# *NOVENA Y DÉCIMA CLASE*

Analytics AoZ

1. Pruebas semiparamétricas
   - ❑ Modelo robusto - ROS (Regression Orden in Statistics)
2. Correlación en Rstudio
   - ❑ Definiciones y Conceptos.
   - ❑ Correlación de Kendall.
   - ❑ Correlación de Spearman.
   - ❑ Correlación de Pearson.
3. Regresión Lineal:
   - ❑ Definiciones y Conceptos.
   - ❑ Supuesto de la Regresión Lineal.
   - ❑ Gráficos Adicionales para ver normalidad.
4. Ejercicio de Regresión Lineal Múltiple
5. Regresiones no Lineales
   - ❑ Regresión Cuadrática
   - ❑ Regresión Polinómica
   - ❑ Otros tipos.
6. Aplicado en Rstudio a una Base de Datos Geológica (Hidrogeológica, Mina).

**REPASO7 y REPASO8** : EJERCICIO PARA AFIANZAR LO APRENDIDO

# PRUEBAS SEMIPARAMÉTRICAS

Analytics AoZ

La estadística es la ciencia que permite tomar decisiones **en situaciones de incertidumbre**.

El propósito de las ***pruebas de hipótesis*** es obtener conclusiones sobre los parámetros de la población (media, proporción u otros) basándonos en los resultados obtenidos en ***muestras aleatorias***.

Las ***pruebas paramétricas*** tienen en cuenta los parámetros en las poblaciones. Para su utilización es necesario de una serie de requisitos o supuestos.

Las ***pruebas no paramétricas*** no necesitan estas condiciones tan rigurosas para su aplicación.

Cuando se dan las condiciones de aplicación, las pruebas paramétricas tienen más potencia que las no paramétricas, pero cuando esto no es así, el riesgo alfa puede ser mayor que el especificado de antemano.

Una estrategia posible sería utilizar siempre pruebas no paramétricas ya que, si se dan las condiciones de la aplicación, la pérdida de potencia no es muy grande y, si no se dan, son los métodos que deben emplearse.

# PRUEBAS SEMIPARAMÉTRICAS

Analytics AoZ

Otra alternativa es la utilización de los denominados **métodos robustos**. Estos métodos son menos potentes que los paramétricos, pero se muestran superiores a los métodos no paramétricos clásicos. La principal de sus ventajas es que ***no se afecta por la existencia de datos anómalos***.

Acordarse de que si se asume que los datos extremos son erróneos, lo cual no es admisible se pierde información y se logra modelar o estimar lo que se desea. Un enfoque más adecuado es comprobar la verificada del dato, alguna sustitución de valor promedio justificada, una interpolación de un dato con respecto a loa adyacentes u otros métodos similares a los  empleados cuando existen datos ausentes o debajo del límite de detección.

En este caso los **métodos robustos** pueden ser de gran utilidad para la realización de inferencias sin tener en cuenta que “depura” los datos extremos.

## Métodos robustos para estimar medidas de centralización

**TABLA 1** Métodos robustos para el cálculo de medidas de posición

| Estimador | Estrategia | Resultado con los datos (1, 2, 3, 4, 5, 100) (media aritmética: 19,2) |
|---|---|---|
| Media $\alpha$-winsorizada muestral | Se sustituye un determinado porcentaje, $\alpha$, (20% generalmente) de valores extremos a cada lado de la muestra por el valor más próximo no sustituido | 3,5 |
| 2 Media $\alpha$-recortada muestral | Se eliminan las k observaciones extremas de cada lado, en lugar de winsorizarlas, calculando la media aritmética de las observaciones restantes | 3,5 |
| Mediana muestral | Divide la distribución en dos partes con el mismo número de elementos | 3,5 |
| 1 Estimador de Huber | Se encuentra dentro de los denominados M-estimadores, que generalizan al estimador de máxima verosimilitud con buenas propiedades de robustez y eficiencia. En este caso se descartan las observaciones que sean mayores (o menores) a una constante | 3,57 |

Un problema conocido es el cálculo de las medidas de posición o centralización en el cual *los datos no siguen una distribución normal*; en esos casos la media no es buen estimador del promedio de los datos.

## Métodos robustos para estimar medidas de dispersión

**TABLA 2** Métodos robustos para el cálculo de medidas de dispersión

| Estimador | Estrategia | Resultado con los datos (1, 2, 3, 4, 5, 100) (media aritmética: 39,6) |
|---|---|---|
| Desviación absoluta mediana estandarizada | Es la mediana de las desviaciones absolutas a la mediana | 2,2 |
| Cuasi desviación típica $\alpha$-winsorizada muestral | En la que se sustituye un determinado porcentaje de valores extremos a cada lado de la muestra por el valor más próximo no sustituido | 1,38 |

Las medidas de dispersión clásicas (varianza y desviación típica) se ven afectadas por las mismas limitaciones que las medidas de posición.

# Métodos robustos para el contraste de hipótesis

Analytics AoZ

**TABLA 3 Métodos robustos para el contraste de hipótesis**

| |
|---|
| Contrastes de hipótesis con una muestra unidimensional |
| Contrastes de hipótesis con dos muestras unidimensionales utilizando intervalos y test basados en medias α-recortadas muestrales |
| Generalización robusta del test de Wilcoxon-Mann-Whitney para datos independientes o apareados |
| Métodos robustos para el análisis de la variancia con uno o varios factores y las comparaciones múltiples entre subgrupos |
| Generalización robusta del test de Kruskal-Wallis |
| Métodos robustos del análisis de la variancia con medidas repetidas |
| Análisis robustos de la correlación y estimación multivariante |
| Análisis robusto de regresión múltiple y de la covariancia |

Con los parámetros antes referidos, se pueden construir intervalos de confianza robustos y realizar contrastes de hipótesis.

El **test de Yuen** es una alternativa de las pruebas t de Student y U de Mann-Whitney para tamaños muestrales pequeños y distribuciones no normales.

**TABLA 4 Ejemplo para la comparación de dos medias**

| Grupo | Días de estancia | Media | Mediana | K-S-L |
|---|---|---|---|---|
| Fallecidos (n = 21) | 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 5, 14, 30, 31, 45 | 7,24 | 2,00 | p < 0,0001 |
| Supervivientes (n = 53) | 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 11, 13, 13, 14, 15, 15, 15, 18, 21, 25, 25, 27, 27, 30, 31, 33, 60, 66, 88 | 11,74 | 3,00 | p < 0,0001 |

K-S-L: prueba de Kolmogorov-Smirnov, con corrección de Lilliefors para la comprobación del supuesto de normalidad.

**TABLA 5 Resultados de las distintas pruebas para la comparación de dos medias**

| Prueba de contraste de hipótesis | Nivel de significación |
|---|---|
| Test de la t de Student | 0,29 |
| Test de Mann-Whitney | 0,07 |
| Test de Yuen (media α-recortada) | 0,04 |

Para la comparación de más de dos medias cuando no se ´puede comparar utilizando el Análisis de Varianza (ANOVA) clásico, debemos recurrir al clásico no paramétrico de Kruskal-Wallis o, mejor a **métodos robustos como la generalización robusta de Welch**, que utiliza medias α-recortadas muestrales.

**TABLA 6** Ejemplo para la comparación de más de dos medias

| Grupo | Días de estancia | Media | Mediana | K-S-L |
|---|---|---|---|---|
| Jóvenes (n = 28) | 8, 30, 55, 4, 3, 1, 1, 2, 2, 1, 2, 2, 1, 3, 1, 1, 2, 3, 2, 1, 2, 1, 3, 4, 1, 2, 3, 4 | 5,18 | 2,00 | p < 0,0001 |
| Maduros (n = 23) | 1, 1, 1, 25, 13, 25, 4, 5, 4, 2, 1, 2, 33, 3, 2, 3, 2, 3, 1, 66, 11, 1, 3 | 9,22 | 3,00 | p < 0,0001 |
| Ancianos (n = 23) | 1, 15, 25, 1, 31, 2, 11, 88, 3, 21, 60, 3, 1, 5, 13, 2, 1, 2, 1, 3, 27, 3, 1, 27 | 15,48 | 3,00 | p = 0,0012 |

K-S-L: prueba de Kolmogorov-Smirnov, con corrección de Lilliefors para la comprobación del supuesto de normalidad.

**TABLA 7** Resultados de las distintas pruebas para la comparación de más de dos medias

| Prueba de contraste de hipótesis | Nivel de significación |
|---|---|
| Test de ANOVA | 0,08 |
| Test de Kruskal-Wallis | 0,09 |
| Método robusto media (α-recortada) | 0,02 |

# Análisis de Correlación y Regresión Lineal

Analytics AoZ

En el caso de la correlación lineal de Pearson no sea adecuada para el análisis, deberíamos recurrir a métodos no paramétricos (Spearman) o, mejor a métodos robustos como **el coeficiente de porcentaje ajustado poblacional  el estimador robusto de regresión medio biponderado**.

**TABLA 8** **Ejemplo para regresión lineal**

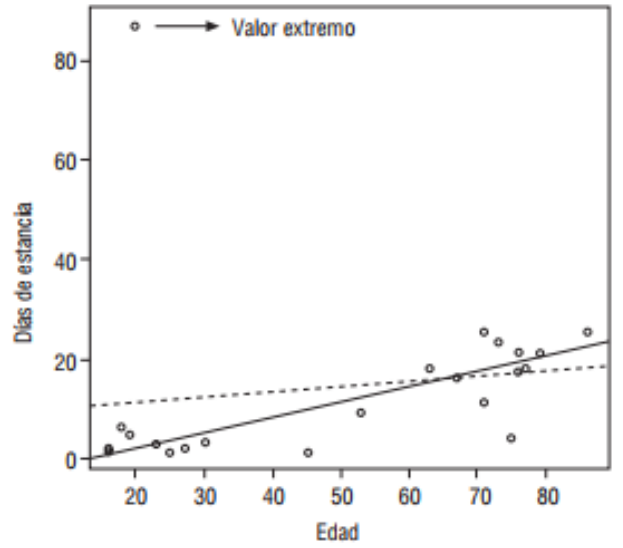
| Variable | Días de estancia | Media | Mediana | K-S-L |
|---|---|---|---|---|
| X (edad) (n = 22) | 63, 79, 53, 20, 23, 18, 19, 16, 45, 30, 16, 67, 71, 73, 71, 76, 77, 75, 27, 86, 76, 25 | 50,27 | 58,00 | p = 0,04 |
| Y (estancia) (n = 22) | 18, 21, 9, 87, 3, 6, 5, 2, 1, 3, 1, 16, 25, 23, 11, 21, 18, 4, 2, 25, 17, 1 | 14,50 | 10,00 | p = 0,002 |

K-S-L: prueba de Kolmogorov-Smirnov, con corrección de Lilliefors para la comprobación del supuesto de normalidad.

**TABLA 9** **Resultados para regresión lineal**

| Prueba de contraste de hipótesis | Coeficiente de correlación | Nivel de significación |
|---|---|---|
| Pearson | 0,15 | 0,50 |
| Spearman | 0,61 | 0,002 |
| Porcentaje ajustado poblacional | 0,70 | 0,0003 |

**FIGURA 1** Rectas de regresión obtenidas por mínimos cuadrados y por el estimador robusto de regresión medio biponderado.




Línea continua: ajuste mediante el estimador robusto de regresión medio biponderado
Línea de trazos discontinuos: ajuste mediante mínimos cuadrados

# MODELO SEMIPARAMÉTRICO ROS (ROBUST ORDEN IN STATISTICS) CASO: COMPLETAR DATOS DE RECURSOS HÍDRICOS EN METALES O NO METALES DISUELTOS O TOTALES DEBAJO DEL LIMITE DE DETECCIÓN

Analytics AoZ

NADA for R
A contributed package for
censored environmental data

https://cran.r-project.org/web/packages/NADA/NADA.pdf

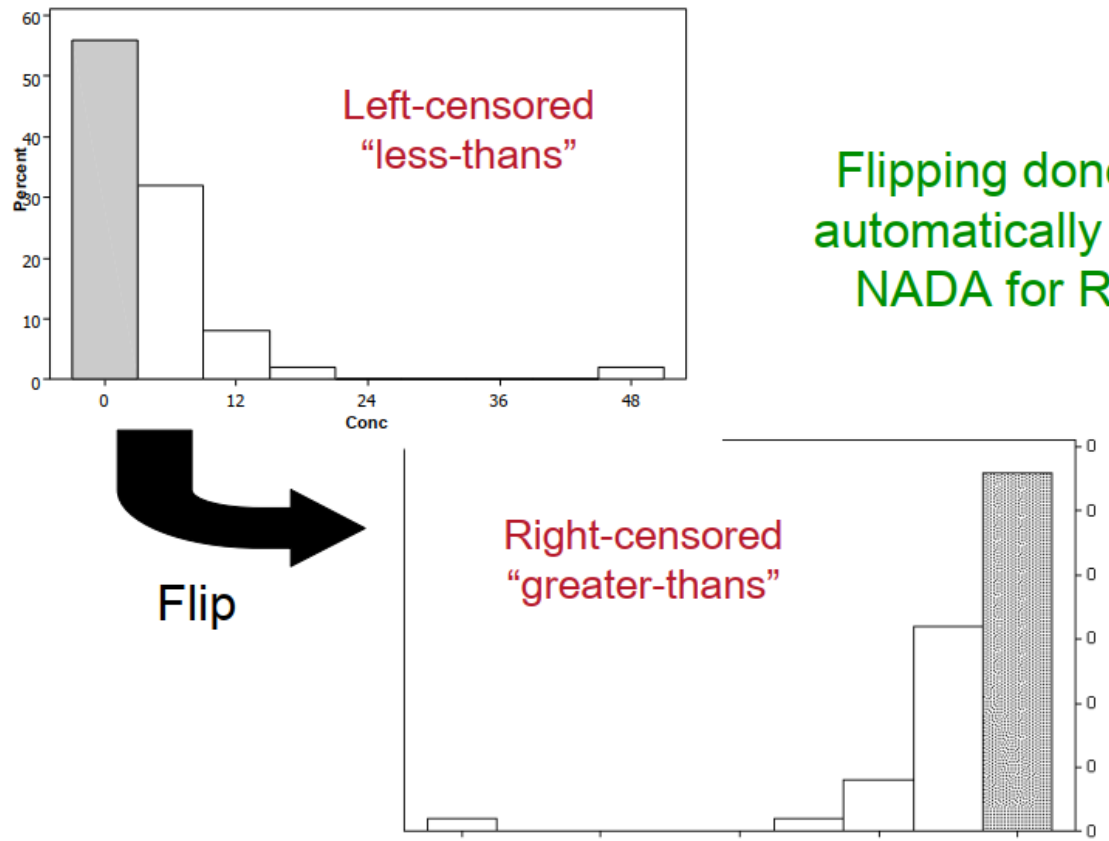
Dennis Helsel
Practical Stats

https://cran.r-project.org/web/packages/nortest/index.html

Lopaka (Rob) Lee
U.S. Geological Survey

# Three Valid Approaches for the Analysis of Censored Data

1. Parametric methods. Assume data follow a specific distribution.
   - Maximum likelihood estimation (MLE)
2. "Robust" methods
   - Regression on Order Statistics (ROS)
3. Nonparametric methods. Based on percentiles, ranks.
   - Kaplan-Meier
   - Wilcoxon score tests
   - Kendall's tau




Flipping done automatically in NADA for R

# R function to compute ROS test



To perform one-sample ROS-test, the R function **ros**() can be used as follow:

```
ros(obs, censored, forwardT="log", reverseT="exp", na.action)
```

**obs :**A numeric vector of observations. This includes both censored and uncensored observations.

**censored:** A logical vector indicating TRUE where an observation in obs is censored (a less-than value) and FALSE otherwise.

**forward:** A name of a function to use for transformation prior to performing the ROS fit. Defaults to log.

**reverseT:** A name of a function to use for reversing the transformation after performing the ROS fit. Defaults to exp.

**na.action:** A function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.omit if that is unset. Another possible value is NULL, no action.

Analytics AoZ

# “LO QUE ESCUCHO LO OLVIDO. LO QUE VEO LO RECUERDO. PERO LO QUE HAGO, LO ENTIENDO.”

**VAMOS A RESOLVER UN EJERCICIO**

# BIBLIOGRAFIA ESPECÍFICA


- Ramalle-Gómara, E., & De Llano, J. A. (2003). Utilización de métodos robustos en la estadística inferencial. *Atención Primaria*, *32*(3), 177.
- Helsel, D., & Lee, L. (2006, August). Analysis of environmental data with nondetects. In *Continuing Education Workshop at the Joint Statistical Meetings. American Statistical Association, Seattle, WA*.
- https://www.practicalstats.com/resources/NADA-resources/NADAforR_Examples.pdf
- Helsel, D. R., & Cohn, T. A. (1988). Estimation of descriptive statistics for multiply censored water quality data. *Water Resources Research*, *24*(12), 1997-2004.
- https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/WR024i012p01997
- Helsel, D. R. (2011). *Statistics for censored environmental data using Minitab and R* (Vol. 77). John Wiley & Sons.