Evaluación de Normalidad

ESTADÍSTICA APLICADA
PROF. MSC. VILMA S. ROMERO ROMERO
5TO CICLO | 09-06-2022

Introducción

Las desviaciones de la normalidad generalmente se presentan de varias formas:

- **Multimodalidad**
  - Tener varias modas sugiere que los datos pueden provenir de dos o más grupos distintos.
  - Los datos pueden ser una mezcla de muestras de diferentes poblaciones.
- **Asimetría**
  - A menudo se produce una falta de simetría cuando los datos se limitan a valores positivos o tienen un umbral límite.
- **Outliers**
  - Indica que se cometió un error de entrada de datos o que algunos datos son fundamentalmente diferentes.

Quantile-Quantile Plot

Gráfico que muestra si un modelo normal es una descripción razonable de la variación en los datos.

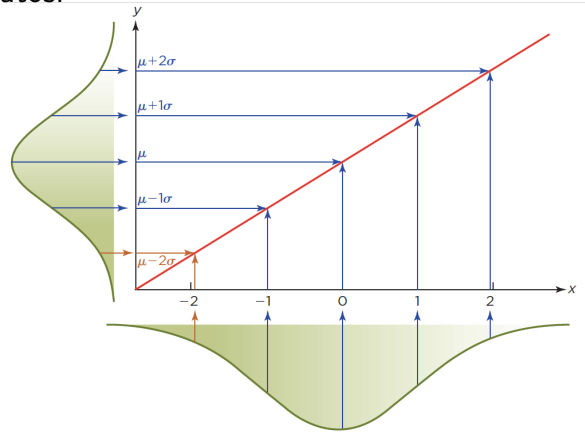


Fig. 1: Eje X: Distribución normal estándar. Eje Y: Distribución normal con parámetros μ y σ. Imagen tomada de Stine and Foster (2014)

UNIVERSIDAD DE LIMA

Características de QQ-Plot

- Si los cuantiles siguen una línea, sugiere un modelo normal para los datos.
- Q-Q plot no es lineal si las formas de las distribuciones difieren.
- Cuanto mayor se desvían los cuantiles de la línea, mayor es la desviación de la normalidad.
  - Se visualiza mejor con niveles de confianza. Si los cuantiles están dentro de estos límites, los datos se consideran normales.

Ejemplos

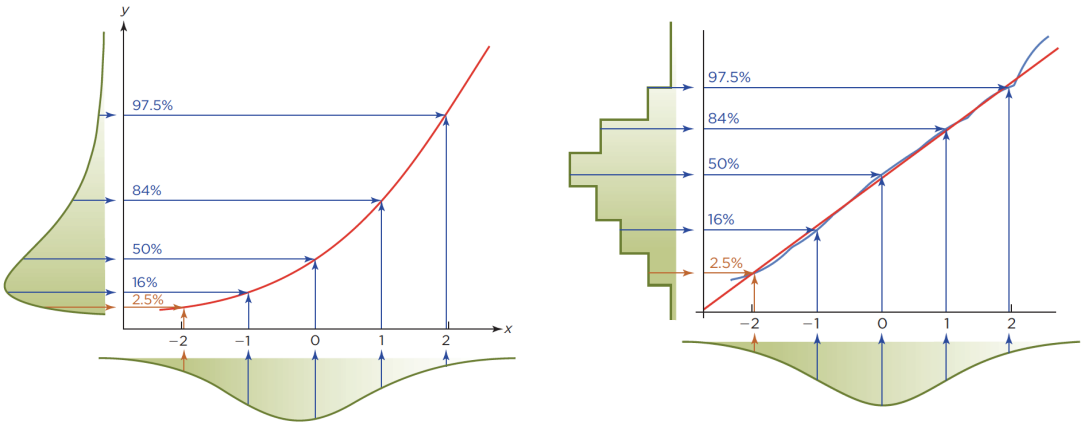
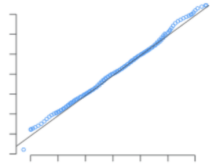


Fig. 2: Izquierda: Q-Q plot para una distribución sesgada. Derecha: Desviación ligera de la línea de referencia, pero sugiere una distribución normal.
Imagen tomada de Stine and Foster (2014)

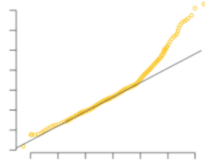
UNIVERSIDAD DE LIMA

Q-Q Plot

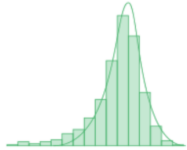
Datos distribuidos normalmente

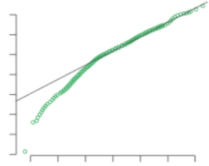


Datos sesgados a la derecha



Datos sesgados a la izquierda





UNIVERSIDAD DE LIMA

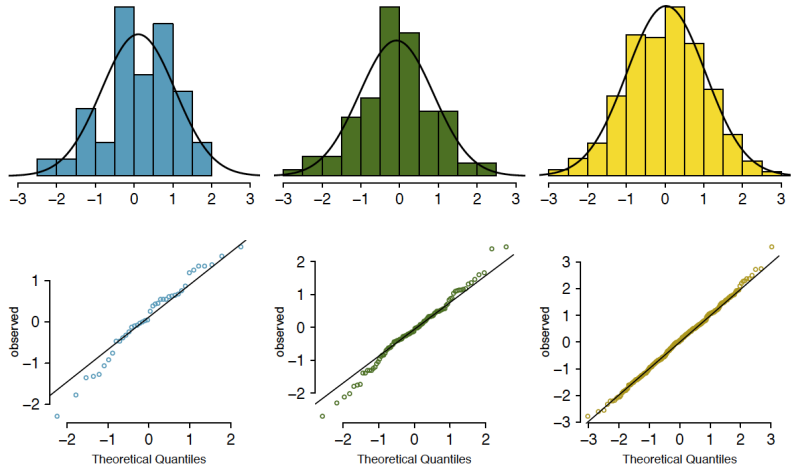


Fig. 3: Histogramas y Q-Q plots para tres conjuntos de datos normales simulados; $n = 40$ (izquierda), $n = 100$ (centro), $n = 400$ (derecha).
Imagen tomada de Diez et al (2015)

UNIVERSIDAD DE LIMA

Prueba de Normalidad

Procedimiento General:

1) Se formulan las hipótesis:
   H_0 : Los datos siguen una distribución normal.
   H_1 : Los datos no siguen una distribución normal.
2) Fijar el nivel de significación α.
3) Obtener el p-value de la prueba de normalidad específica.
4) Si p-value $> \alpha \Rightarrow$ No se rechaza la H_0. Por lo tanto, no existe suficiente evidencia para sugerir que los datos no siguen una distribución normal.

Prueba de Anderson-Darling

Es una prueba que evalúa qué tan bien se ajustan los datos a una distribución específica. Es usuada comúnmente como una prueba de normalidad.

Ejemplo: Evaluar si los puntajes obtenidos por los alumnos provienen de una distribución normal. Los datos se encuentran en el archivo "`Puntajes.csv`". Considerar un nivel de significación del 5%.

1) Se formulan las hipótesis:
   H_0 : Los datos siguen una distribución normal.
   H_1 : Los datos no siguen una distribución normal.
2) $\alpha = 0.05$

```
# Lectura de datos
datos <- read.csv("Puntajes.csv")
```

Prueba de Anderson-Darling en R

3) Prueba de Anderson- Darling

```
library(nortest)
ad.test(datos$Puntaje)
##
##  Anderson-Darling normality test
##
## data:  datos$Puntaje
## A = 0.62109, p-value = 0.1002
```

4) Como p-value $= 0.1002 > \alpha = 0.05$, entonces no se rechaza la H_0. Por lo tanto, no existe suficiente evidencia para sugerir que los puntajes no siguen una distribución normal.

Prueba de Shapiro-Wilk

- Contrasta si un conjunto de datos sigue una distribución normal o no.
- Esta prueba tiene una mayor potencia comparada con las demás pruebas de normalidad.

Ejemplo: Evaluar si los puntajes obtenidos por los alumnos provienen de una distribución normal. Los datos se encuentran en el archivo "`Puntajes.csv`". Considerar un nivel de significación del 5%.

1) Se formulan las hipótesis:
   H_0 : Los datos siguen una distribución normal.
   H_1 : Los datos no siguen una distribución normal.
2) $\alpha = 0.05$

Prueba de Shapiro-Wilk en R

```
# Lectura de datos
datos <- read.csv("Puntajes.csv")
```

3) Prueba de Shapiro-Wilk

```
shapiro.test(datos$Puntaje)
##
##  Shapiro-Wilk normality test
##
## data:  datos$Puntaje
## W = 0.95676, p-value = 0.06508
```

4) Como p-value $= 0.06508 > \alpha = 0.05$, entonces no se rechaza la H_0. Por lo tanto, no existe suficiente evidencia para sugerir que los puntajes no siguen una distribución normal.

UNIVERSIDAD DE LIMA

Recursos Adicionales I

Devore, J. (2019). *Introducción a la probabilidad y estadística para ingeniería y ciencias*. Cengage, 1 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_CENGAGE.

Johnson, R. A. (2012). *Probabilidad y estadística para ingenieros*. Pearson Educación, 8 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_PEARSON.

Kokoska, S. (2015). *Introductory Statistics*. W. H. Freeman and Company, 2 edition.

Mendenhall, W., Beaver, R. J., and Beaver, B. M. (2015). *Introducción a la probabilidad y estadística*. Cengage, 14 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_CENGAGE.

Recursos Adicionales II

Millones, R., Barreno, E., Vásquez, F., and Castillo, C. (2017). *Estadística Descriptiva y Probabilidades: Aplicaciones en la ingeniería y los negocios*. Lima: Fondo Editorial de la Universidad de Lima, 1 edition. Código Biblioteca U.Lima: 519.53 E.

Triola, M. (2018). *Estadística*. Pearson Educación, 12 edition. Tomado de http://webaloe.ulima.edu.pe/portalUL/bi/baseDatosEtech/index.jsp?BD=BI_RUTA_PEARSON.