UNIVERSIDAD DE JAÉN

Material del curso "Análisis de datos procedentes de investigaciones mediante programas informáticos"

Manuel Miguel Ramos Álvarez,

MATERIAL III "DESCRIPCIÓN DE LOS DATOS"

Índice

3. Acercamiento con objeto de resumir la información: análisis descriptivo-exploratorio 2
3.1. Estadísticos descriptivos de utilidad para el resumen descriptivo 3
3.2. Descripción de los datos a partir de la representación gráfica 4
3.2.1. Gráficos básicos 4
3.2.2. Gráficos de asociación entre variables 5
3.2.3. Nuevos formatos de representación 6
3.3. Los supuestos estadísticos. Resumen de los supuestos implícitos en el análisis estadístico de modelización 7
3.4. Análisis de los residuales con fines diagnósticos 8
3.5. Transformaciones de datos y recomendaciones para su selección 10
3.6. Clasificación de las pruebas de análisis no paramétrico 12
3.7. Identificación de los valores extremos 13
3.8. Realización de los supuestos de prácticas 16
3.8.1. Supuesto 1 17

3. Acercamiento con objeto de resumir la información: análisis descriptivo-exploratorio

- La descripción inicial de los datos tiene por objeto reducir la abundante cantidad de información obtenida en las variables medidas; a partir de la tendencia central, la variabilidad o dispersión y la forma en que los datos se agrupan (apuntamiento y curtosis).
- Cuando los datos exhiben puntos extremos, sería preferible optar por una alternativa robusta, como la mediana.
- La síntesis de los datos puede realizarse mediante Análisis Exploratorio de Datos ("*Exploratory Data Análisis*", EDA).
- A veces es necesario llegar más lejos sometiendo los datos a alguna técnica estadística Multivariada de Reducción, como el Análisis Factorial o el Análisis de Cluster.
- Ver Convenciones acerca de las representaciones gráficas en el manual
- En general, será de utilidad:
    - Estadísticos descriptivos de utilidad para el resumen descriptivo: según el tipo de acercamiento a los datos y según las propiedades de interés
    - Análisis de los supuestos del Modelo mediante técnicas EDA y No Paramétricas
    - Transformación de los datos para evitar el incumplimiento de los supuestos
    - Generalidades sobre las técnicas de Análisis de carácter no paramétrico
    - La aproximación EDA (Exploratory Data Análisis) para análisis de residuales y valores extremos
    - Desarrollo de un supuesto de prácticas

3.1. Estadísticos descriptivos de utilidad para el resumen descriptivo

CUADRO 6.2. Adaptado a partir de Ramos, M.M.; Catena, A. y Trujillo, H. (2004). *Manual de Métodos y Técnicas de Investigación en Ciencias Del Comportamiento*. Madrid: Biblioteca Nueva.

A1. Descripción previa
Extremos:
- Mín vs. Máx

Puntos destacados:
- Percentiles.

A2. Forma
- Simetría y Apuntamiento

Histogramas
Distribución Frecuencias

B1. Tendencia Central
- Media

Robustos:
- Mediana

B2. Variabilidad
- Desviación típica
  Error estándar Media

Robustos:
- Desviación Mediana

Diagramas:

Clásico
- Líneas con medias y barras de error

Robusto
- Caja y bigotes

C1. Asociación vs. Correlación

Cuantitativas.
- Pearson

Ordinales
- Rangos Spearman

Nominales
- Kappa de Cohen

Cuantitativa-Dicotómica
- Biserial-puntual

Cuantitativa-Dicotomizada
- Biserial

Cuantitativa-Dicotomizada
- Tetracórica

Diagrama dispersión
Tablas Contingencia/

Decisión fundamental: ¿Estadísticos clásicos versus robustos?

Analizar → Estadísticos descriptivos → Explorar → Dependientes: *Todas* → Aceptar.

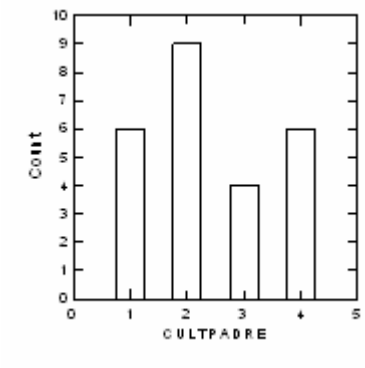
Statistics → Basic Statistics/Tables→ Descriptive statistics→ OK→ Summary:…Variables: X1-Y→OK→ Summary

Interpretación:

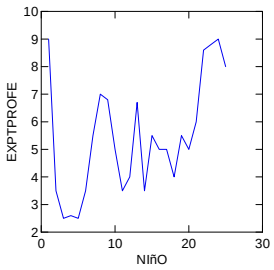
3.2. Descripción de los datos a partir de la representación gráfica

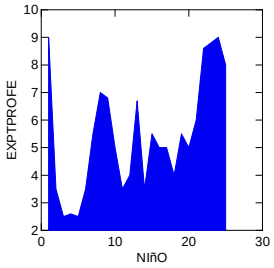
3.2.1. Gráficos básicos

HISTOGRAMAS
Nivel sociocultural padres (1: sin estudios, 2: primarios, 3: secundarios, 4: superiores)



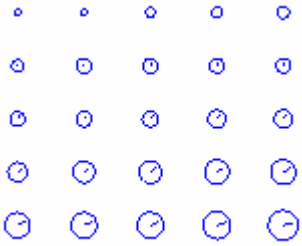
POLÍGONO FRECUENCIAS
Expectativa profesor Rendimiento cada niño del aula.





SIMBÓLICOS
Diagrama **circular** o Gráfico tarta. Nivel sociocultural padres (1: sin estudios, 2: primarios, 3: secundarios, 4: superiores)

Pictograma.
Rendimiento en cada niño



Para Distribuciones de Frecuencias
Analizar → Estadísticos descriptivos → Frecuencias → Variables: X1, X2, X3, X4, Y → Gráficos: Opción Histogramas con curva normal → Aceptar.

Graphs→Histograms→ Variables: X1, X2, X3, X4, Y; Fit Type:Normal→ Aceptar.

Para Variables en su escala original:

- Para el supuesto 3. Si las variables fueran cuantitativas se hace todo igual pero con el tipo "Líneas".
- Gráficos →(Cuadros de Diálogos Antiguos)→ Barras→ Agrupado; Resúmenes para distintas variables; Definir → Las barras representan: *Todas las medidas* (i.e. Tiempo80, Tiempo90, Tiempo100); Eje de categorías: *Variables de agrupación* (i.e. VARX1)→ Aceptar.

- Graphs→(2D Grpahs)→ Means w/Error Plots→ Graph Type: Columns; Mutiple; Unique values; Variables→ Dependent…: *Todas las medidas* (i.e. Tiempo80, Tiempo90, Tiempo100), Grouping…: *Variables de agrupación* (i.e. VARX1) → OK→ Aceptar.

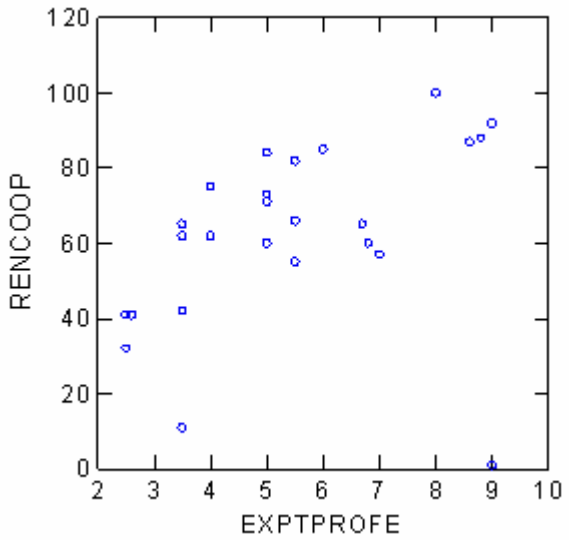
Interpretación:

Interpretación:

3.2.2. Gráficos de asociación entre variables

- Los diagramas de dispersión o nube de puntos [*scatterplot*] son recomendables para estudiar la asociación entre dos o más variables. Por ejemplo, respecto a la relación entre el rendimiento y las expectativas del profesor sería,

Diagramas de dispersión o nube de puntos.
Variables: Rendimiento de los guiños y expectativas profesor.



Para Regresión:

- Graphs → Scatterplots → Regular → Eje Y: Y; Eje X: X1; Fit Type: Linear → OK→ Aceptar.
- Repetir la operación para todos los pares X-Y (es decir, X2-Y, X3-Y, etc.).
- Rastrear todas las interdependencias si es regresión múltiple: Graphs → Matriz Plots → Square Scatter Matrix → Variables en la Matriz: *Todas* → OK→ *Pestaña* Advanced → Fit: Linear→ Aceptar.

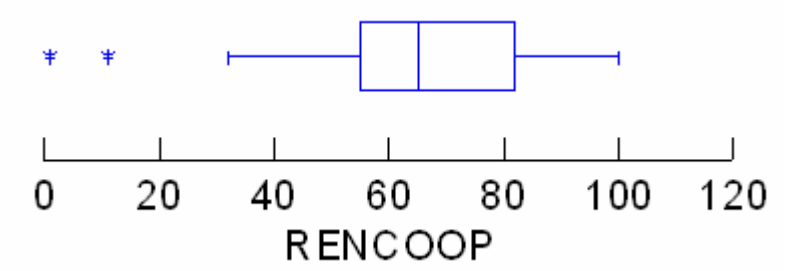
- Gráficos → Dispersion→ Simple → Eje Y: Y; Eje X: X1 → Aceptar.
- Repetir la operación para todos los pares X-Y (es decir, X2-Y, X3-Y, etc.).
- Rastrear todas las interdependencias si es regresión múltiple: Gráficos → Dispersion→ Matricial → Variables en la Matriz: *Todas* → Aceptar.

Interpretación:

3.2.3. Nuevos formatos de representación

- Dentro de EDA han surgido representaciones que condensan una mayor cantidad de información que los precedentes y son más flexibles para alterar el nivel de detalle de la representación.
- Los más importantes son los diagramas de tallo y hojas y los diagramas de caja y barbas.

El diagrama de caja y barba ("**Box Plot"**) en el mismo caso de rendimiento en situación cooperativa es,



Para Variables en su escala original:

- Para el supuesto 3.
- Gráficos → Diagramas de caja → Agrupado; Resúmenes para distintas variables; Definir → Las barras representan: *Todas las dependientes medidas*; Eje de categorías: *Variables de agrupación* → Aceptar.

- Graphs→ 2D Grapas→Box Plots→Box-Whiskers→Multiple→ Variables→ Dependent: Tiempo80, Tiempo90, Tiempo100; Grouping: VarX1→ OK→Aceptar

Interpretación:

3.3. Los supuestos estadísticos. Resumen de los supuestos implícitos en el análisis estadístico de modelización

(**Cuadro 7.2** adaptado a partir de Ramos, M.M.; Catena, A. y Trujillo, H. (2004). *Manual de Métodos y Técnicas de Investigación en Ciencias Del Comportamiento*. Madrid: Biblioteca Nueva.).

B) SUPUESTOS GENERALES

1) **Normalidad**. Las observaciones se extraen de poblaciones distribuidas según la Normal para cada grupo.
   - Pruebas de Bondad de Ajuste.

2) **Homocedasticidad**. El numerador y denominador de la razón F son estimaciones de la misma varianza poblacional, σ_ε^2. De ahí que las varianzas en los diferentes tratamientos tienen que ser iguales.
   - Prueba de Hartley, Levene, **Brown-Forsythe**.
   - Supuesto de Esfericidad respecto a Homogeneidad de Varianzas-Covarianzas según la Prueba de Mauchley.

3) Respecto a cada nivel j, el **error E_{ij}** ha de cumplir algunas condiciones:
   - Es **independiente** del resto de errores.
   - Se distribuye según una **Normal** dentro de cada población de tratamiento $N(0,\sigma_\varepsilon^2)$. Es decir con **media cero y varianzas equivalentes**.
   - Prueba Rachas, Corr. Intraclases, Durbin-Watson.

4) La ecuación estructural del modelo refleja una composición **aditiva** de las fuentes de variación.

- Para el supuesto 3.
- Analizar → Pruebas No paramétricas → Chi-Cuadrado, Rachas, K-S para una muestra (Normal, etc.)
- En módulos específicos (i.e. Modelo Lineal General que es para ANOVA aparecen las de homocedasticidad, esfericidad y Normalidad).

Sobre Normalidad:
Statistics → Basics Statistics/Tables→ Descriptive statistics→ Variables: FreqY→ *Pestaña* Normality *y seleccionar* Kolmogorov-Smirnov, Shapiro-Wilks' W (también efectúa la prueba de Lilliefors aunque no se menciona explícitamente) → Histograms

Sobre homogeneidad Varianzas (Para el Supuesto 2):
Statistics → Basics Statistics/Tables→ Breakdown & …→ Variables: Dependent: Frecuenc; Grouping: Mes, Tipo→ OK → OK → *Pestaña* ANOVA & Tests → Levene → Brown-Forsythe.

→ Ver Ramos, Catena & Trujillo (2004) para otras pruebas más específicas.

Interpretación:

3.4. Análisis de los residuales con fines diagnósticos

- El análisis exploratorio de los gráficos que representan a los residuales nos permitirá una primera aproximación a los supuestos, así como una mejor comprensión de los datos. Por ejemplo, esto facilita la detección de puntos extremos.
- La primera impresión nos la puede proporcionar el análisis exploratorio convencional, es decir la representación de la vd frente a la vi así como los histogramas de distribución de frecuencias.
- Sin embargo, a veces estos gráficos no son aparentes y de hecho pueden llegar a enmascarar algunos problemas, por lo que son preferibles los de EDA a partir de los residuales (errores) estandarizados, pues tienen ventajas interpretativas al ser libres de escalas. Realizaremos una **terna de gráficos residuales**.
    - Representan los residuales estandarizados (eje Y) en función de las predicciones (Y') o pronósticos de un modelo.
    - Los residuales estandarizados (eje Y) en función de la variable predictora/independiente (X),
    - Mejor aún un gráfico basado en la probabilidad normal (*normal probability plot*) o tipo p-p, que ordena los residuales estandarizados desde el más negativo hasta el más positivo y los representa contra su valor esperado asumiendo una distribución Normal estándar. Este último puede, además servir, para contrastar otros modelos de distribución diferentes.
- **Interpretación**:
    - Desde el punto de vista del diagnostico de los supuestos, los gráficos de los residuales deberían exhibir una **forma aleatoria**, es decir no deberían seguir ningún patrón.
    - Igualmente también podríamos detectar puntos extremos, datos para los cuales el error es desproporcionado en relación al conjunto de datos.
    - En el gráfico de los errores según la probabilidad normal, deberíamos esperar que éstos se ajustasen a una línea recta con pendiente unidad.

En Statistica

- Statistics → Advanced Linear/...→ General Linear Models → OK → Variables: Dependent Variables: FeqY; Continuous pred: X1, X2, X3, X4 → OK→ OK → *Pestaña* More results → *Pestaña* Residuals 1→ Resids for default polots: Standardized → *Botón* Pred & resids; *Botón* Normal→ *Pestaña* Residuals 2→ X(var/pred/res): X1; Y(var/pred/res): z Resid. → Scatterplot of selected X → *(Hacer lo mismo para los otros predictores)*.

➢ Sería recomendable inspeccionarlos conjuntamente para facilitar su interpretación (i.e. pulsando sobre el icono de la izquierda que los agrupa en el Workbook donde están ubicados).

En SPSS (es algo más laborioso)

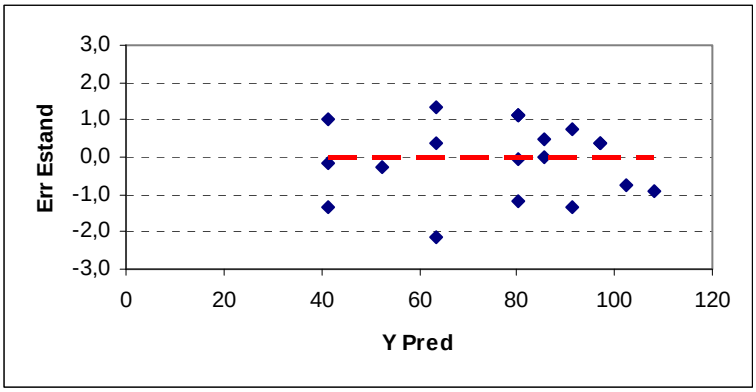
- Analizar → Regresión lineal → Dependiente: Y; Independientes: X1, X2, X3, X4; Guardar → Valores pronosticados: No Tipificados; Residuos: No Tipificados, Tipificados; Continuar → Aceptar
- En el fichero de datos se generan las columnas pre_1, res_1 y zre_1. Realizamos entonces los gráficos de dispersión:
    - Valor pre_1 vs zre_1
    - Valor X1 vs zre_1, Valor X2 vs zre_1, Valor X3 vs zre_1, Valor X4 vs zre_1

Interpretación:

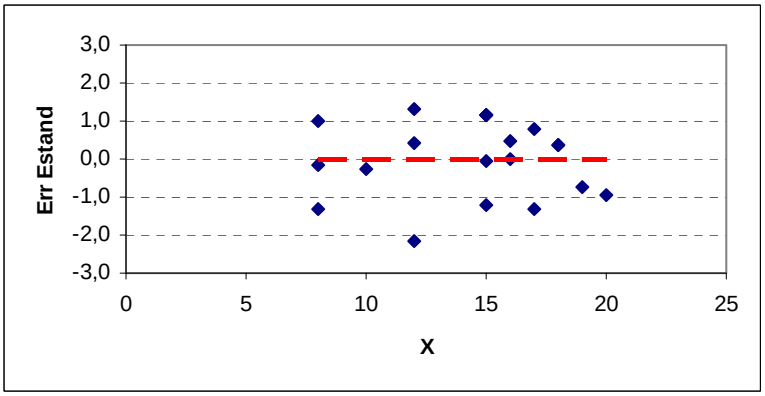
Análisis exploratorio de los residuales con fines diagnósticos

FIGURA 7.1. Tomada de Ramos, M.M.; Catena, A. y Trujillo, H. (2004). *Manual de Métodos y Técnicas de Investigación en Ciencias Del Comportamiento*. Madrid: Biblioteca Nueva.

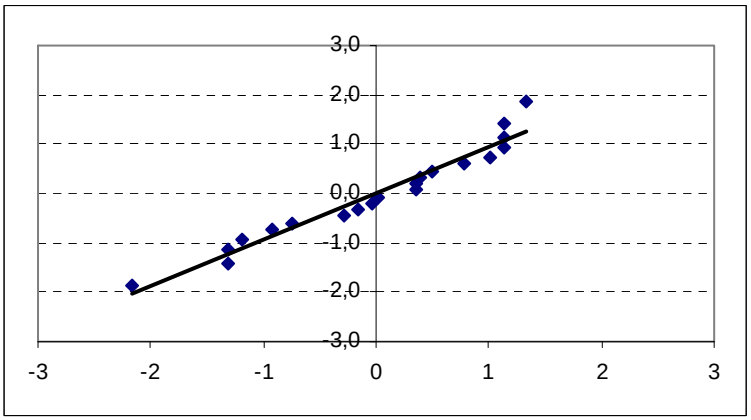
A



B



C



3.5. Transformaciones de datos y recomendaciones para su selección

- Realizamos dos representaciones gráficas, cuya inspección nos proporciona una aproximación al cumplimiento de los supuestos, especialmente aditividad, normalidad y homocedasticidad:
- De medias en función de varianzas o desviaciones
- Distribución de frecuencias de los errores $Y_{ij} - \overline{Y}_{ij}$.
- **Interpretación**:
- Lo ideal sería que no hubiera un patrón sistemático de relación entre medias y varianzas/desviaciones y que la distribución de los Errores fuera aleatoria.
- Por el contrario, tendremos sospechas del incumplimiento de los supuestos si medias y varianzas/desviaciones exhiben una tendencia lineal y los Errores exhiben una distribución clara, una normal o bien se agrupan de manera asimétrica.

- Una vez decidido, realizamos un diagnóstico de Box-Cox para decidir el tipo de transformación más adecuada, según la función:

$$\log(S_j) = \beta_0 + \beta_1 \cdot \log(\overline{Y}_j)$$

- El resto del proceso aparece en el siguiente cuadro:

(**Cuadro 7.8** adaptado a partir de Ramos, M.M.; Catena, A. y Trujillo, H. (2004). *Manual de Métodos y Técnicas de Investigación en Ciencias Del Comportamiento*. Madrid: Biblioteca Nueva.).

TRANSFOR-MACIÓN	EXPRESIÓN	INVERSA	CONDICIONES
Raíz	$Y'_{ij} = \sqrt{Y_{ij}} = Y_{ij}^{1/2}$	$Y_{ij} = Y'^{2}_{ij}$	* Medias y varianzas iguales. * Distribución de Poisson.
Logarítmica	$Y'_{ij} = \log_a Y_{ij}$	$Y_{ij} = a^{Y'_{ij}}$	*Medias y des.típicas proporcionales. * Datos asimétricos.
Recíproca	$Y'_{ij} = \frac{1}{Y_{ij}} = Y_{ij}^{-1}$	$Y_{ij} = Y'^{-1}_{ij}$	*Medias al cuadrado y des.típicas proporcionales.
Arco Seno	$Y'_{ij} = \text{arcsen}\sqrt{Y_{ij}}$	$Y_{ij} = \left(sen Y'_{ij}\right)^2$	* Medias y varianzas proporcionales. * Patrón de diamante de los errores.
Logit	$Y'_{ij} = \log \frac{Y_{ij}}{1-Y_{ij}}$	$Y_{ij} = \frac{e^{Y_{ij}}}{e^{Y_{ij}}+1} \equiv \frac{1}{1+e^{-Y_{ij}}}$	* Alternativa de la angular (arsen).
Z de Fisher	$Y'_{ij} = \frac{1}{2}\log\left(\frac{1+Y_{ij}}{1-Y_{ij}}\right)$	$Y_{ij} = \frac{e^{2Y'_{ij}}-1}{e^{2Y'_{ij}}+1}$	* Para variables que se obtienen como correlaciones. * Errores en forma de diamante.

ESTIMAR β_1 CON LA FUNCIÓN:

$$\log(S_j) = \beta_0 + \beta_1 \cdot \log(\overline{Y}_j)$$

PROPORCIÓN DESVIACIÓN Y MEDIA	β_1	$1-\beta_1$	TRANSFORMACIÓN RECOMENDADA
$\sigma \propto \mu^2$	2	-1	Recíproca
$\sigma \propto \mu^{3/2}$	1,5	-0,5	Recíproca Raíz
$\sigma \propto \mu$	1	0	Logarítmica
$\sigma \propto \mu^{1/2} \equiv \sigma^2 \propto \mu$	0,5	0,5	Raíz
$\sigma \propto cte.$	0	1	Ninguna

- Estimar Medias y Varianzas en Hoja de Cálculo (i.e. Excel), sus correspondientes logaritmos, aplicar la ecuación de regresión de arriba (i.e. "=PENDIENTE" en Excel) y tantear en la tabla según el valor estimado para la pendiente.
- Tantear las transformación sugerida junto con las más destacadas una por una y comprobar si se devuelve una relación lineal donde no la había. Realizamos las tres transformaciones mediante [Transformar→Calcular... ó en la caja Functions asociada a la variable en Statistica] y creamos tres nuevas variables, según las expresiones:
  - 1/X1, SQRT(X1) y LN(X1) mejor que LG10(X1)
- Entonces comparamos los gráficos de Diagnóstico para determinar la ganancia ahocicada a las transformaciones respecto a la var. Sin transformar y decidir finalmente si interesa realmente alguna de ellas o no.
- En Statistica hay una prueba específica: Statistics → Industrial Statistics & Six Sigma --> Experimental Design (DOE) → Central Composite, non-factorial, surface designs →Analyze design → Variables→ Dependent →Independent →OK → *Pestaña* Box-Cox→ Box-Cox Transformation.

Interpretación:

3.6. Clasificación de las pruebas de análisis no paramétrico

(Cuadro 7.9 adaptado a partir de Ramos, M.M.; Catena, A. y Trujillo, H. (2004). *Manual de Métodos y Técnicas de Investigación en Ciencias Del Comportamiento*. Madrid: Biblioteca Nueva.).

A) GENERAL

A.1 Una muestra.

A.1.1. **Bondad de ajuste** y análisis de los supuestos.

- Kolmogorov-Smirnov, Prueba χ^2, Lilliefors. Para Normalidad.
- Rachas. Para aleatoriedad.

A.1.2. En torno al parámetro de **posición.**

- De los signos.
- De la mediana.
- De los cuantiles.

A.2. Dos muestras.

A.2.1. De carácter **independiente**.

Sobre la forma general distribución:

- Prueba χ^2.
- Rachas de Wald-Wolfowitz

En torno al parámetro de **posición.**

- Prueba de la Mediana.
- U de Mann-Whitney.

A.2.2. **Relacionadas** o dependientes.

En torno al parámetro de posición o tendencia central.

- De los signos.
- Wilcoxon.

A.3. k-muestras.

A.3.1. De carácter independiente

- ANOVA de Kruskal-Wallis

A.3.2. **Relacionadas** o dependientes.

- Prueba de Friedman.

➢ Para el supuesto 3. • Analizar → Pruebas No paramétricas → … • Statistics → Nonparametrics→ …

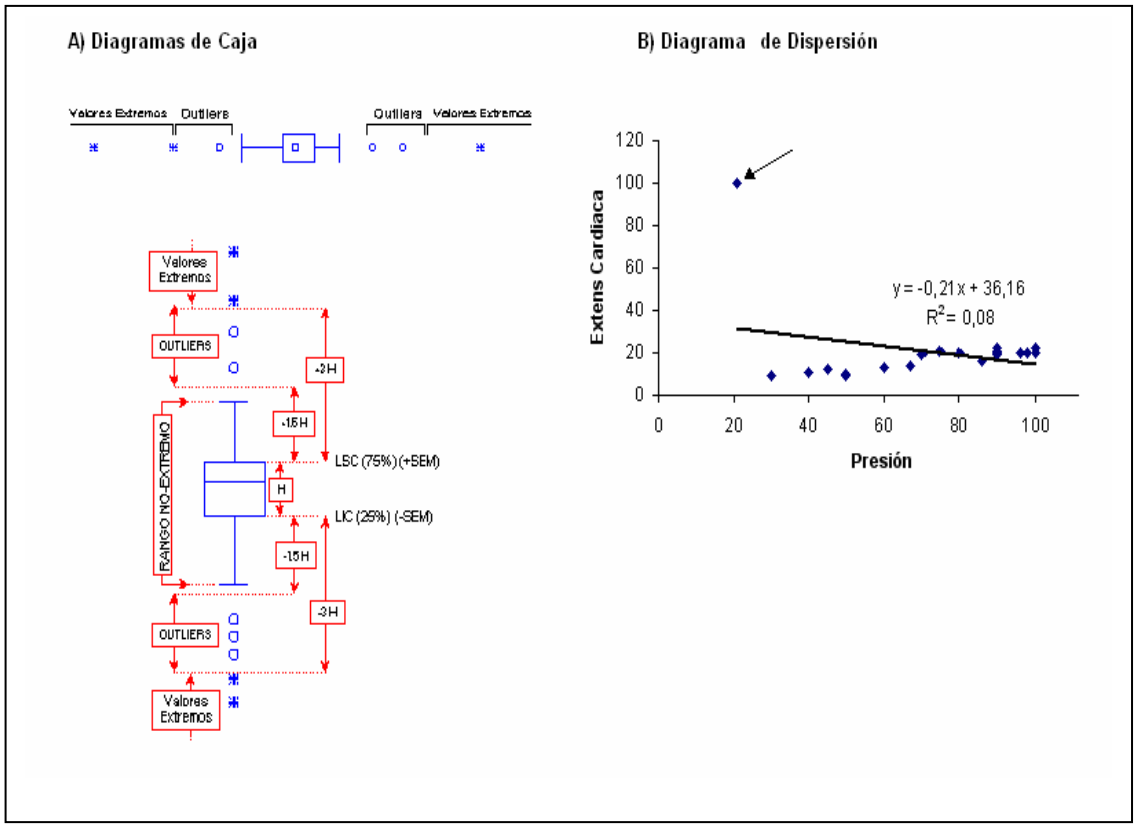
Interpretación:

3.7. Identificación de los valores extremos

Este análisis opera en dos pasos:

- Los valores que son identificados mediante análisis descriptivo se conocen como puntos extremos. Esto se hace normalmente mediante las representaciones gráficas oportunas (Diagrama de cajas en ANOVA y de dispersión en Correlacional), tal y como se refleja en la siguiente figura:

FIGURA 7.3. Tomada de Ramos, M.M.; Catena, A. y Trujillo, H. (2004). *Manual de Métodos y Técnicas de Investigación en Ciencias Del Comportamiento*. Madrid: Biblioteca Nueva.



- Para diferenciar entre "*outliers*" y valores extremos. Los segundos están dos veces más allá del criterio de desviación acotado.

Para Variables en su escala original (Hacerlo para el Supuesto 3):

- Gráficos → Diagramas de caja → Agrupado; Resúmenes para distintas variables; Definir → Las barras representan: *Todas las dependientes medidas*; Eje de categorías: *Variables de agrupación* → Aceptar.

- Graphs→ 2D Graphs→Box Plots→Box-Whiskers→Multiple→ Variables→ Dependent: Tiempo80, Tiempo90, Tiempo100; Grouping: VarX1→ OK→Aceptar

➢ En el supuesto 1, diagramas de dispersión o mejor gráficos de diagnóstico de residuales.

Interpretación:

Identificación de los valores extremos-II

- Posteriormente se procede a un análisis más exhaustivo si en la etapa previa tenemos sospechas de puntos extremos. Los puntos identificados mediante análisis residual se denominan puntos de influencia indebida y exigen técnicas específicas.

1) Identificación de valores inusuales en el predictor mediante el análisis de la influencia relativa.

- Distancia entre una observación en el predictor, digamos i, y su influencia en su propia predicción: influencia relativa (*levers*) y si tiene un valor elevado, nos indicará que se está ignorando al resto de predictores durante la estimación de la observación. Cuando el valor es elevado entonces se entiende como un punto de apalancamiento. La fórmula para regresión simple es simplemente una adaptación de la anterior.

Criterios interpretativos para considerar la posibilidad de un punto extremo:

- $h_{ii} > 2\overline{h}_{ii}$ ó $h_{ii} > 3\overline{h}_{ii}$
- Si la muestra es grande: $h_{ii} > 0,2$

Efecto:
Sobre la tasa de Error Tipo I.

2) Identificación de valores inusuales en el modelo mediante el análisis de los residuales.

- El objetivo ahora es determinar en qué medida un par de datos concreto afecta a las estimaciones básicas de modelización. Podríamos calcular los errores o residuales que se obtienen al diferenciar un modelo Ampliado que omite el par evaluado ($\widehat{Y}_{i.i}$) y otro Compacto que sí lo incluye ($\widehat{Y}_i$). Pero la lógica de contraste de Hipótesis no nos vale sin más pues se altera la estimación de los parámetros. Una fórmula que evita este problema viene dada por lo que se conoce como la **Distancia de Cook**:

Criterios interpretativos para considerar la posibilidad de un punto extremo:

- $D_i > \dfrac{4}{NP_{AMP}N}$
- Si la muestra es grande: $D_i > 1$ ó $D_i > 2$
- Para Di muy distanciados de los inmediatamente precedentes en magnitud.

Efecto:
- Sesgo en los parámetros estimados.
- Sobre la tasa de Error Tipo I o sobre el Tipo II.

Identificación de los valores extremos-III.
Decisiones en torno a los valores extremos detectados.

- Todas las pruebas estadísticas que hemos incluido no se deben de tomar como instrumentos para descartar automáticamente observaciones sino meramente para detectar puntos extremos y en general puntuaciones indebidas o "sospechosas", pero la interpretación de las mismas sólo puede corresponder a motivos teóricos.
- Si el punto extremo es claramente un error de codificación entonces obviamente habría que corregirlo.
- Pero salvo este caso, todos los demás merecen una consideración especial. De hecho, podría suceder que los valores extremos identificados provoquen un cambio en el modelo de partida, lo que evidentemente sería enriquecedor.
- Por el contrario, si omitimos observaciones de este tipo evidentemente estamos contribuyendo al sostenimiento de un modelo teórico incompleto.
- La conclusión metodológica es clara, se debe de informar de los puntos extremos detectados y efectuar los análisis estadísticos correspondientes con y sin dichos valores. *La mera omisión de la información para la que uno no tiene explicaciones es un proceder contradictorio con la motivación que nos llevó a invertir recursos para realizar la investigación. Este principio incrementa la probabilidad de que los resultados publicados no se confirmen y por ende que la propuesta teórica no se persiga en el futuro.*

Analizar → Regresión lineal → Dependiente: Y; Independientes: X1, X2, X3, X4; Guardar → Distancias: De Cook, Valores de Influencia; Continuar → Aceptar

---

Statistics → Advanced Linear/...→ General Linear Models → OK → Variables: Dependent Variables: FeqY; Continuous pred: X1, X2, X3, X4 → OK→ OK → *Pestaña* More results → Predicted and residuals.

- "STATISTICA will compute the (default) 95% Prediction intervals and 95% Confidence limits, the Standardized predicted and Standardized residual score, the Leverage values, the Deleted residual and Studentized deleted residual scores, Mahalanobis and Cook distance scores, the DFFITS statistic, and the Standardized DFFITS statistic."

Interpretación:

3.8. Realización de los supuestos de prácticas

Apartados recomendados en general
1.- Descripción básica basada en las frecuencias
2.- Descripción básica tanto gráfica (Dispersión vs. Cajas) como estadística. Decidir entre estadísticos clásicos vs. Robustos.
3.- Presencia de puntos extremos.
4.- Análisis de los supuestos y posibles transformaciones.
5.- Con el modelo identificado se realiza un análisis de los residuales.
Los ejemplos se ilustran con SPSS

3.8.1. Supuesto 1

A) Ejemplo para Ciencias:

1.- Descripción básica basada en las frecuencias

[Analizar|Estadísticos descriptivos|Frecuencias] [Gráficos…: Opción Histogramas con curva normal]
Aparentemente no hay puntos extremos destacados y las desviaciones de la normalidad no parecen preocupantes puesto que no son asimétricas, si acaso uniformes.

2.- Descripción básica tanto gráfica (Dispersión vs. Cajas) como estadística. Decidir entre estadísticos clásicos vs. Robustos
[Gráficos…|Dispersión]
Rastreamos todas las relaciones: Y-X1, Y-X2, Y-X3 e Y-X4
Que el de X1-Y sugiere una transformación pues se asemeja a una exponencial.
Que en algunos de ellos, i.e. X2-Y podría haber puntos extremos.
Que las dos últimas variables parecen redundantes pues son muy similares los diagramas.

[Analizar|Estadísticos descriptivos|Descriptivos …] [Opciones con más estadísticos]
Se puede apreciar que algunas variables exhiben cierta asimetría/curtosis tanto mayor cuanto más alejado de la unidad el valor, ya en sentido positivo ya en sentido negativo. Pero la forma de la distribución la veremos mejor al tratar el tema de los supuestos.

[Analizar|Estadísticos descriptivos|Explorar…] [Opciones con más estadísticos] y todas juegan el papel de variables dependientes.
Al comparar Media y Mediana se aprecia que en este caso hay algunas discrepancias y que por lo tanto el recurso a los estadísticos robustos, del tipo de la mediana, podría ser una buena opción, aunque esto se descarta con una visión general.

3.- Presencia de puntos extremos
Nos valemos de la opción de regresión lineal para estimar las distancias de valores inusuales
[Analizar|Regresión lineal] [Guardar…: Distancia de Cook y valores de influencia]
Case 11 podría ser sospechoso pues está por encima de 1 en Cook y tiene de los valores más elevados en cuanto a la medida de influencia (por encima de 0,4); aunque no parecen puntos muy extremos y podrían ser genuinos.

4.- Análisis de los supuestos y posibles transformaciones

Mejor dejarlo para el Módulo siguiente con el modelo definitivo, no obstante:
1º estimamos el modelo SATURADO, que es el de referencia, en el ejemplo el que incluye todos los predictores, pero además con InvAlt en lugar de la Altura, luego un modelo de regresión múltiple y aprovechamos para pedirle que guarde las predicciones, los residuales y los residuales tipificados y en el fichero de datos se generan las columnas pre_1, res_1 y zre_1
Realizamos entonces los gráficos:
Valor pre_1 vs zre_1
Valor X1 vs zre_1
Valor X2 vs zre_1
Valor X3 vs zre_1
Valor X4 vs zre_1
Un gráfico p-p para zre_1

De Normalidad, miramos los resultados aportados en la opción de Explorar vista más arriba, lo que nos lleva a concluir que al menos X1 y X3 podrían ser significativos e incumplir el supuesto. Los gráficos de los residuales así lo sugieren también.
De homocedasticiad y otros los veremos al analizar los gráficos de residuales en el módulo siguiente, una vez que tengamos el modelo.

De las transformaciones

- Dado el contexto de regresión, es preferible tantear las transformaciones más destacadas una por una y comprobar si se devuelve una relación lineal donde no la había.
- Realizamos las tres transformaciones mediante **[Transformar|Calcular…]** y creamos tres nuevas variables, según las expresiones:
  1/Alt
  SQRT()
  LN() mejor que LG10()
- ✓ Entonces acudimos nuevamente a la opción **[Gráficos…|Dispersión]** y comparamos los diagramas de dispersión de las transformaciones frente a la variable Altitud original, donde se aprecia que la mejor transformación para devolver linealidad es la del tipo Inverso.

5.- Con el modelo identificado se realiza un análisis de los residuales.
Lo veremos en el módulo siguiente, una vez estimado el modelo de regresión oportuno.

B) Ejemplo para Humanidades:

1.- Descripción básica basada en las frecuencias

[Analizar|Estadísticos descriptivos|Frecuencias] [Gráficos…: Opción Histogramas con curva normal]
Aparentemente no hay puntos extremos destacados y las desviaciones de la normalidad no parecen preocupantes puesto que no son asimétricas, si acaso uniformes.

2.- Descripción básica tanto gráfica (Dispersión vs. Cajas) como estadística. Decidir entre estadísticos clásicos vs. Robustos

[Gráficos…|Dispersión]
Rastreamos todas las relaciones: Y-X1, Y-X2, Y-X3 e Y-X4
Que el de X1-Y sugiere una transformación pues se asemeja a una exponencial.
Que en algunos de ellos, i.e. X2-Y podría haber puntos extremos.
Que las dos últimas variables parecen redundantes pues son muy similares los diagramas.

[Analizar|Estadísticos descriptivos|Descriptivos …] [Opciones con más estadísticos]
Se puede apreciar que algunas variables exhiben cierta asimetría/curtosis tanto mayor cuanto más alejado de la unidad el valor, ya en sentido positivo ya en sentido negativo. Pero la forma de la distribución la veremos mejor al tratar el tema de los supuestos.

[Analizar|Estadísticos descriptivos|Explorar…] [Opciones con más estadísticos] y todas juegan el papel de variables dependientes.
Al comparar Media y Mediana se aprecia que en este caso hay algunas discrepancias y que por lo tanto el recurso a los estadísticos robustos, del tipo de la mediana, podría ser una buena opción, aunque esto se descarta con una visión general.

3.- Presencia de puntos extremos

Nos valemos de la opción de regresión lineal para estimar las distancias de valores inusuales
[Analizar|Regresión lineal] [Guardar…: Distancia de Cook y valores de influencia]
Case 11 podría ser sospechoso pues está por encima de 1 en Cook y tiene de los valores más elevados en cuanto a la medida de influencia (por encima de 0,4); aunque no parecen puntos muy extremos y podrían ser genuinos.

4.- Análisis de los supuestos y posibles transformaciones

Mejor dejarlo para el Módulo siguiente con el modelo definitivo, no obstante:
1º estimamos el modelo SATURADO, que es el de referencia, en el ejemplo el que incluye todos los predictores, pero además con InvCulpab en lugar de la Altura, luego un modelo de regresión múltiple y aprovechamos para pedirle que guarde las predicciones, los residuales y los residuales tipificados y en el fichero de datos se general las columnas pre_1, res_1 y zre_1
Realizamos entonces los gráficos:
Valor pre_1 vs zre_1
Valor X1 vs zre_1
Valor X2 vs zre_1
Valor X3 vs zre_1
Valor X4 vs zre_1
Un gráfico p-p para zre_1

De Normalidad, miramos los resultados aportados en la opción de Explorar vista más arriba, lo que nos lleva a concluir que al menos X1 y X3 podrían ser significativos e incumplir el supuesto. Los gráficos de los residuales así lo sugieren también.
De homocedasticiad y otros los veremos al analizar los gráficos de residuales en el módulo siguiente, una vez que tengamos el modelo.

De las transformaciones

- Dado el contexto de regresión, es preferible tantear las transformaciones más destacadas una por una y comprobar si se devuelve una relación lineal donde no la había.
- Realizamos las tres transformaciones mediante **[Transformar|Calcular…]** y creamos tres nuevas variables, según las expresiones:
  1/Culpab
  SQRT()
  LN() mejor que LG10()
  - ✓ Entonces acudimos nuevamente a la opción **[Gráficos…|Dispersión]** y comparamos los diagramas de dispersión de las transformaciones frente a la variable Culpab original, donde se aprecia que la mejor transformación para devolver linealidad es la del tipo Inverso.

✓

5.- Con el modelo identificado se realiza un análisis de los residuales.
Lo veremos en el módulo siguiente, una vez estimado el modelo de regresión oportuno.

Volver Principio