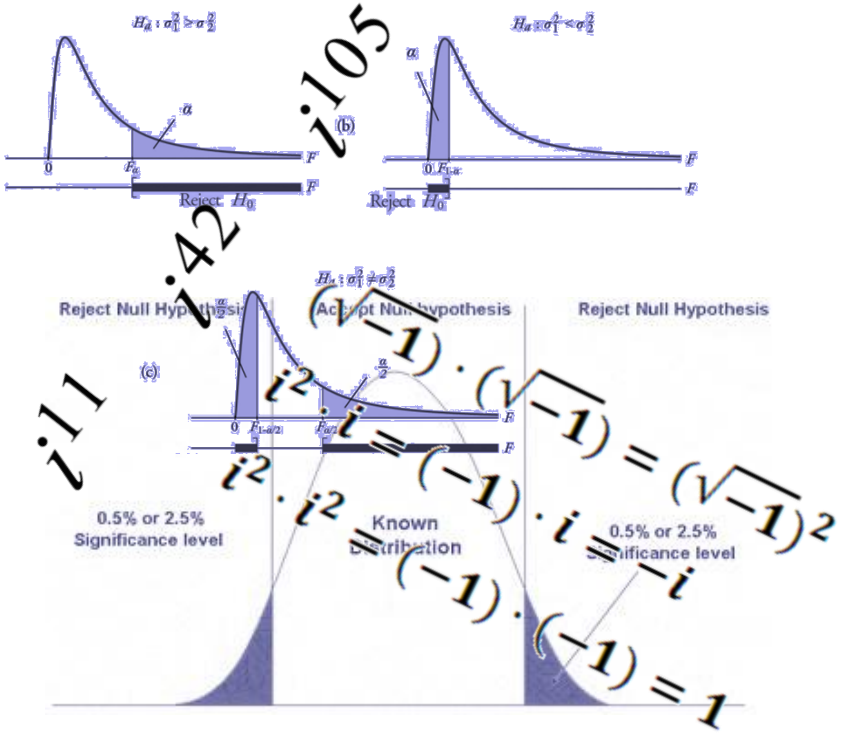


Capítulo 14

Análisis de datos (II)

Investigación de mercados

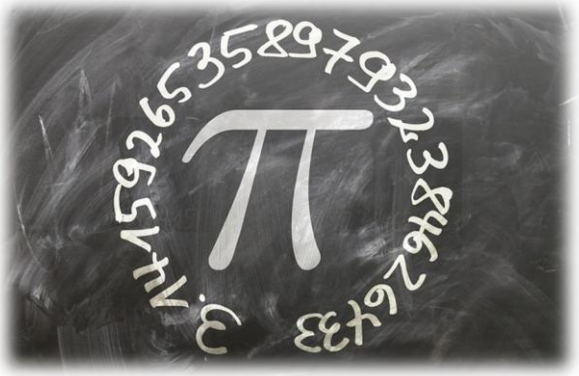
Prof. Verónica Rosendo Ríos

"Order is heaven's law"

ALEXANDER POPE

CONTENIDOS

1. Análisis de tablas estadísticas.
   - Distribución de frecuencias y porcentajes
   - Estadísticos asociados con la distribución de frecuencias
2. Tabulación cruzada
   - Tabulación cruzada de dos variables
   - Tabulación cruzada de tres variables
   - Estadísticos asociados con las tabulaciones cruzadas
3. Segmentación
   - Análisis clúster
   - Análisis factorial
4. Otros análisis multivariantes
   - Correlaciones
   - Regresiones



★DISTRIBUCIÓN DE FRECUENCIAS:

El análisis estadístico se puede dividir en varios grupos:

- **El análisis estadístico univariate** prueba hipótesis que involucran solo a una variable.
- **El análisis estadístico bivariante** prueba hipótesis que implica a dos variables.
- **El análisis estadístico multivariante** prueba hipótesis que implican a múltiples (tres o más) variables o conjunto de variables.

★ Distribución de frecuencias:

Grado de satisfacción global con una cadena nacional de tiendas

Etiqueta	Valor	Frecuencia (N)	Porcentaje	Porcentaje válido	Procentaje acumulado
Poco satisfecho	1	26	11.7	11.7	11.7
	2	34	15.3	15.3	27.0
	3	56	25.2	25.2	52.2
	4	62	27.9	27.9	80.1
Muy satisfecho	5	44	19.8	19.8	99.9
	9	2	0.01	Perdidos	100.0
	Total	223	100	100.0	

Estadísticos asociados con la distribución de frecuencias

- Medias de **localización:**

Media X=3,27

Mediana: La mediana es el percentil 50 (3)

Moda: Representa el pico más alto de la distribución. (4)

- Medias de **variabilidad:**

Rango: la diferencia entre los valores más grandes y más pequeños de la muestra. 5-1=4

Rango intercuartílico: Es la diferencia entre el percentil 75 y el 25; 4-2=2

Varianza/Desviación típica: $S_x = \sqrt{1,61} = 1.269$

- Medias de **forma:**

Asimetría

Curtosis: medida del pico relativo o planitud de la curva

✓ Estadísticos asociados con las tabulaciones cruzadas

Interés de los encuestados por la oferta de actividades culturales

		Edad		
Interés	<35	35-60	>60	Total
Interesado	24	44	55	123
No interesado	37	23	27	87
Total	61	67	82	210

Chi-Cuadrado (X^2): Nos ayuda a determinar si existe una asociación sistemática entre las dos variables. La hipótesis nula H_0 es que no hay asociación entre las variables.

Coeficiente Phi (ø): tabla de 2x2

Coeficiente de contingencia: fuerza de asociación en una tabla de cualquier tamaño $C = \frac{\sqrt{X^2}}{X^2 + n}$

Cramer's V: una versión modificada del coeficiente de correlación de phi, ø, y se utiliza en tablas de tamaño mayor a 2x2. V toma valores entre 0 y 1.

✓ Segmentación

★ 1. Análisis clúster:

Examina un conjunto completo de relaciones interdependientes, sin hacer distinción entre variables dependientes e independientes.

Los pasos a seguir para realizar un análisis clúster son los siguientes:

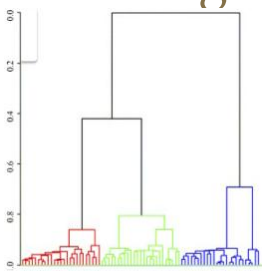
1. **Formulación del problema**

   - Seleccionar las variables en las que se basa la agrupación

2. **Selección de la distancia (medida de similaridad) y del procedimiento**

   Distancia: Euclidea, Manhattan o la distancia de Chebychev.

   El resultado obtenido se puede mostrar a través de diferentes medios, el más común es un gráfico visual denominado dendograma.



3. Selección del número de clústeres

No hay reglas específicas pero el tamaño relativo de los clústeres en el dendograma es un indicativo importante.

4. Interpretación de los clústeres

Examinar el **centroide** del clúster

Centroides

Cluster	V1	V2	V3	V4	V5	V6
1	3.1	6.3	3.0	6.0	1.7	3.2
2	3.0	1.5	3.8	1.7	5.7	3.6
3	5.9	3.4	5.9	3.1	3.5	6.2

★2. Análisis factorial

Uso: Reducción de datos *cuando las variables están correlacionadas*. Las relaciones entre muchas variables interrelacionadas se examinan y representan en términos de unos pocos factores subyacentes. Ej.: Imagen de marca en función de los atributos de la marca.

Modelo

El gado de varianza que comparten las variables se denomina **comunalidad.**

$$F_i = W_{i1}X_1 + W_{i2}X_2 + W_{i3}X_3 + \ldots + W_{ik}X_k$$

donde:

F_i= factor o variable

W_i = peso o carga factorial

X= factor o variable observable

k= número de variables

➢ Estadísticos asociados con el análisis factorial

- ✓ **Matriz de correlación**.
- ✓ **La prueba de esfericidad de Barlett's:** las variables poblacionales no están correlacionadas.
- ✓ **Test de medida de adecuación de Kaiser-Meyer-Olkin (KMO)**: Índice que evalúa la idoneidad de realizar o no un análisis factorial. (≥0,5).
- ✓ **Eigenvalue**. Representa la varianza total explicada por cada factor.
- ✓ **Cargas factoriales**. Las cargas factoriales son las correlaciones entre las variables y los factores.
- ✓ **Residuos**. Diferencias entre las correlaciones observadas en la matriz inicial y las reproducidas en la estimación de la matriz factorial.

- ✓ Si las correlaciones entre todas las variables son pequeñas, el análisis factorial puede no ser apropiado.

➢ Procedimiento para realizar el análisis factorial

Imagine que un investigador desea determinar los beneficios subyacentes que los consumidores buscan a partir de la compra de un coche. Se entrevistó a una muestra de 125 encuestados mediante entrevistas en concesionarios. A los encuestados se les pidió que indicaran la relativa importancia que asignaban a diferentes atributos en la compra de un coche utilizando una escala de Likert de siete puntos (1=muy importante, 7=poca importancia), siendo: V_1=Seguridad; V_2 = Tamaño; V_3 = Capacidad; V_4 = Emisiones; V_5 = Conectividad; V6 = Precio; V7 = Diseño.

Barlett = 1010.47, significación = 0.0000; KMO= 0.82. Estos valores indican la idoneidad de llevar a cabo el análisis factorial. Los resultados del análisis factorial se muestran en la siguiente tabla:

Variable	Factor	Autovalores	% Varianza	Varianza Acumulado
V_1	1	3.1	44.3	44.3
V_2	2	2.1	30.0	74.3
V_3	3	0.67	9.7	84
V_4	4	0.38	5.4	89.4
V_5	5	0.35	5.3	94.7
V_6	6	0.27	4.0	98.7
V_7	7	0.09	1.3	100.0

Variables	Factor 1	Factor2
V_1 Seguridad	0,89	0,21
V_2 Tamaño	0,27	0,74
V_3 Capacidad	0,23	0,56
V_4 Emisiones	0,01	0,30
V_5 Conectividad	0,02	0,27
V_6 Precio	0,59	0,15
V_7 Diseño	0,15	0,35

*El paso final implica el análisis de la bondad de ajuste del modelo. Para ello se examinan los **residuos**.*

★ Otros análisis multivariantes

★ 1. Correlaciones

Correlación simple

La correlación mide la fuerza de asociación entre dos variables métricas (de intervalo o razón) cuando hay relaciones lineales. Ej.: X/Y (calidad/precio)

Por ejemplo, imagine que un investigador desea explicar la opinión del encuestado con respecto a un producto (medida en una escala de Likert de 7 puntos donde 1 = valoración muy negativa; 7 = valoración muy positiva del producto) con respecto al precio (Likert de 7 puntos donde 1 = muy caro y 7 = muy barato) y la calidad (Likert de 7 puntos donde 1 = muy poca calidad y 7 = mucha calidad):

Opinión del consumidor sobre un producto

Encuestado	Opinión del producto	Calidad	Precio
1	6	7	3
2	7	7	7
3	6	5	5
4	3	4	1
5	6	5	6
6	4	6	1
7	5	6	7
8	2	2	4
9	7	6	7
10	6	6	17

R^2 cercano a 1; hay fuerte correlación

★ Otros análisis multivariantes

★ 1. Correlaciones

Correlación parcial

El coeficiente de correlación parcial ($r_{xy.z}$) mide el grado de asociación entre dos variables (X, Y) después de controlar los efectos de una o más variables adicionales (Z).

- ✓ Muy útil para identificar asociaciones espurias

r_{yx1}=0.893 r_{yx2}=0.632 r_{x1x2}=0.456

Las correlaciones parciales serían:

$$r_{yx1.x2} = \frac{r_{xy} - (r_{xz})(r_{yz})}{\sqrt{1-r^2_{xz}}\sqrt{1-r^2_{yz}}} = \frac{0.893 - (0.456)(0.632)}{\sqrt{1-(0.456)^2}\sqrt{1-(0.632)^2}} = \frac{0,604}{0,889 \quad 0,774}$$

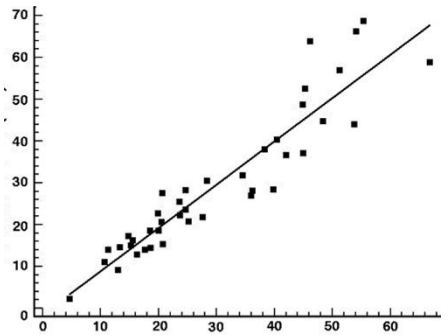
$r_{yx1.x2}$= 0.69

★2. Regresiones

2.1. Regresión lineal simple

- ✓ Hipótesis iniciales:
  - ✓ Normalidad.
  - ✓ Linealidad.
  - ✓ Homocedasticidad.
  - ✓ Independencia.

Diagrama de dispersión



✓ Determinar si la relación entre las dos variables es lineal (linealidad).

Modelo lineal: $Y= ß_0 + ß_1X+e$ ó $Y_i= a+bx_i +e$

donde:

Y= variable dependiente

X= variable independiente

$ß_0$= constante (punto en el que la recta corta el eje vertical).

$ß_1$=pendiente de la línea (magnitud del efecto que X tiene sobre Y)

e= error asociado con la observación. Residuos (distancia de la observación a la recta de regresión)

✓ Estimación de los parámetros y estandarización de las variables.

Media de 0 y una varianza de 1. Cuando los datos están estandarizados, la constante, a, asume un valor de 0. El término coeficiente ß se usa para indicar el coeficiente de regresión estandarizado. Imaginemos los siguientes resultados, donde se observa un valor de ß de 0,961.

R			0.961		
R^2			0.922		
R^2 Ajustado			0.921		
Error estandar			2.329		
	GI		Análisis de la varianza Suma de cuadrados		Media cuadrática
Regresión	1		5442.989		5442.989
Residual	10		458.786		7.198978
F=759.834	Sig. of F=0.0000				
Variable	B	SE_B	Beta (ß)	t	Significancia de t
Precio	0.48971	0.07185	0.961	6.815	0.0000
(Constante)	1.05434	0.74335		1.41	0.1452

✓ Significación estadística

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

La hipótesis nula establece que no hay relación lineal entre las dos variables

Precio (Y)= 1.05434 + 0.4891 (calidad)

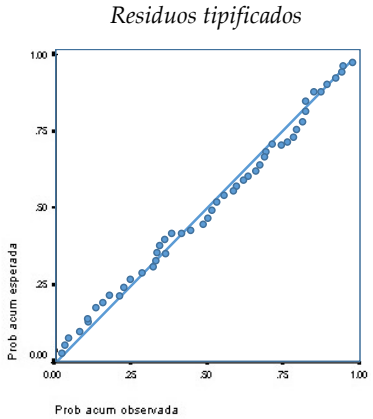
✓ Porcentaje de variabilidad explicado.

Este coeficiente tiene un rango entre 0 y 1. En el ejemplo anterior ($R^2 = 0,922$),están muy relacionados, el precio explica un 92% de la variabilidad de la calidad.

✓ Análisis de los residuos. Comprobación de las suposiciones iniciales.

★ Normalidad, independencia y homocedasticidad: Examen de residuos.

Normalidad: Examinando los residuos tipificados y el histograma (para cada valor fijo de X, la distribución de Y es normal); prueba de normalidad de Kolmogorov-Smirnov o de Shapiro-Wilk.

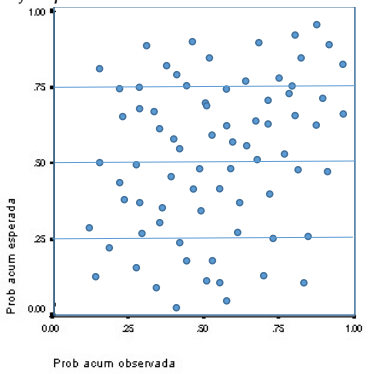


Independencia: Las observaciones se han tomado de forma independiente y los errores no están correlacionados. Durbin-Watson.

★ Residuos

Homocedasticidad, se debe realizar un diagrama de dispersion de las estimaciones tipificadas (ZPRED, valores predichos por el modelo) frente a los residuos tipificados (ZRESID). Ningún patrón.

Ejemplo de homocedasticidad



2.2. Regresión múltiple

- ✓ Una sola variable dependiente y dos o más variables independientes.
- ✓ Linealidad, homocedasticidad, independencia y normalidad.
- ✓ **R^2 Ajustado**. Número de variables independientes y el tamaño muestral
- ✓ **Coeficientes de regresión parciales**
  - ✓ $Y= a + b_1X_1 + b_2X_2 \ldots + b_kX_k + e$

2.2. Regresión múltiple

- ✓ Una sola variable dependiente y dos o más variables independientes.
- ✓ Linealidad, homocedasticidad, independencia y normalidad.
- ✓ **R^2 Ajustado**. Número de variables independientes y el tamaño muestral.
- ✓ **Coeficientes de regresión parciales**
  - ✓ $Y= a + b_1X_1 + b_2X_2 \ldots + b_kX_k + e$

$Y= 1.05434 + 0.47910\ X_1 + 0.\ 28865X_2$

Precio (Y)= 1.05434 + 0.479910 (calidad)+ 0. 28865 (imagen de marca)

Con respecto a la hipótesis nula en las regresiones múltiples, vendría dada por:

$H_0 : ß_1=ß_2=ß_3=\ldots=ß_k=0$

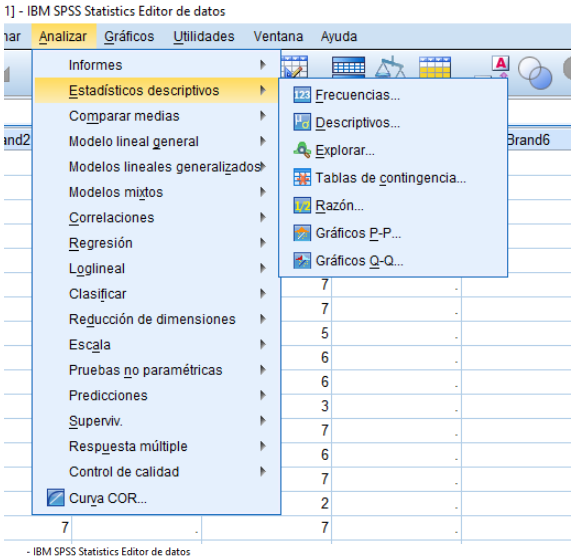
- ✓ Multicolinealidad

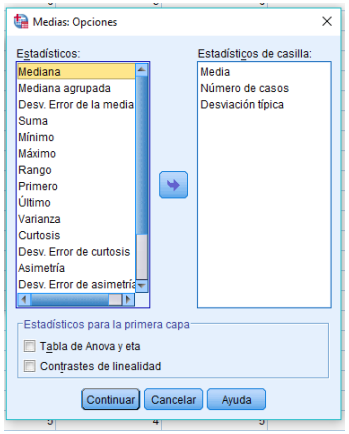
  Inter-correlaciones muy altas entre las variables independientes. Elegir.

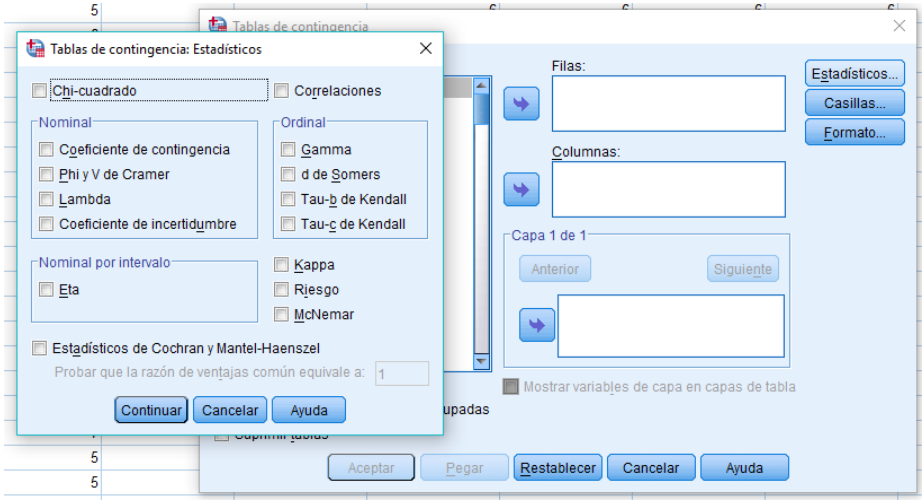
Análisis univariante, bivariante y multivariante (SPSS V.19)

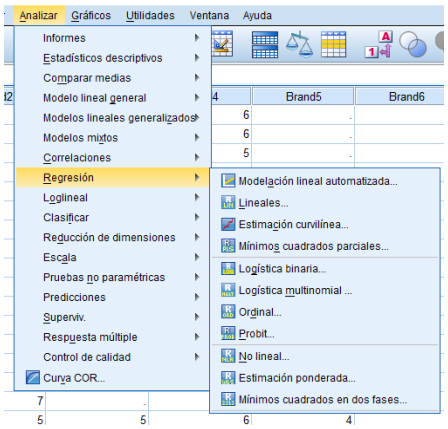
Análisis univariante	
Variables no métricas	Menú: Analizar/Estadísticos descriptivos/Frecuencias (Análisis: frecuencias, porcentajes, porcentaje acumulado, etc)
Variables métricas	Menú: Analizar/Estadísticos descriptivos/Descriptivos (Análisis: rango, suma, media, moda, mediana, varianza, desviación típica, etc)
Análisis bivariante	
Tabulación cruzada de frecuencias y porcentajes	Menú: Analizar/Estadísticos descriptivos/Tablas de contingencia/Casillas/Recuentos y porcentajes
Prueba de la chi cuadrado	Menú: Analizar/Estadísticos descriptivos/Tablas de contingencia/Estadísticos/Chi cuadrado
Coeficiente de correlación entre rangos	Menú: Analizar/Correlaciones/Bivariadas/Spearman
Coeficiente de correlación lineal	Menú: Analizar/Correlaciones/Bivariadas/Pearson
Coeficiente alfa de Cronbach	Menú: Análisis/ Escalas/Análisis de fiabilidad/ Alfa
Análisis de varianza	Menú: Analizar/ Comparar medias/ Medias/Opciones/Tabla de Anova Menú: Analizar/Comparar medias/Anova de un factor
Análisis multivariante	
Análisis factorial de componentes principales	Menú: Analizar/Reducción de dimensiones/Factor
Análisis factorial de correspondencias	Menú: Analizar/Reducción de dimensiones/Escalamiento óptimo
Análisis de regresión simple/ múltiple	Menú: Analizar/Regresión/Lineal

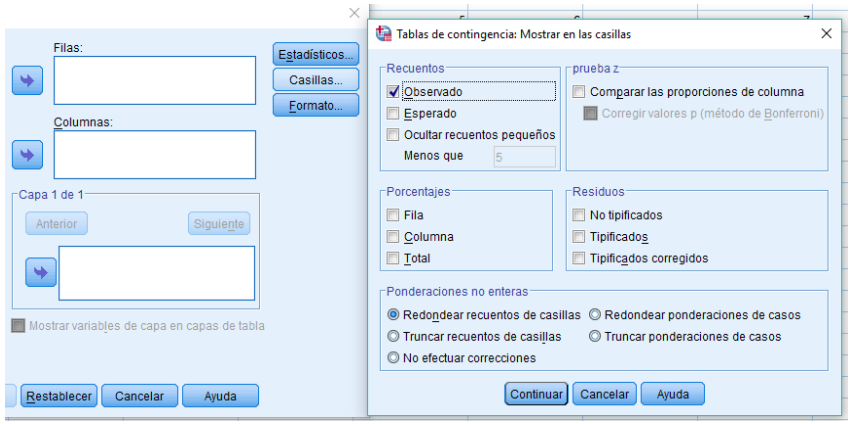
CAPÍTULO 14. ANÁLISIS DE DATOS II











REFERENCIAS BIBLIOGRÁFICAS

- Brown, T. J., y Suter, T. (2012): *MR*. South Western, Cenage Learning. USA.
- Hair, J.; Bush, R., y Ortinau, D. (2006): *Marketing research. Within a changing environment*. Revised International Edition (3rd ed.). McGraw-Hill, New York, USA.
- Malhotra, N. K. (1996): *Marketing Research. An Applied Orientation*. 2nd ed. Prentice-Hall International. USA.
- Malhotra, N. K. (2012): *Basic Marketing Research*, 4th Edition, Prentice Hall, USA.
- Rosendo-Ríos, V., y Pérez del Campo, E. (2013): *Business Research Methods: Theory and Practice*. ESIC Editorial. España.
- Rosendo-Ríos, V.; de Esteban, J., y Antonovica, A. (2012): *MR: Development ofTheoretical Concepts for Market Research I and II*. South Western, Cenage Learning. USA.
- Zikmund, W. G.; Babin, B. J.; Carr, J. C., y Griffin, M. (2013): *Business Research Methods*. 9th Edition. South Western, Cenage Learning. USA.