 3, por lo que estamos ante el caso de una distribución achatada o platocúrtica. Por último, el coeficiente de asimetría revela una cola hacia la derecha. Para hablar de una distribución simétrica, el coeficiente debe ser cercano a 0[20].

[19] Si fuera mucho mayor a 3, estaríamos en el caso de una distribución alargada o leptocúrtica.

[20] Si fuera un valor negativo, sería una distribución asimétrica con cola a la izquierda o negativa.

Figura 11-5: Análisis de residuos

Residuos	
Media	1.152E-13
Error típico	12.57933
Mediana	-23.01549
Moda	-84.96058
Desviación estándar	240.98529
Varianza de la muestra	58073.911
Curtosis	0.5669928
Coeficiente de asimetría	0.7842308
Rango	1288.8263
Mínimo	-491.9064
Máximo	796.91994
Suma	4.229E-11
Cuenta	367

Ya que no se observa una distribución normal de residuos, se evidencia que existen variables importantes que no han sido incluidas en el modelo y que están alterando la distribución aleatoria de los residuos.

Lecturas recomendadas

- Anderson, D., Sweeney, D., y Williams, T. (2008). *Estadística para administración y economía* (10.ª ed., cap. 14, 14.8-14.9; cap. 15, 15.7-15.8; cap. 16). Cengage Learning Editores.
- Levin, R., y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed. revisada; cap. 13, 13.5). México: Pearson Educación de México; Prentice Hall.
- Gujarati, D. (2004). *Econometría básica* (4.ª ed., cap. 9). McGraw-Hill Companies.

12. Bibliografía

Anderson, D., Sweeney, D., y Williams, T. (2008). *Estadística para administración y economía* (10.ª ed.). Cengage Learning Editores.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4.ª ed.). SAGE Publications.

Gujarati, D. (2004). *Econometría básica* (4.ª ed.). McGraw-Hill Companies.

Hernández, R., Fernández, C., y Baptista, P. (2010). *Metodología de la investigación*. México DF: McGraw-Hill

Levin, R., y Rubin, D. (2010). *Estadística para administración y economía* (7.ª ed. revisada). México: Pearson Educación de México; Prentice Hall.

Lind, D. A., Marchal, W. G., y Wathen, S. A. (2008). *Estadística aplicada a los negocios y la economía* (13.ª ed.). México: McGraw-Hill Interamericana.

Ministerio de Educación. (2018). Padrón de instituciones educativas [censo]. Recuperado de http://escale.minedu.gob.pe/documents/10156/85e996dc-6359-418c-bde6-81b2c898fceb

Pontificia Universidad Católica del Perú. (2018). *Drive* con bases de datos (BD) y solucionarios (SOL) de *Apuntes de clase: Métodos de Investigación Cuantitativa*. Recuperado de goo.gl/d1G6N1