Análisis Multivariante de Datos

Normalidad multivariante

Vanesa Jordá

Grado en Economía

Curso 2019-2020

Test de Normalidad en datos univariantes

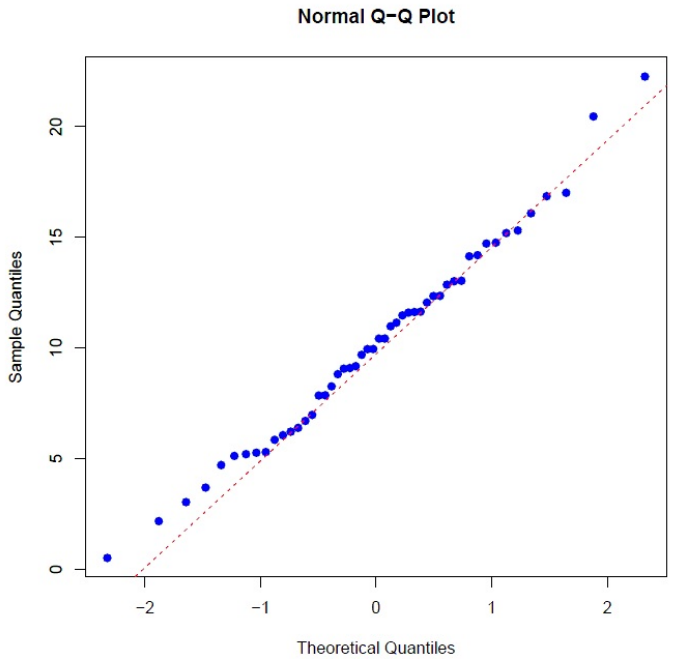
Existen varios test de normalidad para datos univariantes.

1. Basados en la función de distribución empírica.
   - Kolmogorov-Smirnov-Lilliefors.
   - Cramer Von Misses.
   - Anderson-Darling.
2. Bera-Jarque que analiza el grado de asimetría y de curtosis.
3. Shapiro-Wilks (métodos gráficos - QQ plot).

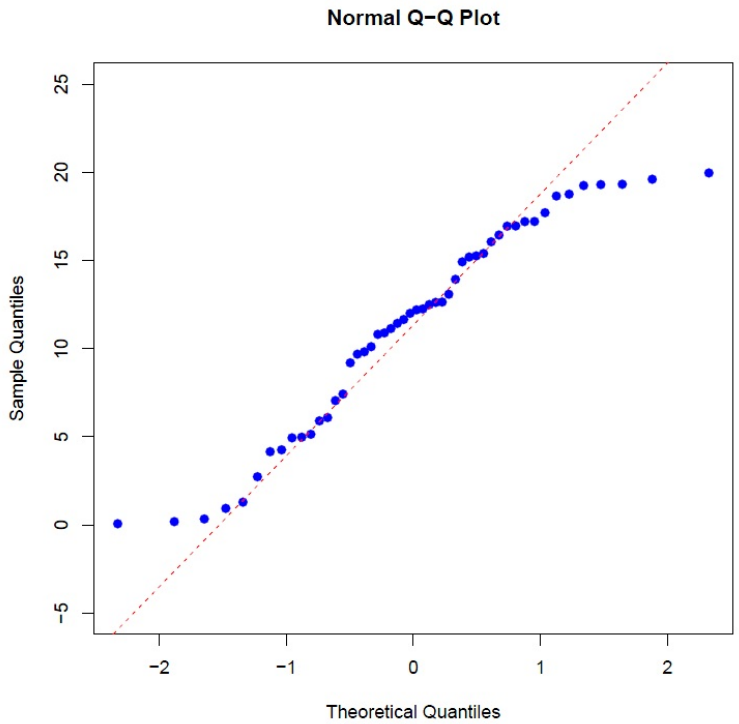
Normalidad en datos univariantes: Q-Q plot

- El gráfico Q-Q (quantile-quantile) compara los percentiles teóricos de la normal estándar con los empíricos.
- Si los datos se ajustan a la recta de pendiente 1 entonces siguen una N(0,1).
- Dependiendo de cómo se ajusten los datos a la recta nos informa sobre distintas características de la distribución.

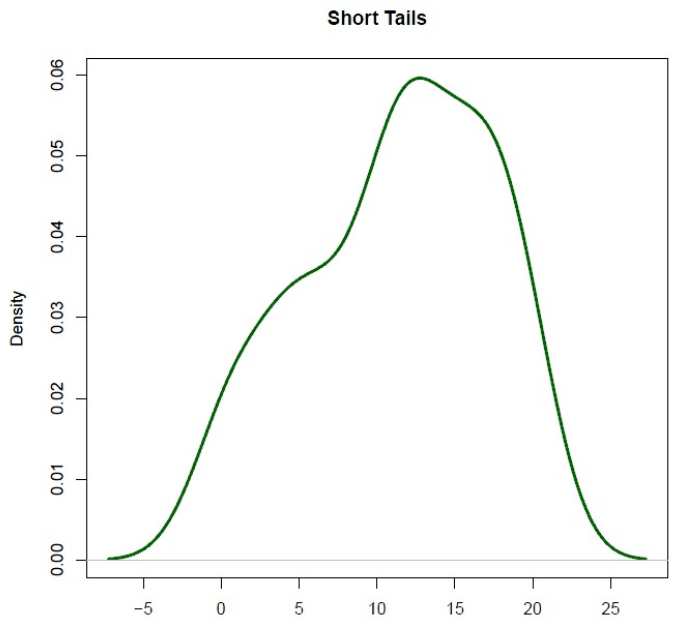
Normalidad en datos univariantes: Q-Q plot



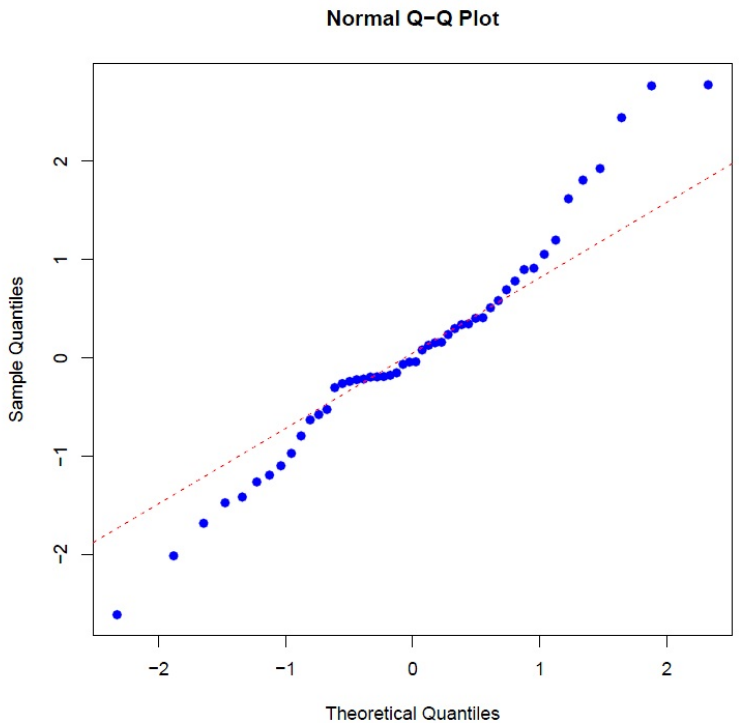
Colas cortas Q-Q plot



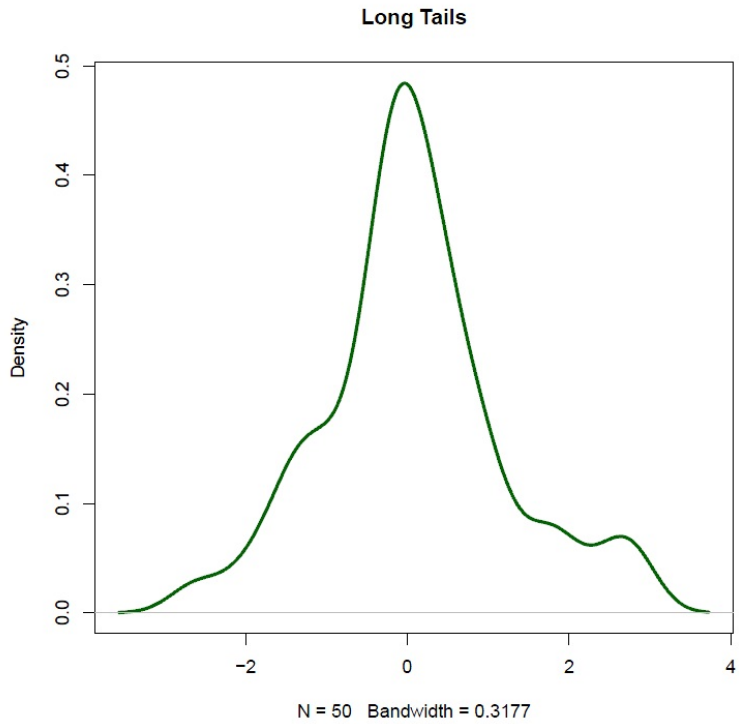
Colas cortas



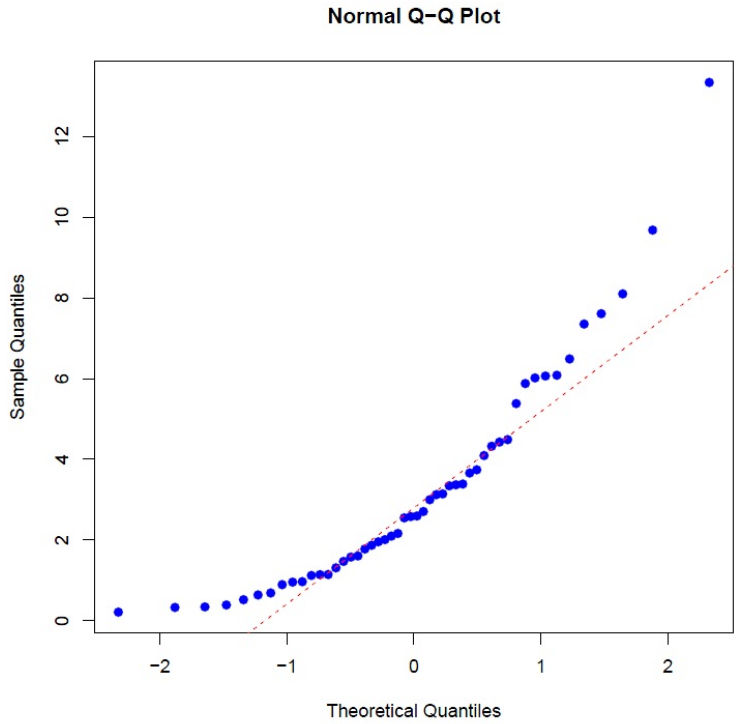
Colas pesadas



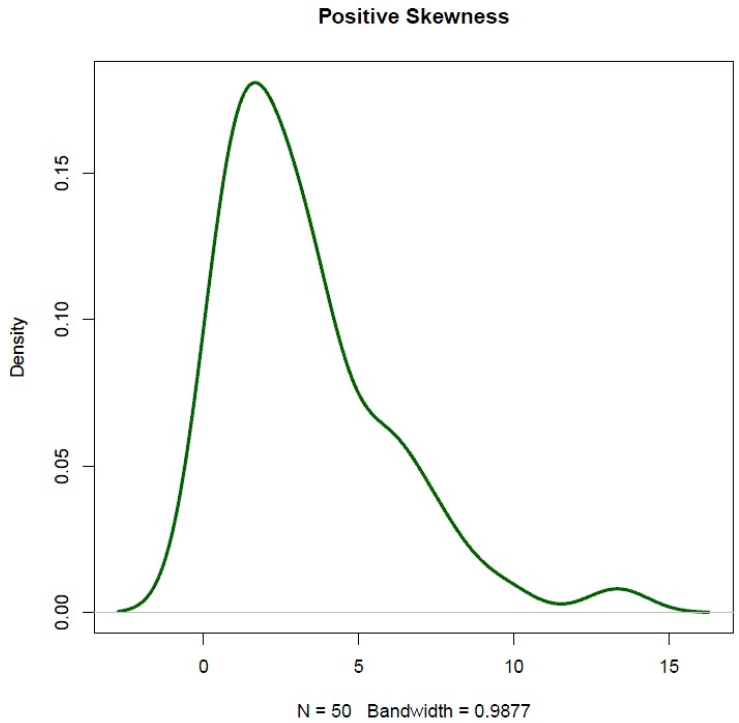
Colas pesadas



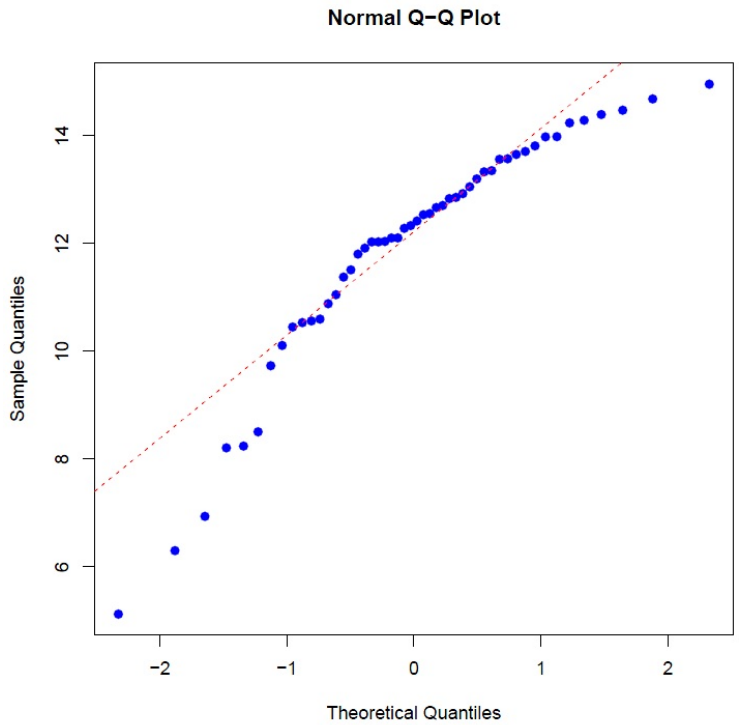
Asimetría positiva (cola derecha pesada)



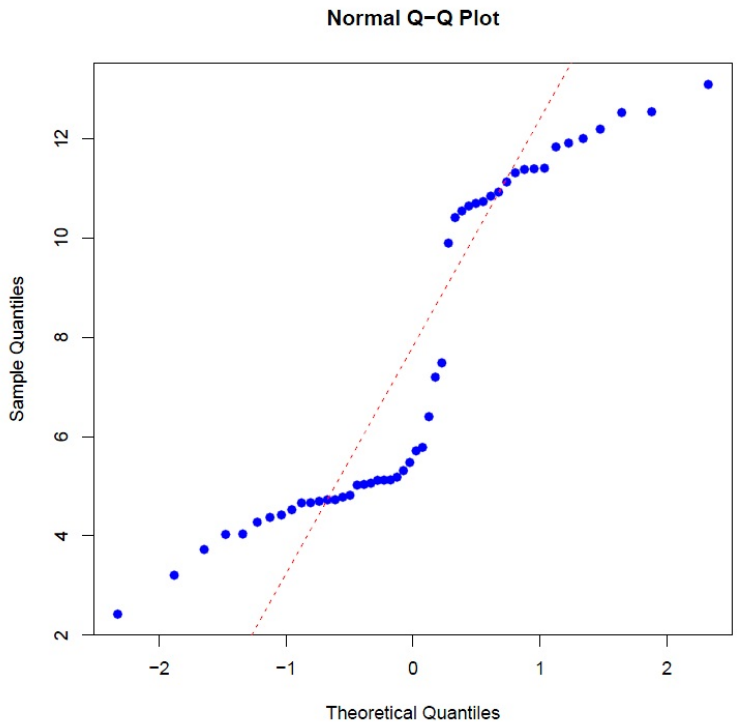
Asimetría positiva (cola derecha pesada)



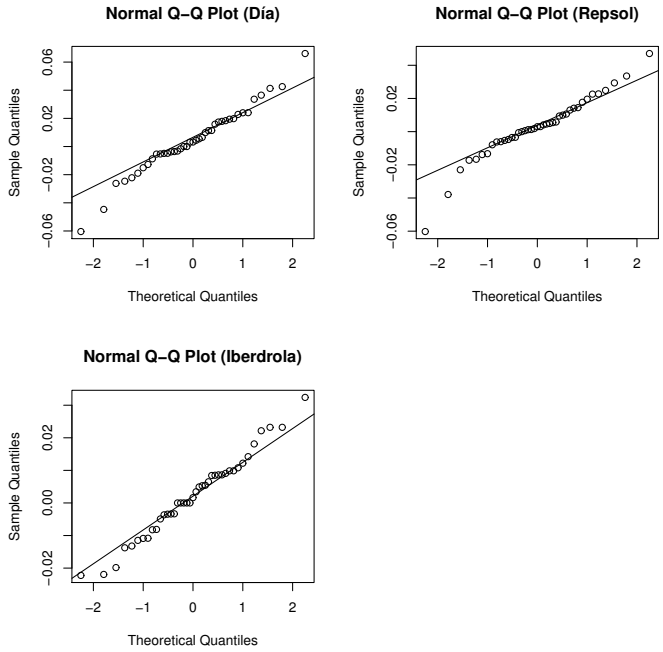
Asimetría negativa (cola izquierda pesada)



Distribuciones bimodales



Normalidad en datos univariantes: Q-Q plot



Contrastes de Normalidad multivariante

- Test de Mardia (Mardia, 1970). Extensión multivariante de los test de asimetría y curtosis.
- Test de Henze-Zirkler (Henze, N., & Zirkler, B.,1990). Basado en la función característica.
- Test de Royston (1992). Extensión multivariada del test de Shapiro-Wilks.
- Test de Doornik, J. A., & Hansen (2008). Basado en transformaciones de los coeficientes de asimetría y curtosis.
- E- statistic – Test de Székely, G. J., & Rizzo (2005). Basado en la distancia euclídea de los datos observados y los quantiles teóricos de la distribución normal multivariada.

Contrastes de Normalidad multivariante

Test de Mardia (Mardia, 1970)

- Extensión multivariante de los test de asimetría y curtosis.
- Supongamos que tenemos una muestra p-dimensional iid de tamaño $n : x_1, ..., x_n$

  $$\hat{\gamma}_{1,p} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} m_{ij}^3 \quad \hat{\gamma}_{2,p} = \frac{1}{n} \sum_{i=1}^{n} m_{ii}^3$$

  donde $m_{ij} = (x_i - \bar{X})' S^{-1} (x_j - \bar{X})$ es la distancia de Mahalanobis.
- $(n/6)\hat{\gamma}_{1,p} \sim \chi^2_{\frac{p(p+1)(p+2)}{6}}$
- La distribución del segundo estadístico viene dada por: $\hat{\gamma}_{2,p} \sim N(p(p+2), \sqrt{8p(p+2)/n})$
- Corrección en muestras pequeñas $(n < 20)$

  $$\frac{nk}{6}\hat{\gamma}_{1,p} \sim X^2_{\frac{p(p+1)(p+2)}{6}} \text{ donde } k = \frac{(p+1)(n+1)(n+3)}{n(n+1)(p+1) - 6}$$

Contrastes de Normalidad multivariante

Test de Henze-Zirkler (Henze y Zirkler, 1990)

- Compara la función carcterística empírica con la asociada a la distribución normal.
- La función característica es una función que toma valores complejos y viene dada por la siguiente expresión:

$$\varphi_X(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itX} f_X(x)$$

- Estdístico

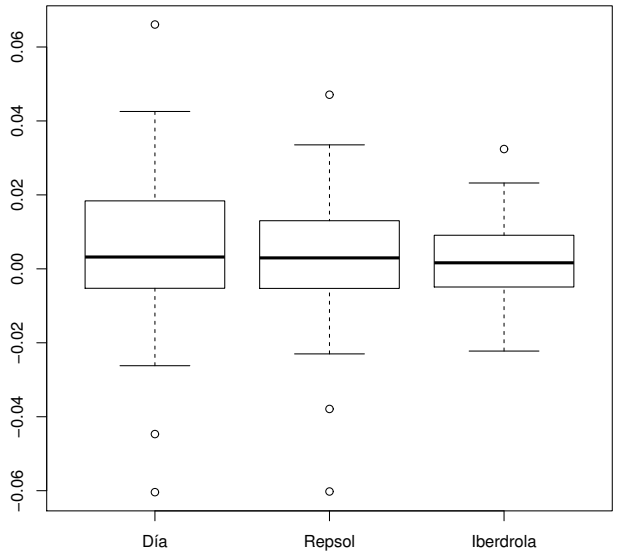
$$HZ = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{\frac{\beta^2 m_{ij}}{2}} - 2(1+\beta^2)^{\frac{-p}{2}} \sum_{i=1}^{n} e^{\frac{-\beta^2 m_{ii}}{2(1+\beta^2)}} + n(1+2\beta^2)^{\frac{-p}{2}}$$

- El estadístico HZ sigue un distibución lognormal de parámetros $\hat{\mu}$ y $\hat{\sigma}$.
- El estadístico de Wald para testar normalidad será entonces:

$$\frac{\log(HZ) - \log(\hat{\mu})}{\log(\hat{\sigma})}$$ que sigue una distribución normal estándar.

Datos atípicos (outliers)

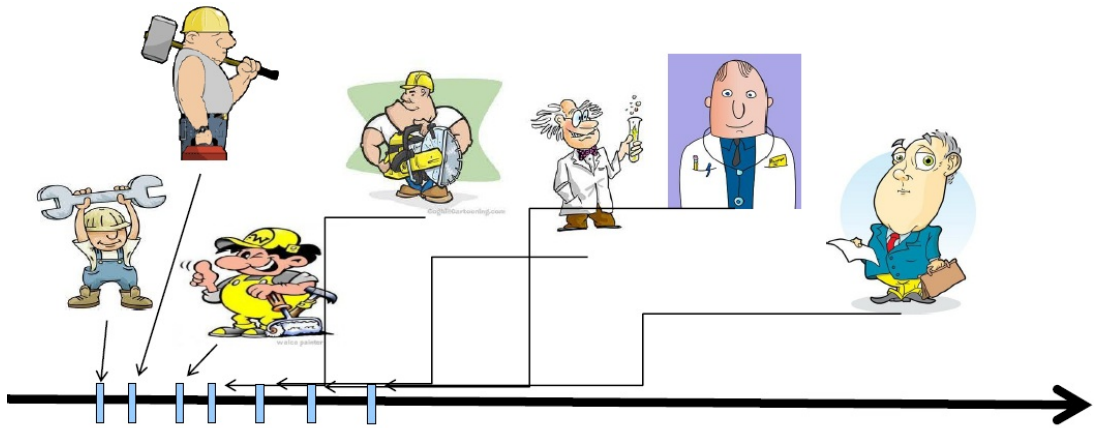
- Datos univariantes. Si suponemos normalidad en los datos, entonces un outlier será un dato que tipificado esté por encima de 2 en valor absoluto.
- Gráficos boxplot.



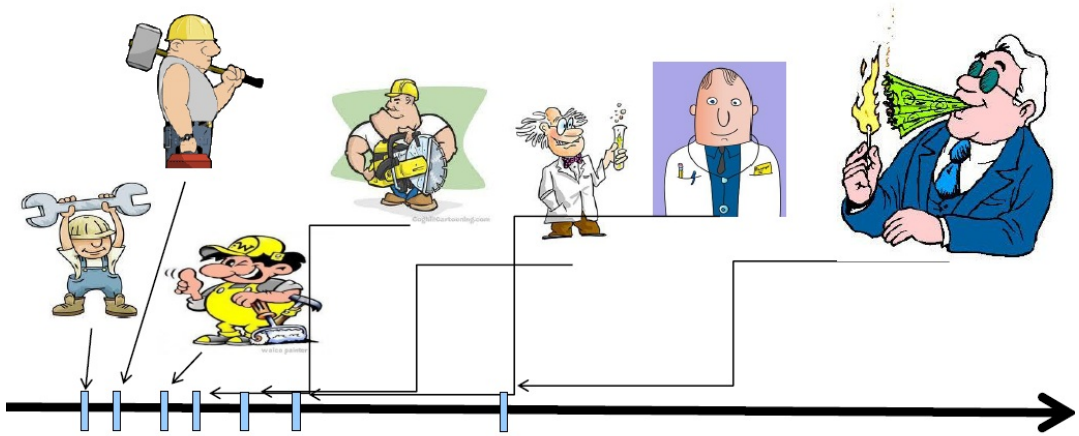
Outliers en datos multivariantes

- En ocasiones, pese a que no se observan outliers en las distribuciones unidimensionales de cada una de las variables por separado, es posible que sí existan outliers multidimenisonales. (Gráfico elipse)
- Distancia de Mahalanobis:
  1. Calculamos la distancia de Mahalanobis robusta (rMD(xi)),
  2. Calculamos el percentil 97.5 de la distribución chi-cuadrada,
  3. Declaramos rMD(xi) $>$ Q como posible outlier.
- Distancia de Mahalanobis 'ajustada' (Filzmoser et al., 2005).
  1. Calculamos la distancia de Mahalanobis robusta (rMD(xi)),
  2. Calculamos el percentil 97.5 de la distribución chi-cuadrada 'ajustada',
  3. Declaramos rMD(xi) $>$ AQ como posible outlier.

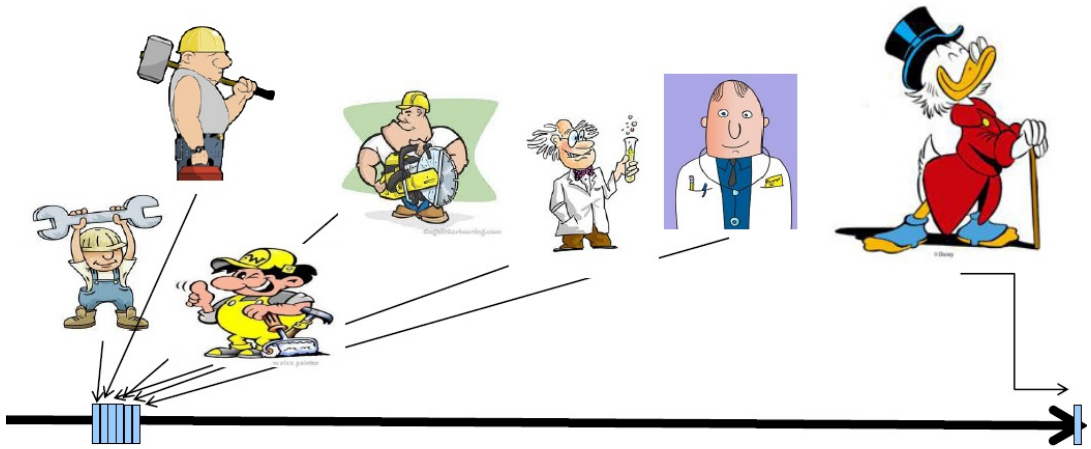
Outliers en datos multivariantes



Outliers en datos multivariantes



Outliers en datos multivariantes



Outliers en datos multivariantes

Distancia de Mahalanobis robusta:

- La distancia de Mahalanobis se usa para determinar los outliers pero a su vez la media y la varianza se ven afectadas por ellos.
- Robusto: modelar el patron general de los datos.
- Se selecciona un $h \in \left[\frac{n+p+1}{2}, n\right]$ $(h \approx 0.75n)$.
- El objetivo es encontrar la submuestra de ese tamaño que minimiza el determinante de la matriz de varianzas y covarianzas.
- Se computa la media y la varianza de dicha submuestra.
- Se calcula la distancia de Mahalanobis con dichos parámetros.
- Para $D_i^2 > X^2_{p,0.975}$ se declara el dato como outlier.

Referencias

- Doornik, J. A., & Hansen, H. (2008). An omnibus test for univariate and multivariate normality. Oxford Bulletin of Economics and Statistics, 70, 927-939.
- Filzmoser, P., Garrett, R. G., & Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. Computers & Geosciences, 31(5), 579-587.
- Henze, N., & Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. Communications in Statistics-Theory and Methods, 19(10), 3595-3617.
- Korkmaz, S., Goksuluk, D., & Zararsiz, G. MVN: An R Package for Assessing Multivariate Normality. Disponible en http://cran.r-project.org/web/packages/MVN/vignettes/MVN.pdf.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. Biometrika, 57(3), 519-530.
- Royston, P. (1992). Approximating the Shapiro-Wilk W-test for non-normality. Statistics and computing, 2(3), 117-119.
- Székely, G. J., & Rizzo, M. L. (2005). A new test for multivariate