Ploteando No Detectados (ND)

Presentation by Ing. A.Otiniano

Ing. A.Otiniano

UNI

2024-07-08

Analytics AoZ

Analytics AoZ

Section 1

Objetivos

Analytics AoZ

Objetivos

- Conocer y usar los **boxplots**.

Analytics AoZ

Objetivos

- Conocer y usar los **boxplots**.
- Conocer y usar los **X-Y scatterplots**.

Analytics AoZ

Objetivos

- Conocer y usar los **boxplots**.
- Conocer y usar los **X-Y scatterplots**.
- Interpretar y usar las **Probability Density Functions (pdf)**.

Analytics AoZ

Objetivos

- Conocer y usar los **boxplots**.
- Conocer y usar los **X-Y scatterplots**.
- Interpretar y usar las **Probability Density Functions (pdf)**.
- Interpretar y usar las **Cumulative Distribution Functions (cdf)**.

Analytics AoZ

Objetivos

- Conocer y usar los **boxplots**.
- Conocer y usar los **X-Y scatterplots**.
- Interpretar y usar las **Probability Density Functions (pdf)**.
- Interpretar y usar las **Cumulative Distribution Functions (cdf)**.
- Conocer los **plots de probabilidad**.

Analytics AoZ

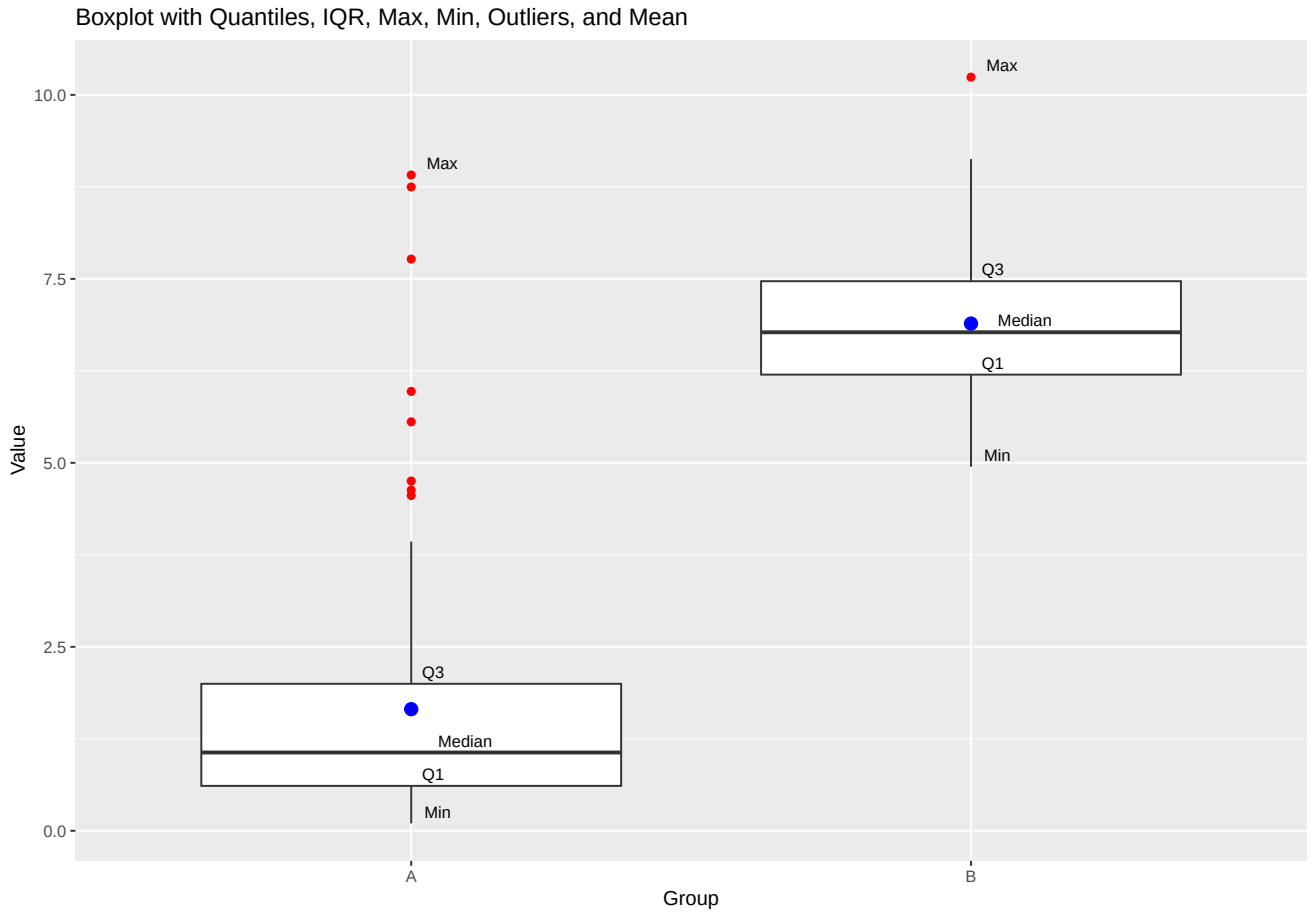
Section 2

Boxplots

Analytics AoZ

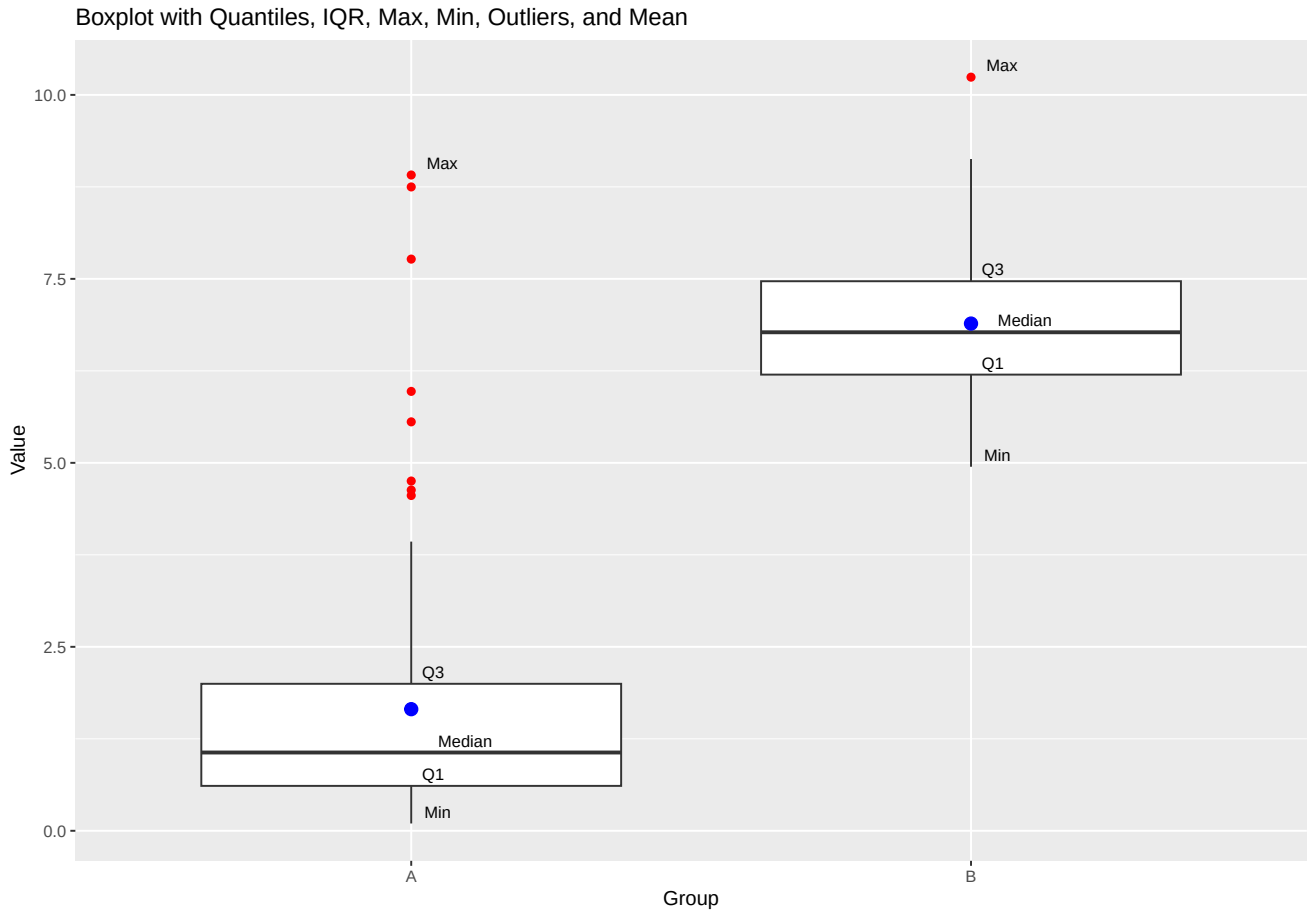
Boxplots estructura

1. Centrado



Boxplots estructura

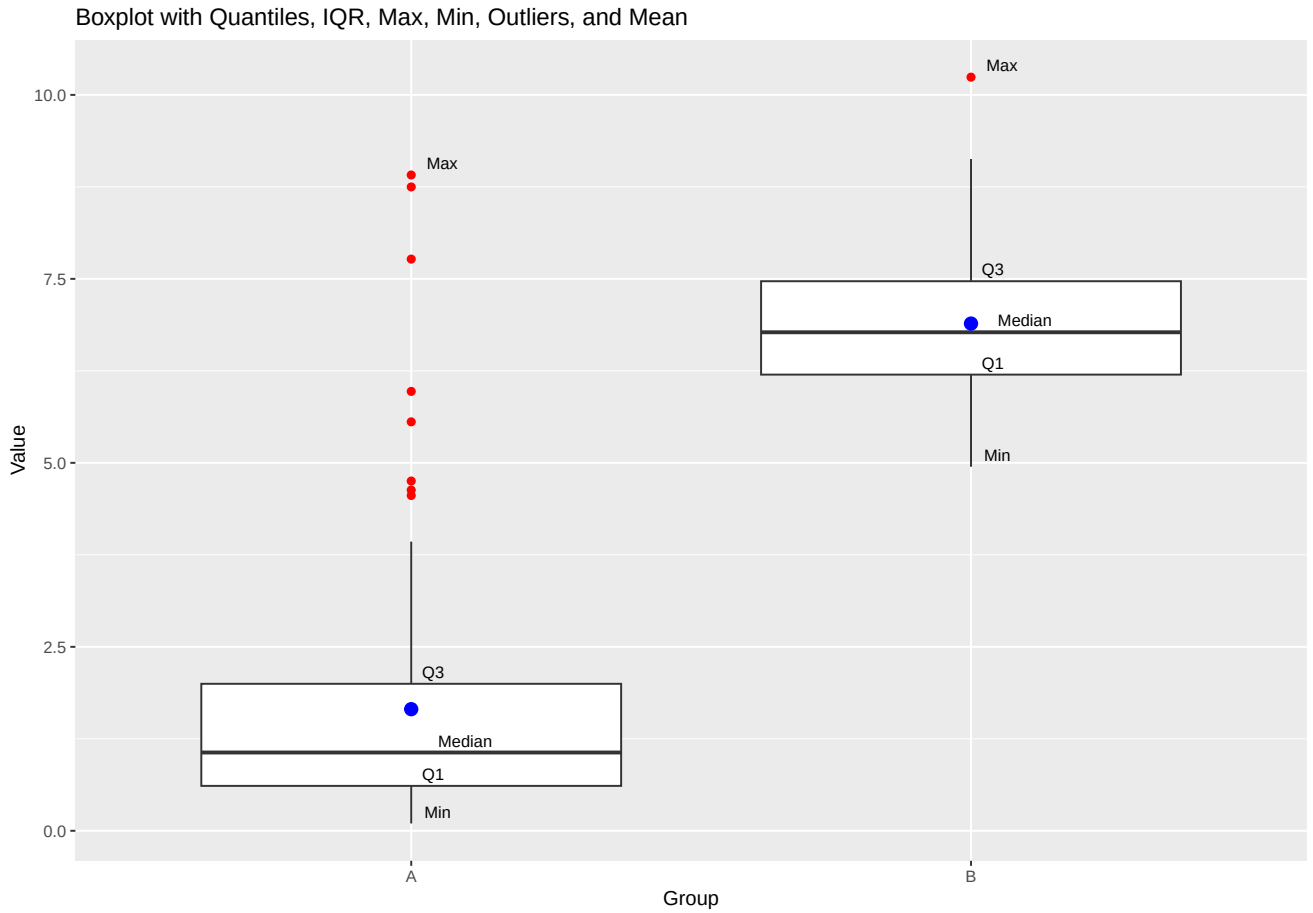
1. Centrado
2. Variabilidad (IQR)



Analytics AoZ

Boxplots estructura

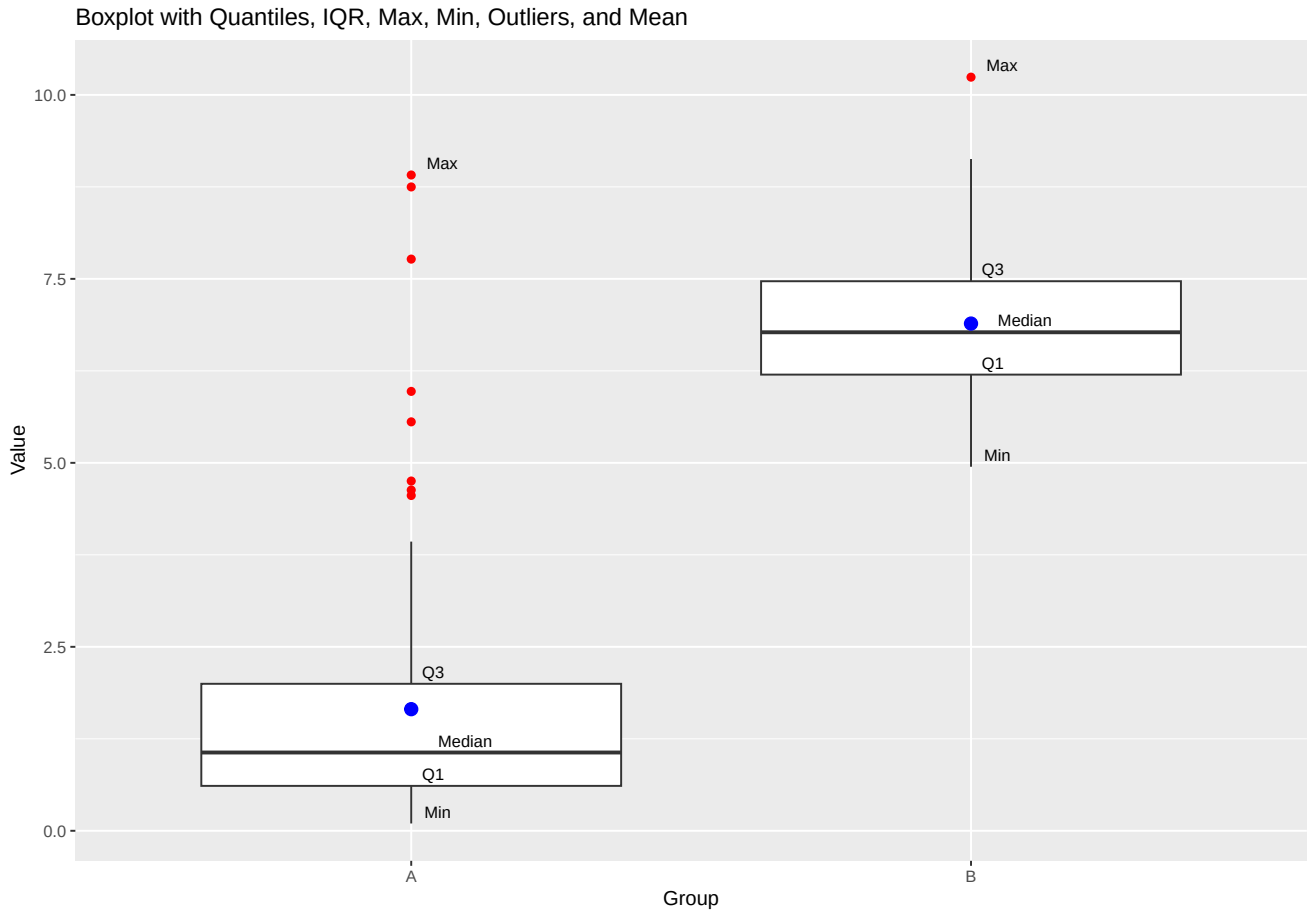
1. Centrado
2. Variabilidad (IQR)
3. *Asimetría*



Analytics AoZ

Boxplots estructura

1. Centrado
2. Variabilidad (*IQR*)
3. *Asimetría*
4. Outliers - Faroutliers

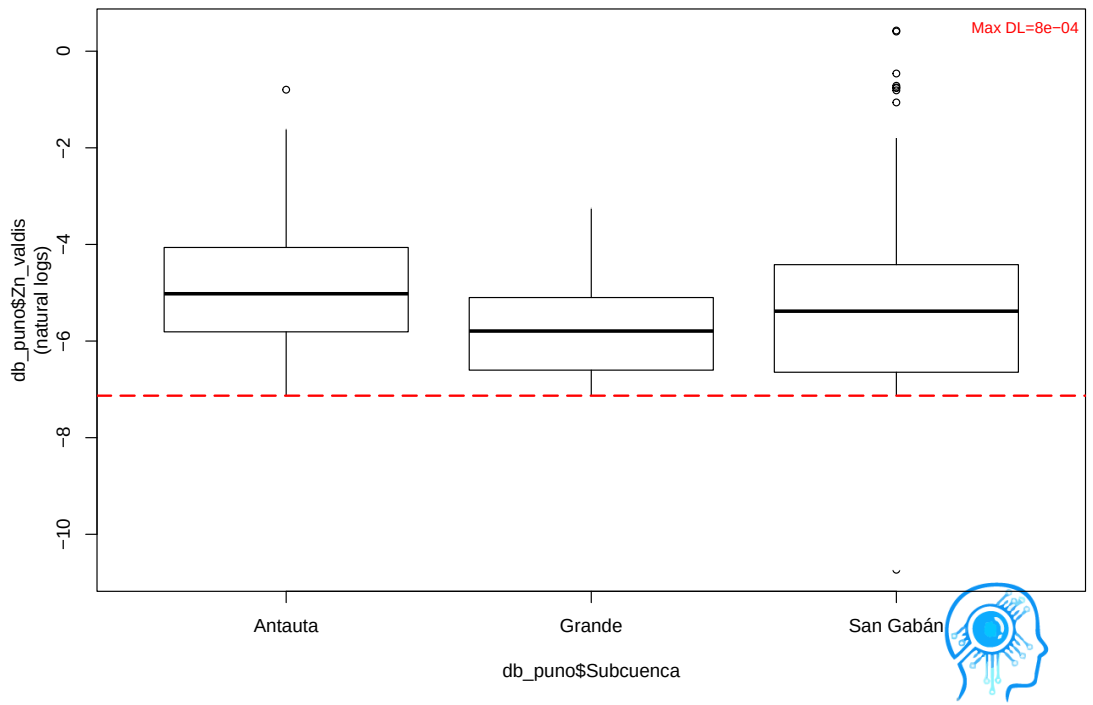


Analytics AoZ

Boxplot con ND

```
cboxplot(mina.Zn,mina.cen.Zn,xgroup=Subcuenca,LOG=TRUE)
```

a) Buena forma de ilustrar diferencias entre grupos.



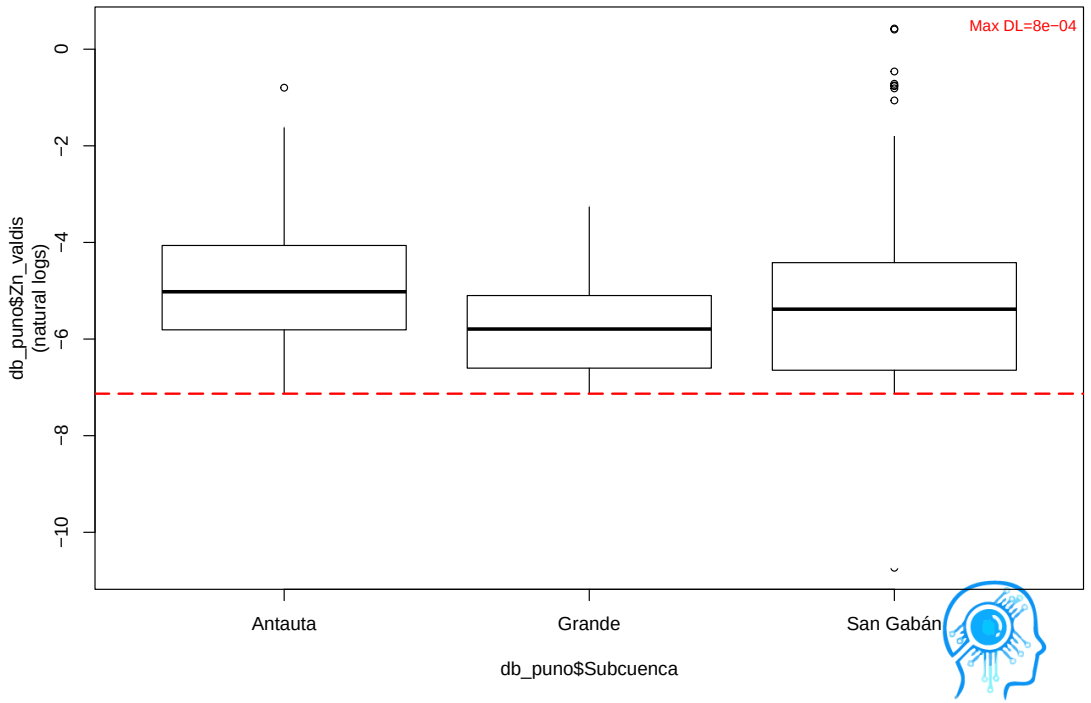
Analytics AoZ

Boxplot con ND

```
cboxplot(mina.Zn,mina.cen.Zn,xgroup=Subcuenca,LOG=TRUE)
```

a) Buena forma de ilustrar diferencias entre grupos.

b) Sobre el máximo Ld, es identico a un boxplot que podría haber dibujado para la misma data sin límite de detección.



Analytics AoZ

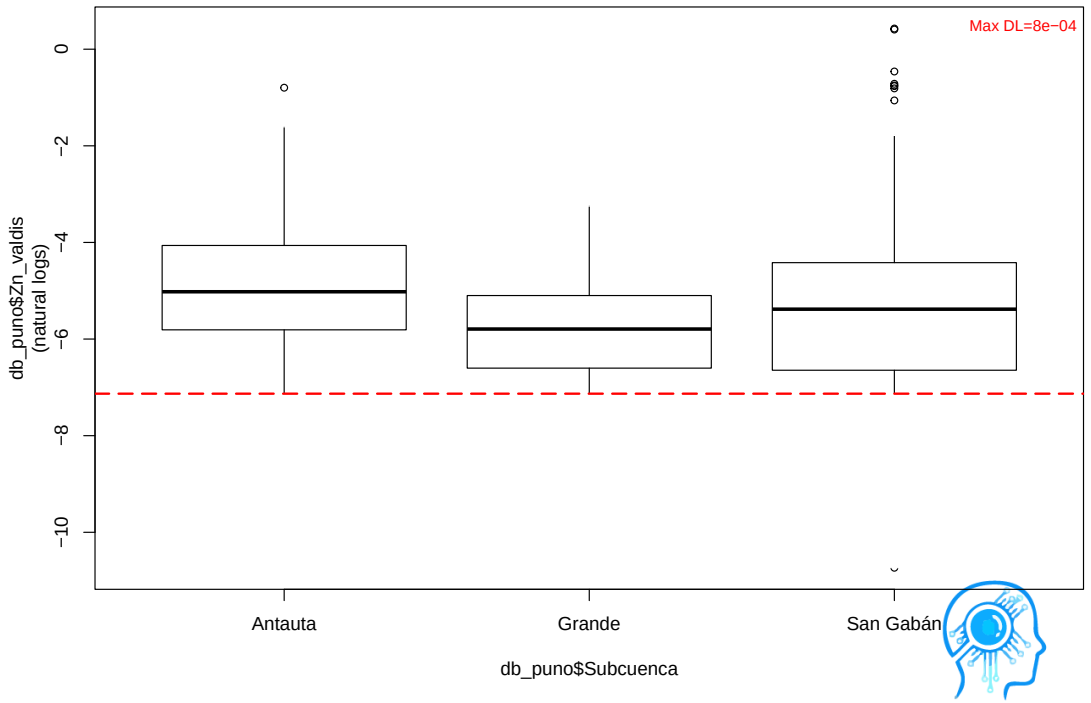
Boxplot con ND

```
cboxplot(mina.Zn,mina.cen.Zn,xgroup=Subcuenca,LOG=TRUE)
```

a) Buena forma de ilustrar diferencias entre grupos.

b) Sobre el máximo Ld, es identico a un boxplot que podría haber dibujado para la misma data sin límite de detección.

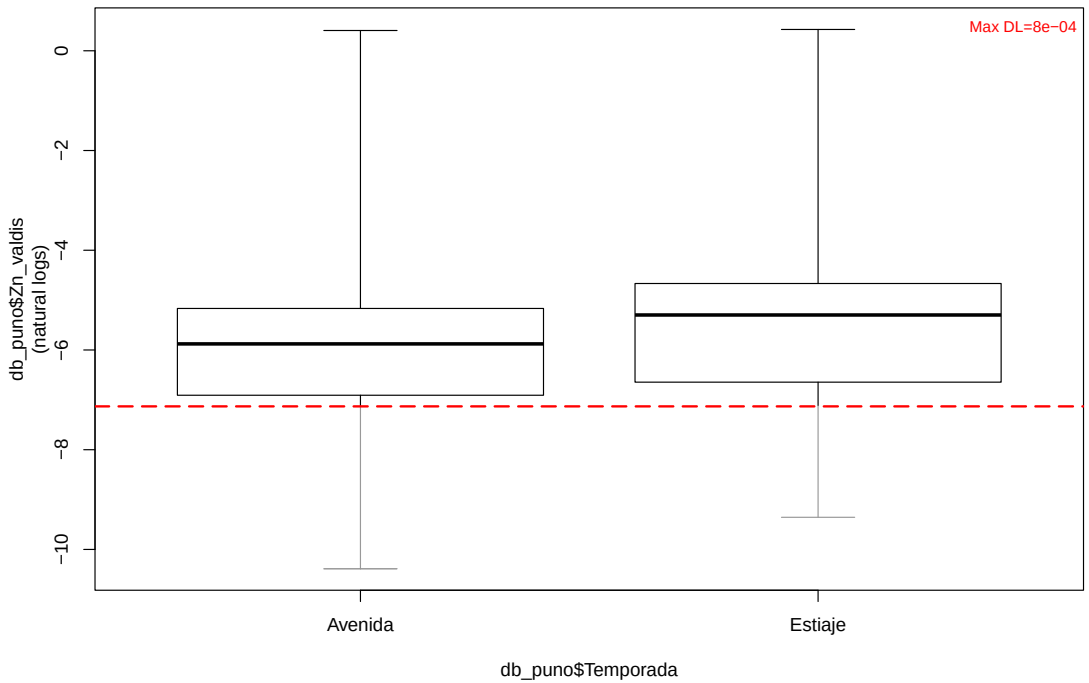
c) No estimaciones debajo del **maxLd** son mostradas por defecto.



Analytics AoZ

Boxplot con ND2

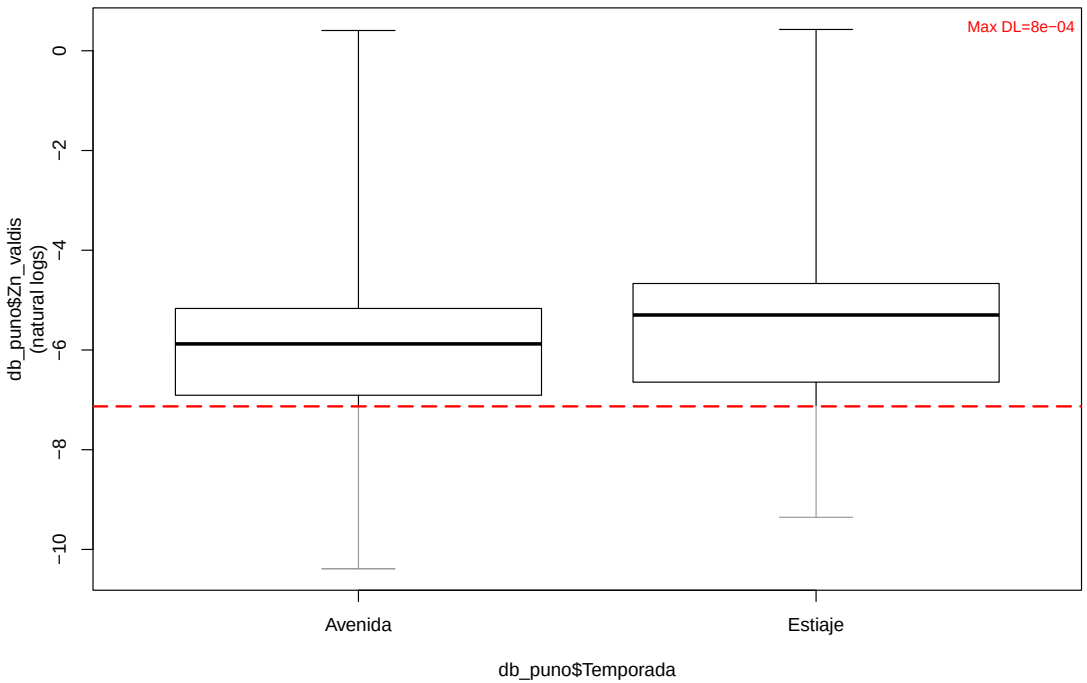
1. Porción debajo del max Ld es estimado con ROS y mostrada con `shown=TRUE`.



Analytics AoZ

Boxplot con ND2

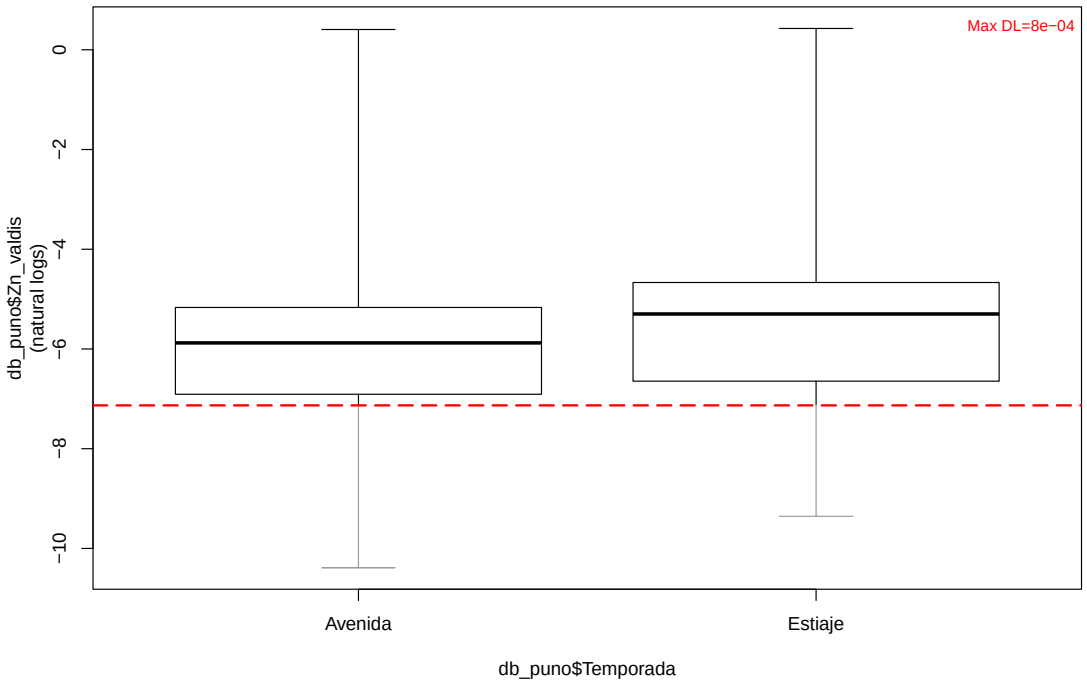
1. Porción debajo del max Ld es estimado con ROS y mostrada con `shown=TRUE`.
2. Estimados son sombreados con gris para indicar incertidumbre.



Analytics AoZ

Boxplot con ND2

1. Porción debajo del max Ld es estimado con ROS y mostrada con `shown=TRUE`.
2. Estimados son sombreados con gris para indicar incertidumbre.
3. No usar `mimax=TRUE` da por defecto boxplot con outliers.



Analytics AoZ

Section 3

X-Y Scatterplots

Analytics AoZ

X-Y Scatterplots with ND

```
cenxyplot(Zn, Zn_cen, Fe, Fe_cen, log="xy")
```

1. Puntos detectados son ploteados individualmente.

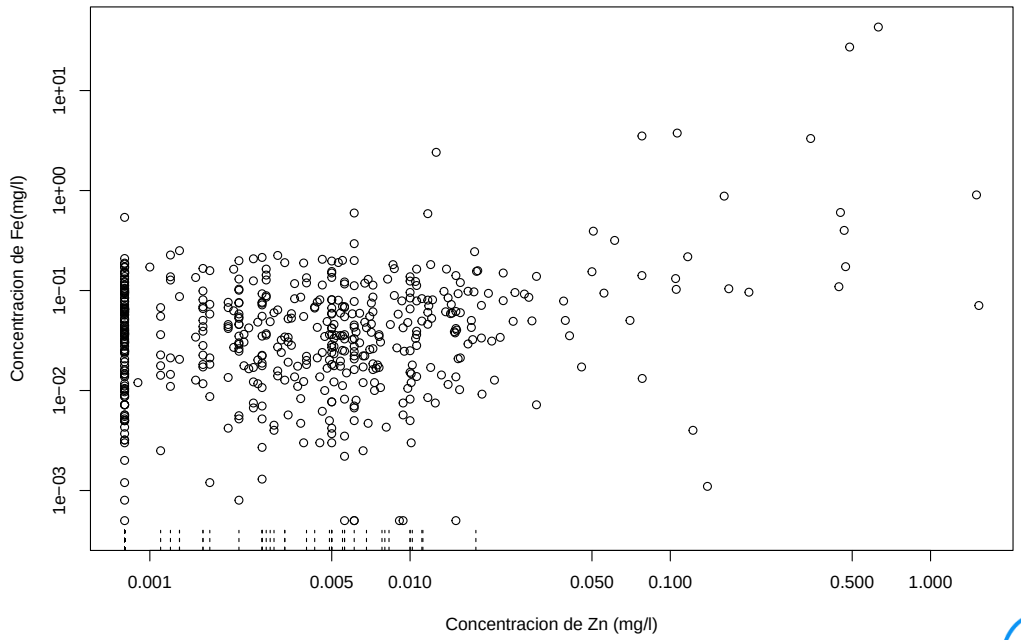


Figure 1: X-Y Scatterplot LogscaleXY

Analytics AoZ

X-Y Scatterplots with ND

```
cenxyplot(Zn, Zn_cen, Fe, Fe_cen, log="xy")
```

1. Puntos detectados son ploteados individualmente.
2. No detectados son mostrado como un intervalo (líneas punteadas).

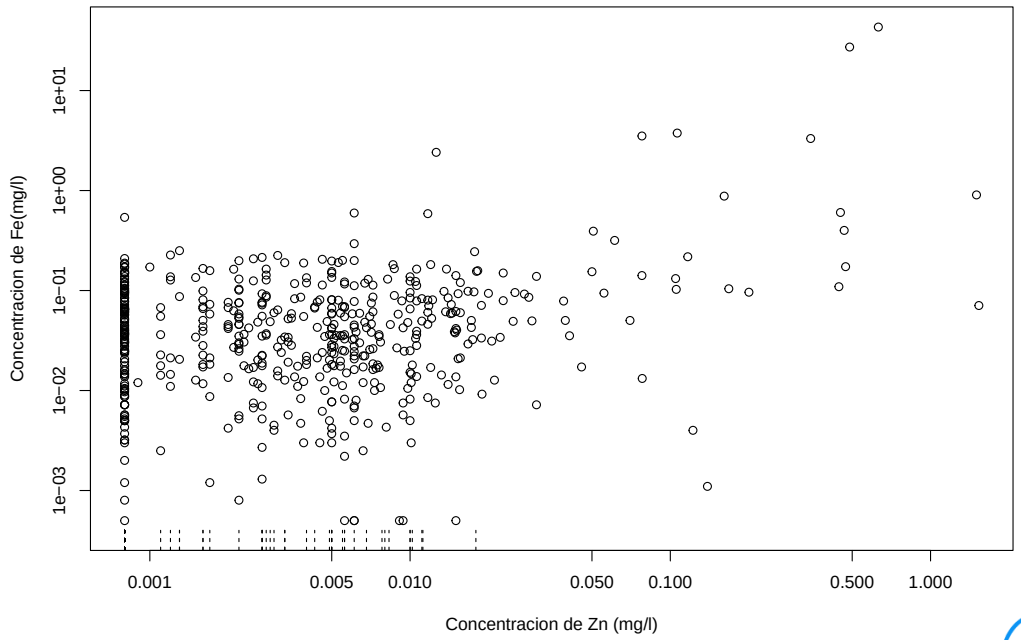


Figure 1: X-Y Scatterplot LogscaleXY

Analytics AoZ

X-Y Scatterplots with ND

```
cenxyplot(Zn, Zn_cen, Fe, Fe_cen, log="xy")
```

1. Puntos detectados son ploteados individualmente.
2. No detectados son mostrado como un intervalo (líneas punteadas).
3. El eje x e y están en escala logarítmica.

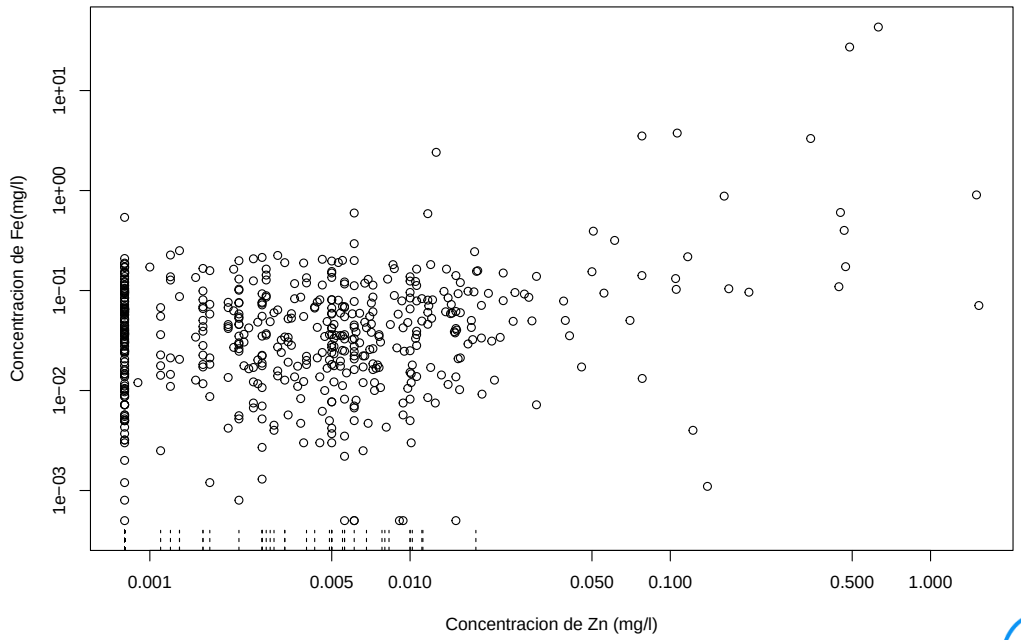
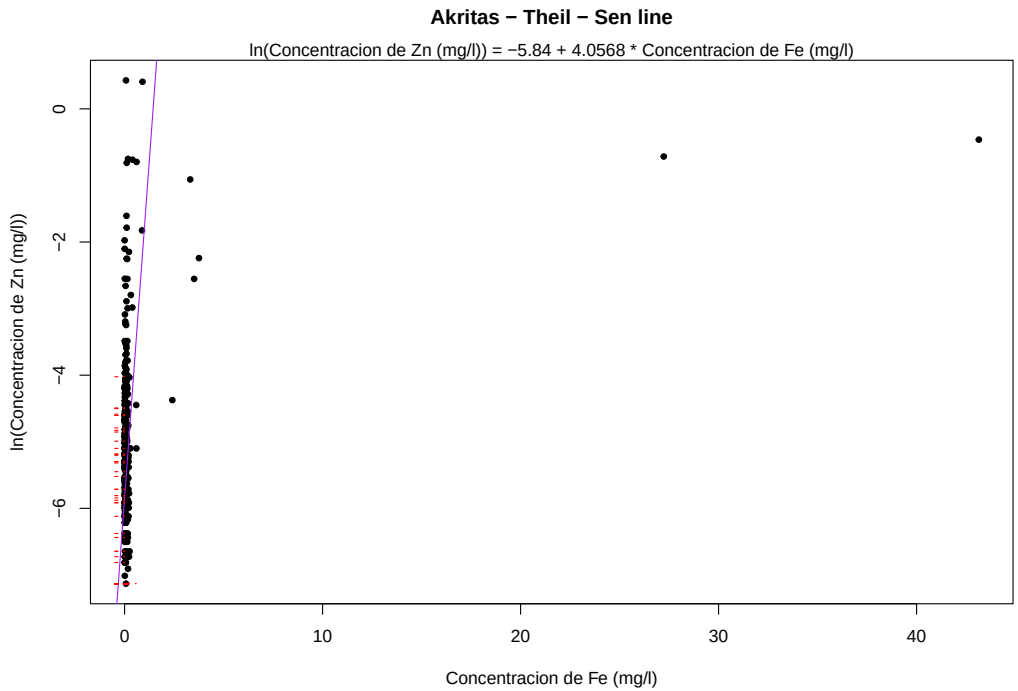


Figure 1: X-Y Scatterplot LogscaleXY

Analytics AoZ

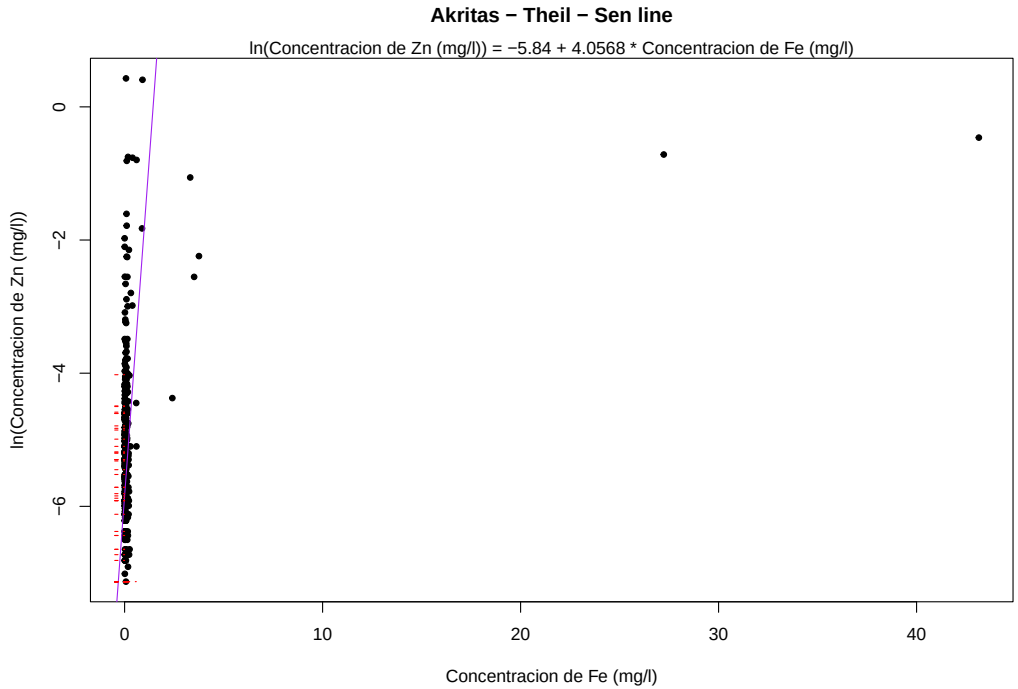
X-Y Scatterplots with ATS line

1. El formato es:
   `ATS(y,ycen, x, xcen, LOG=FALSE)`



X-Y Scatterplots with ATS line

1. El formato es: `ATS(y,ycen, x, xcen, LOG=FALSE)`
2. Existen algunos no detectados como lineas punteadas en la base.



Analytics AoZ

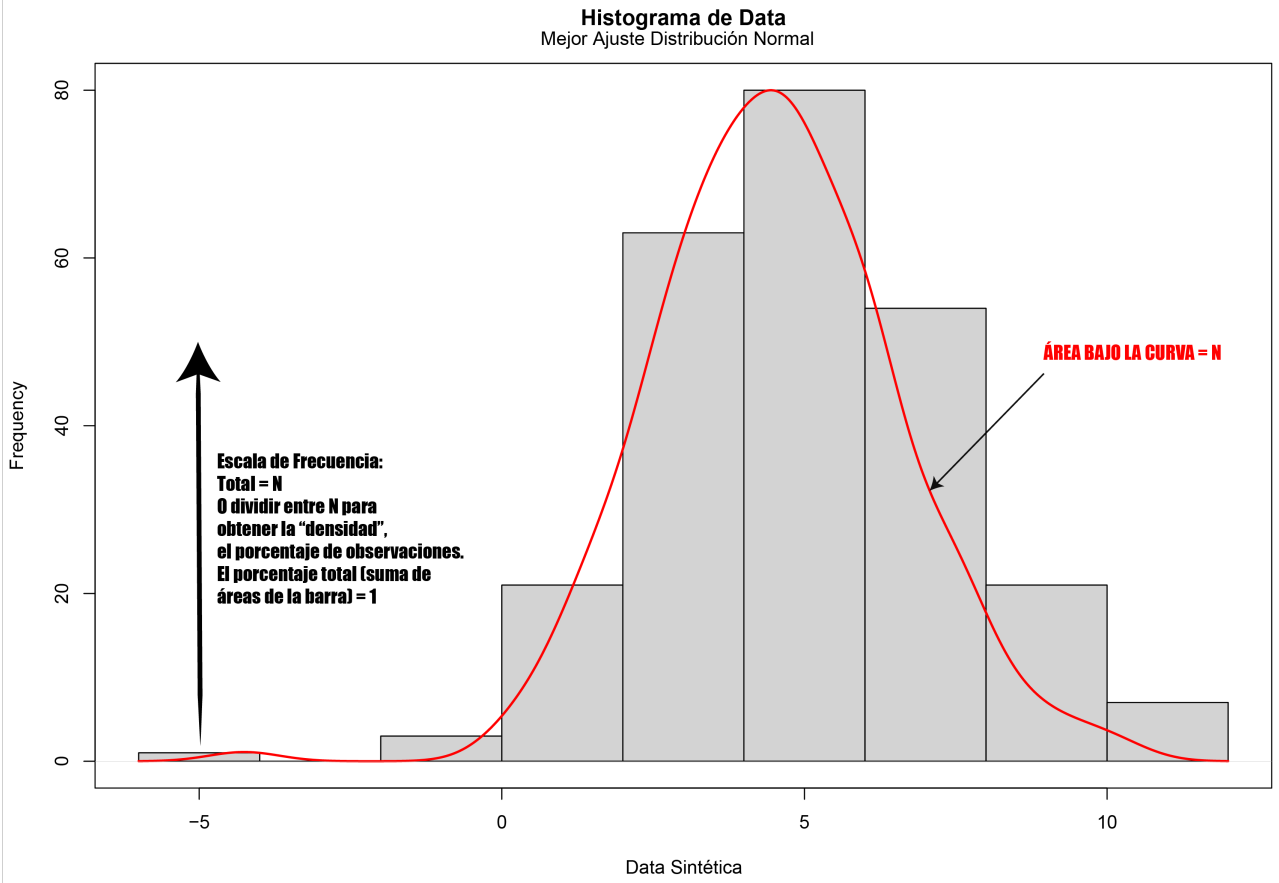
Section 4

Función de Probabilidad de Densidad (pdf)

Analytics AoZ

PDF

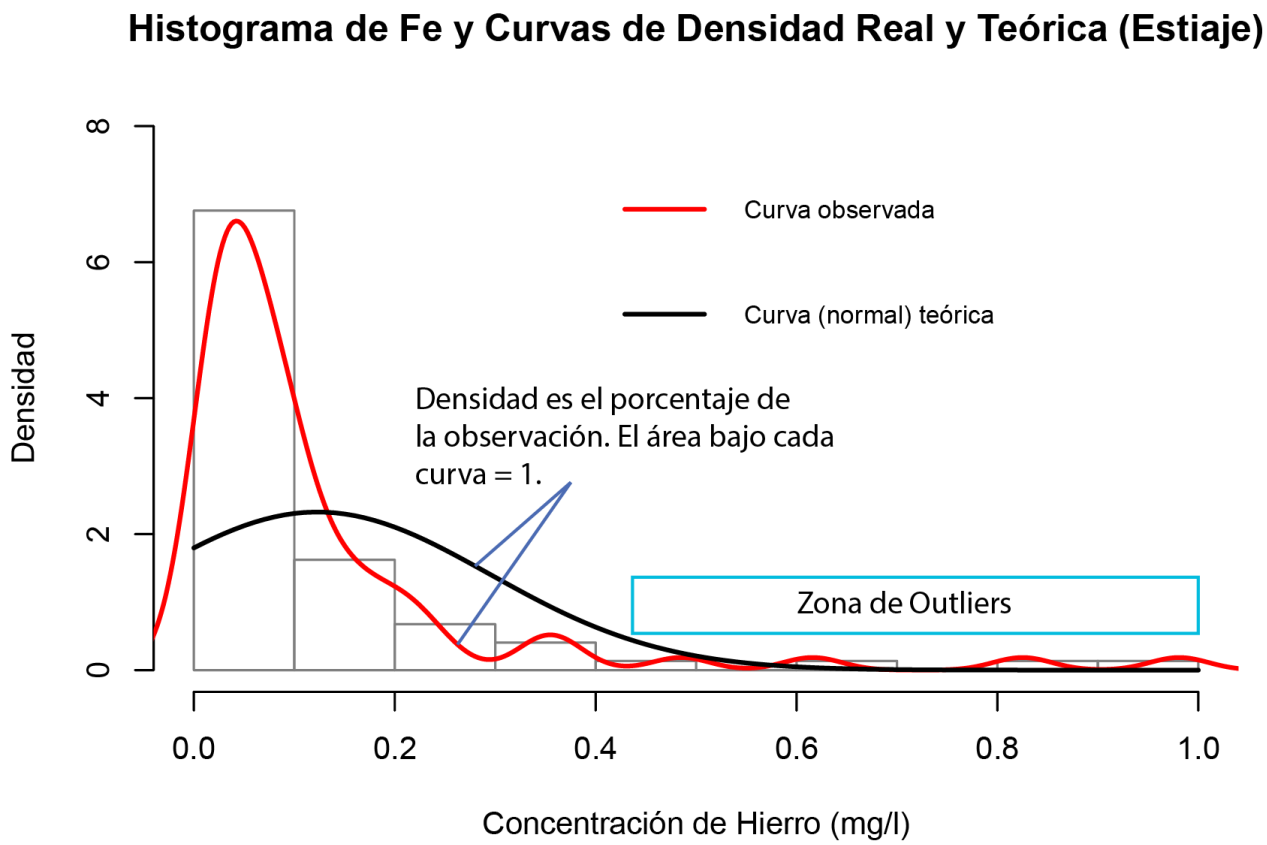
1. La familiar "curva de campana o Gausiana" de la distribución normal.



Analytics AoZ

PDF

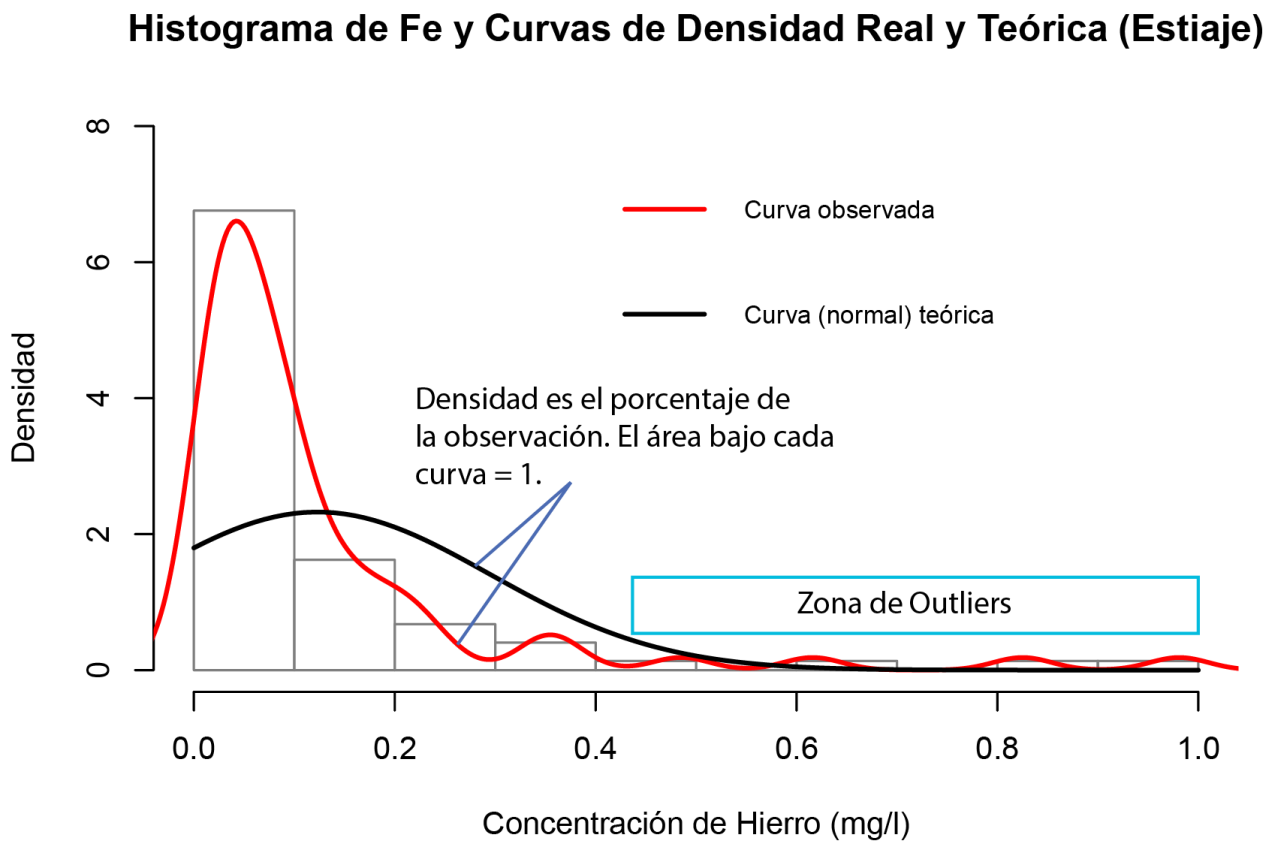
1. Una forma más realista para la data con no detectados son distribuciones asimétricas.



Analytics AoZ

PDF

1. Una forma más realista para la data con no detectados son distribuciones asimétricas.
2. Dos distribuciones simétricas comunes son Lognormal y Gamma.

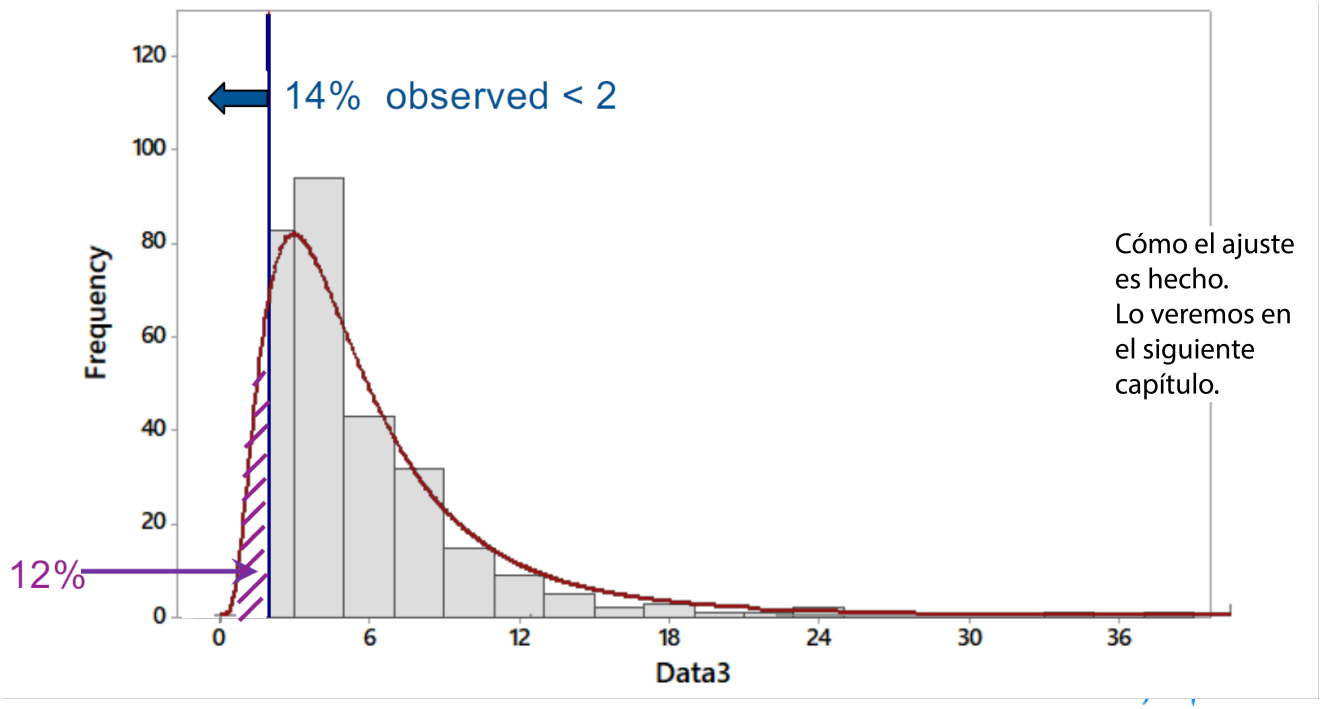


Analytics AoZ

PDF para datos censurados

Histograma: una barra no es dibujada para datos censurados. No deectados no son mostrados. No existen valores para el 14% de data inferior, solo conocemos que estos son <2.

1. El ajuste de la distribución debería tener un % dabajo de 2 similar al % en el set de datos.



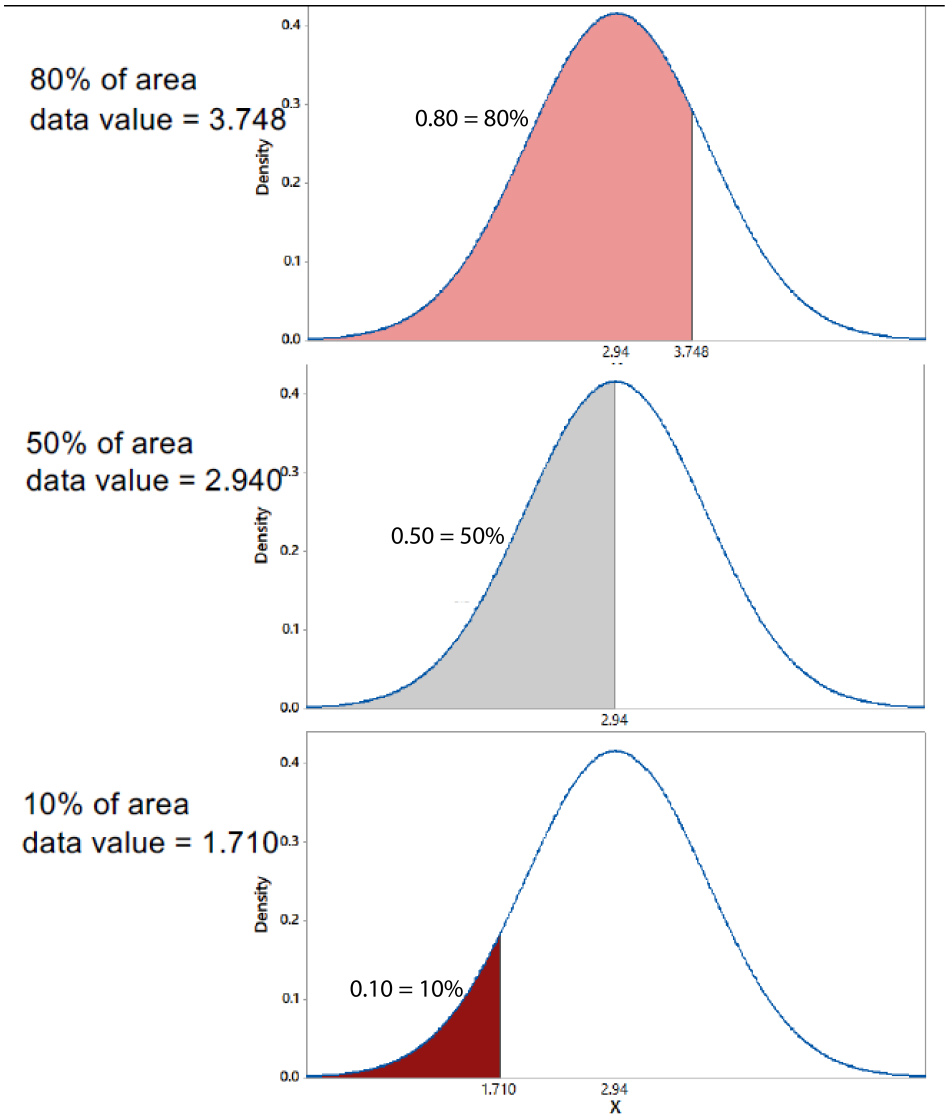
Analytics AoZ

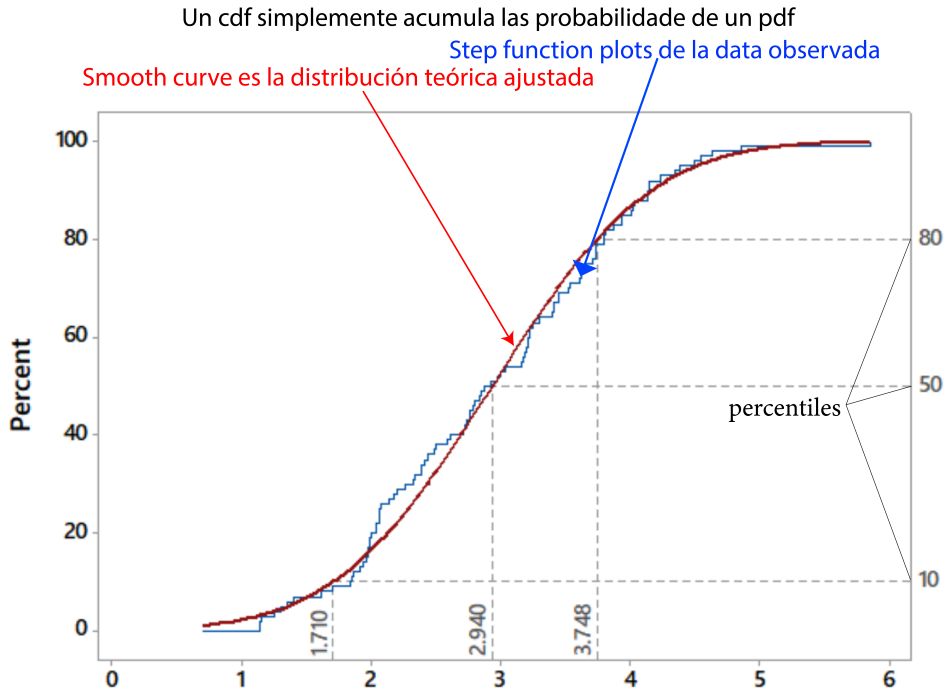
Section 5

Función de Probabilidad de Densidad Acumulada (cdf)

Analytics AoZ

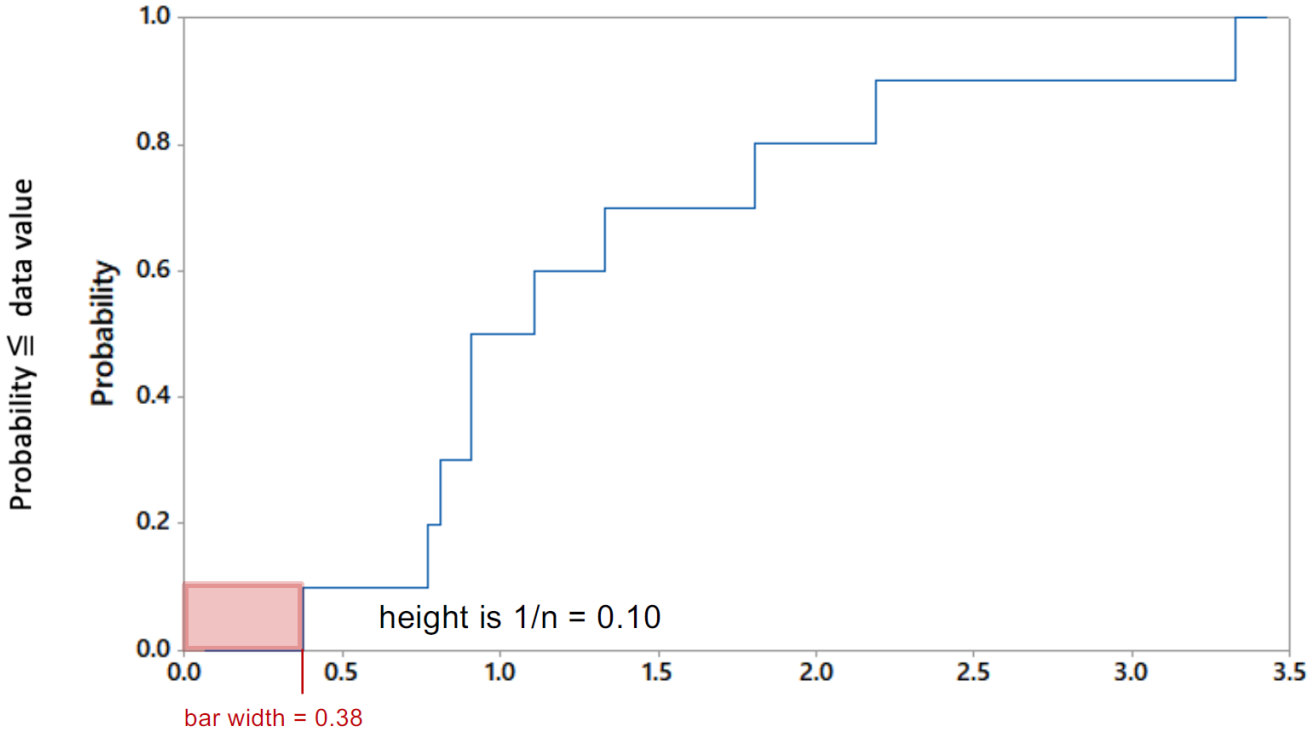
CDF: Cummulative Distribution Functions





CDF: Cummulative Distribution Functions

No NDs para comenzar, n = 10
3.33 2.19 1.81 1.33 1.11 0.91 0.91 0.81 0.77 0.38
Los **saltos** son 1/n

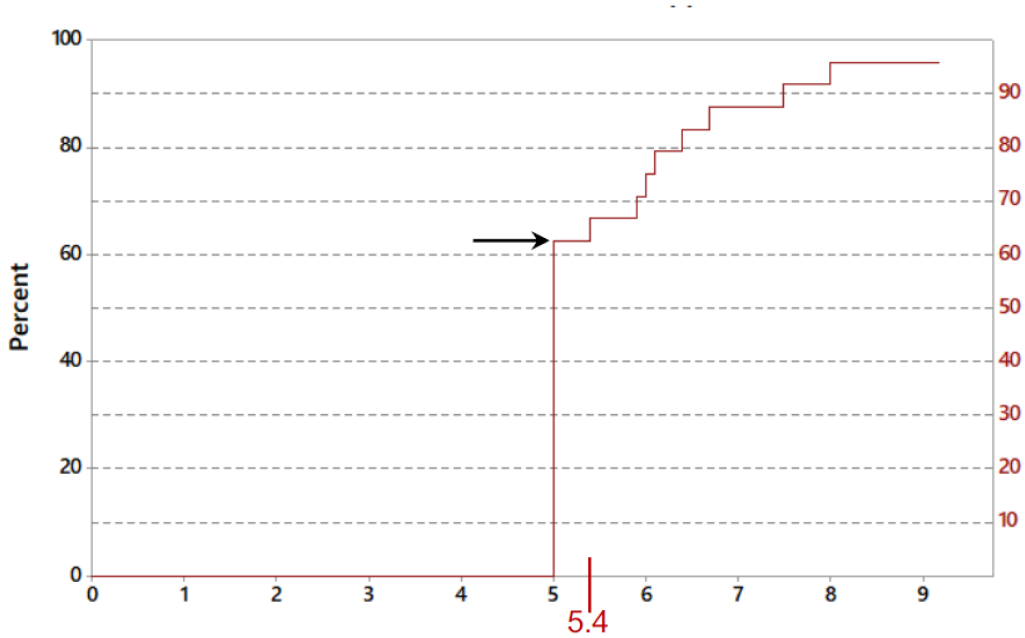


Analytics AoZ

CDF para datos censurados

Concentraciones Background Cu

```
> enparCensored(Copper.ppb, Censored)

Based on Type I Censored Data

----------------------------------------

Censoring Level(s):              5    (only 1 DL)

Estimated Parameter(s):          mean     = 5.6750000
                                 sd       = 1.1177544
                                 se.mean = 0.1457466

Estimation Method:               Kaplan-Meier

Sample Size:                     24

Percent Censored:                62.5%

Median:                          <5
```



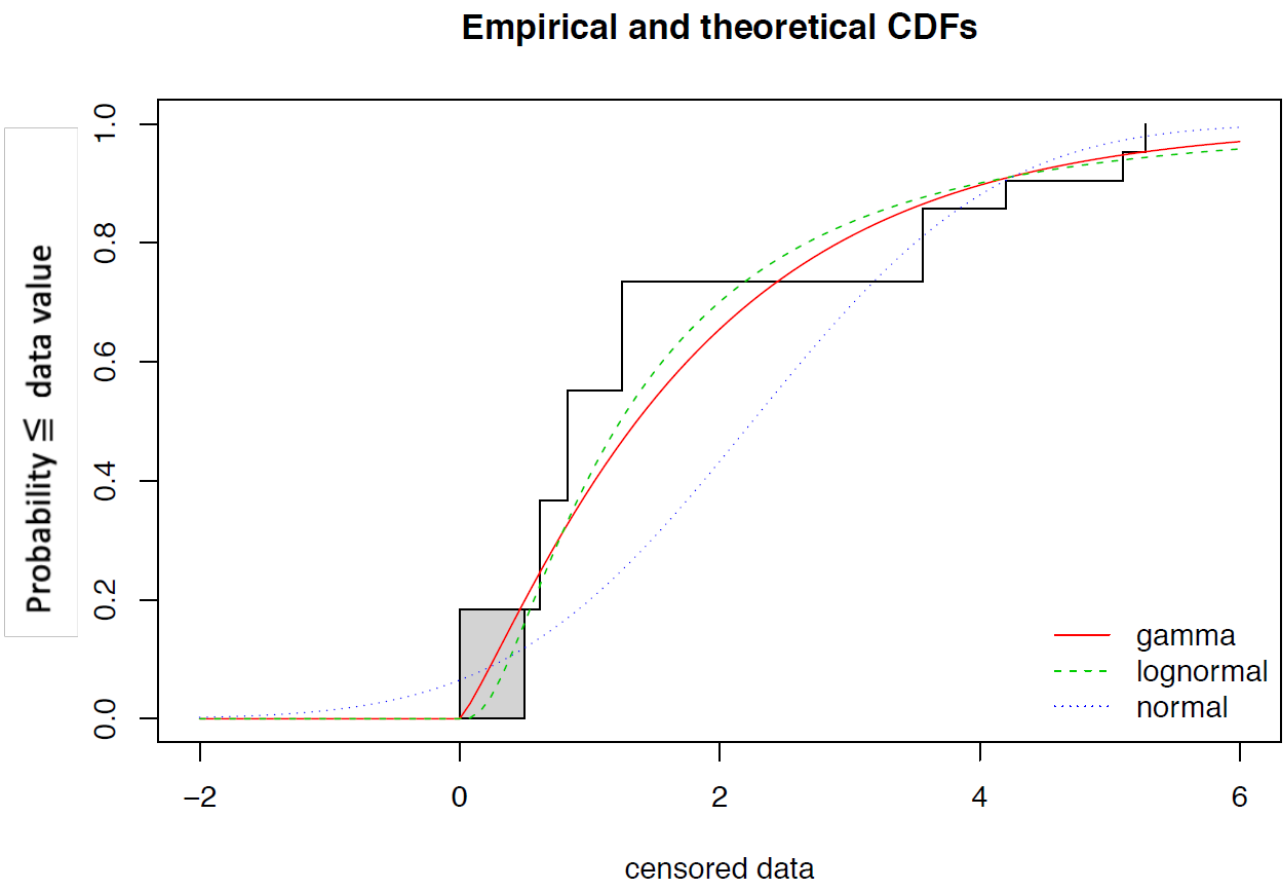
62.5% de la concentración de Cu son <5.

La primera observación detectada empieza en la flecha, con el valor de 5.4. Su alto = 1/n

Analytics AoZ

Mejor Ajuste cdf para datos censurados

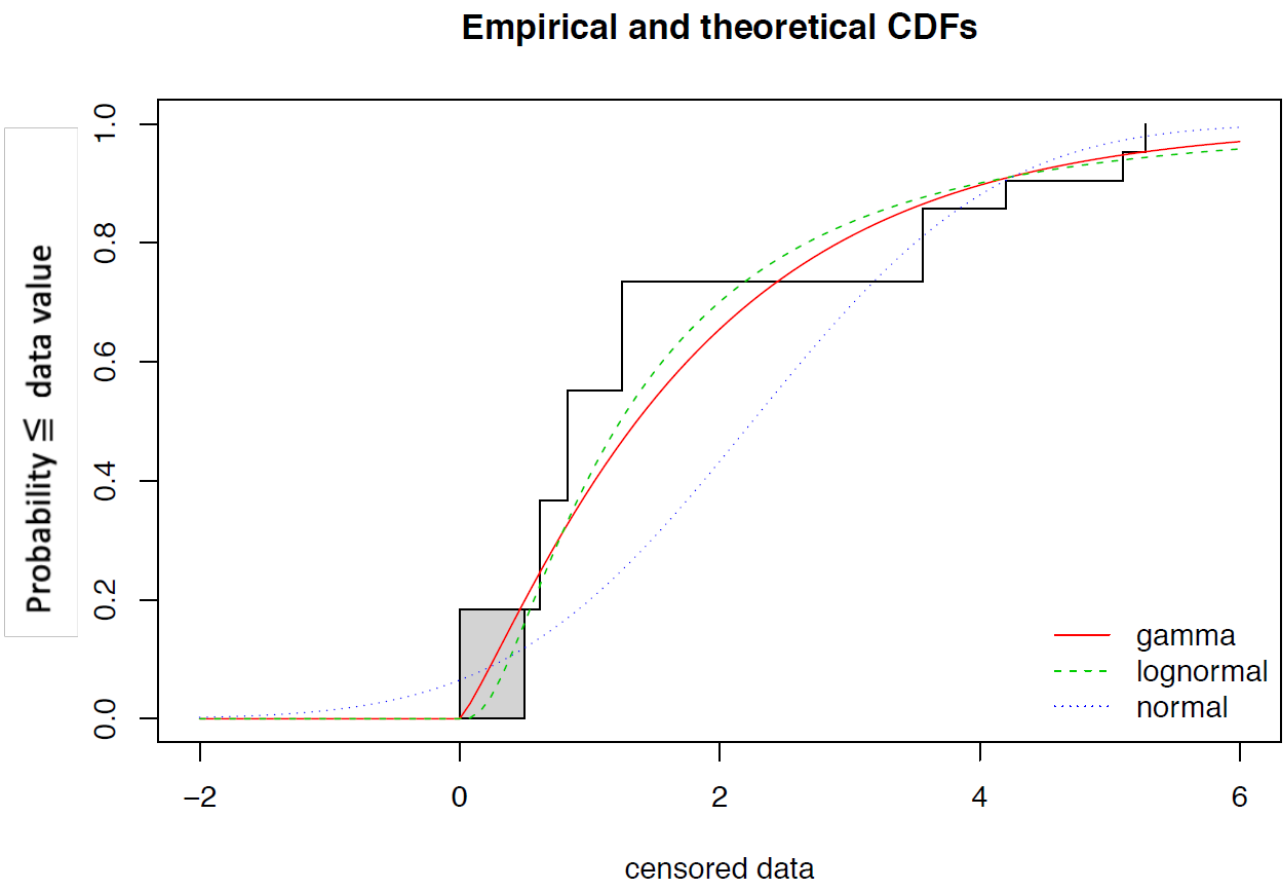
1. Data es mostrada como una step function. Caja color plomo esta debajo del menor Ld.



Analytics AoZ

Mejor Ajuste cdf para datos censurados

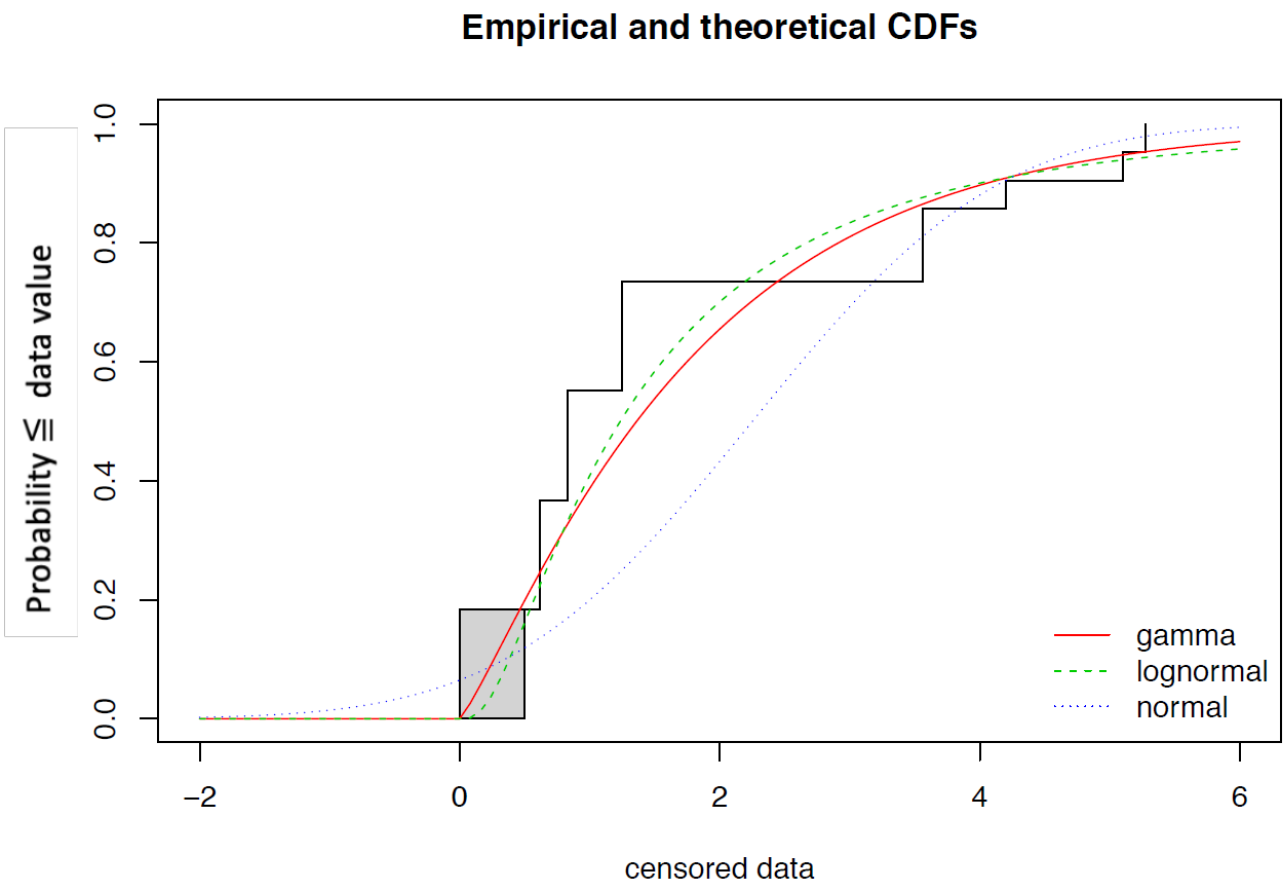
1. Data es mostrada como una step function. Caja color plomo esta debajo del menor Ld.
2. Gamma parece el mejor ajuste, 2nd lognormal.



Analytics AoZ

Mejor Ajuste cdf para datos censurados

1. Data es mostrada como una step function. Caja color plomo esta debajo del menor Ld.
2. Gamma parece el mejor ajuste, 2nd lognormal.
3. Nota: solo la distribución normal estima debajo de 0.



Analytics AoZ

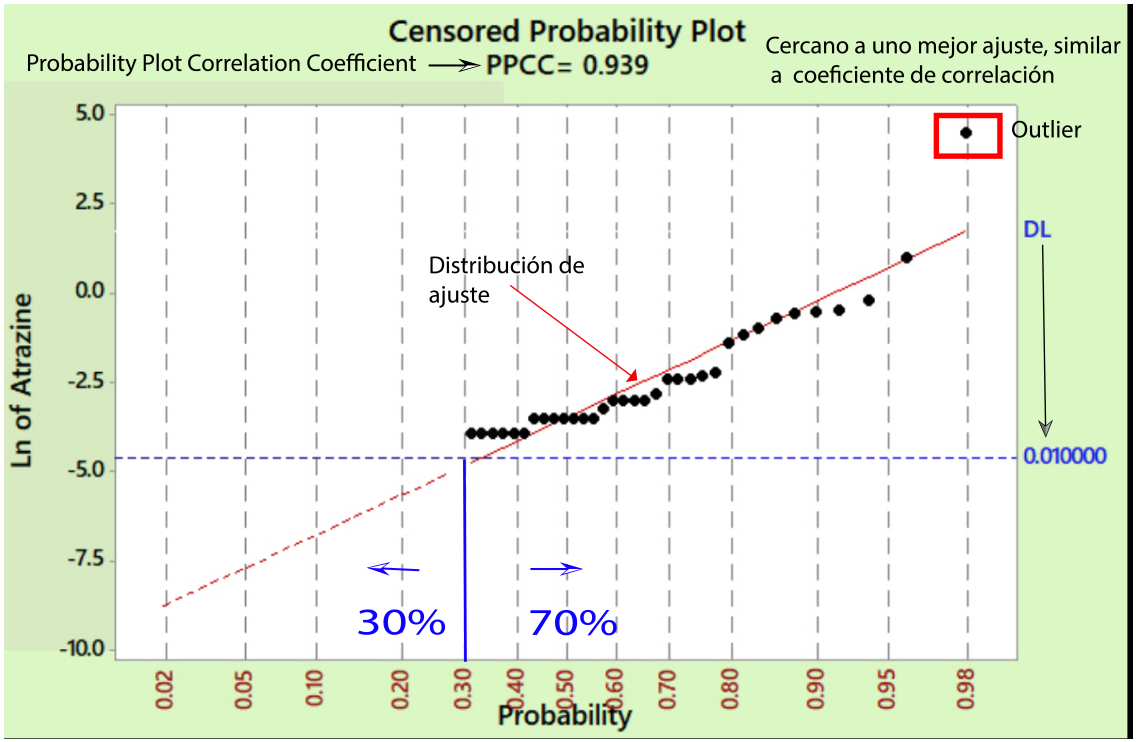
Section 6

Plots de Probabilidad (or Q-Q plots)

Analytics AoZ

Q-Q plots para data con NDs

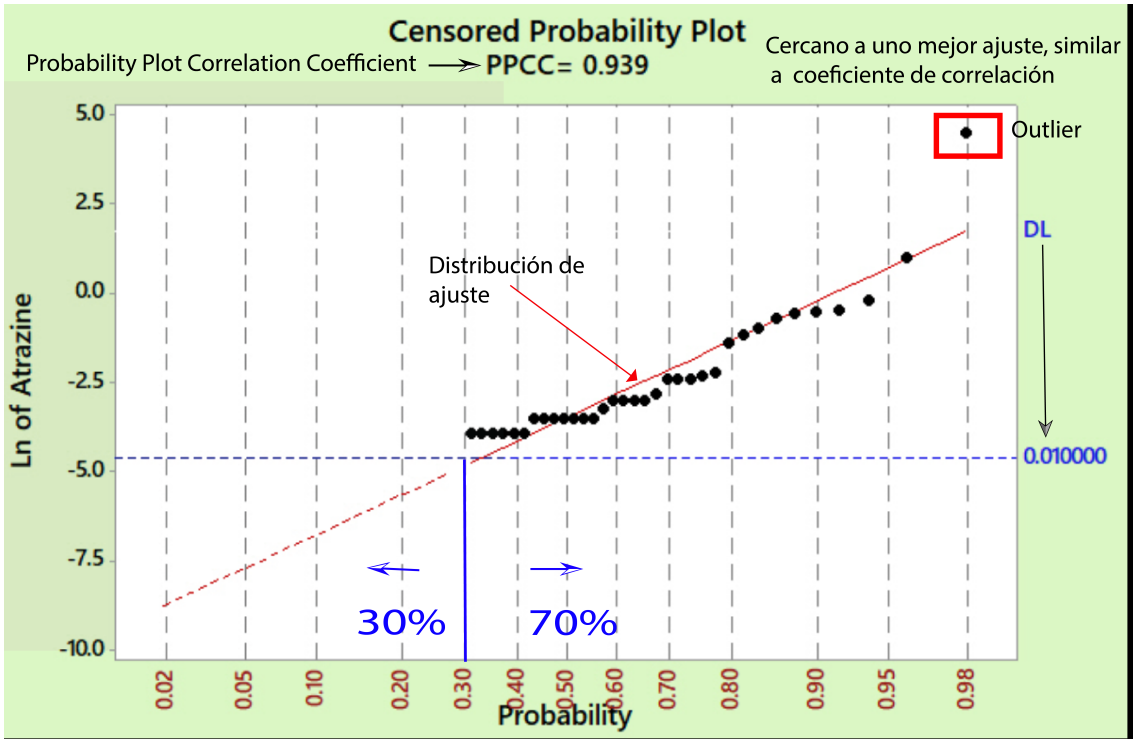
1. Plots de probabilidad <= valores de data para observaciones detectdas en el eje X.



Analytics AoZ

Q-Q plots para data con NDs

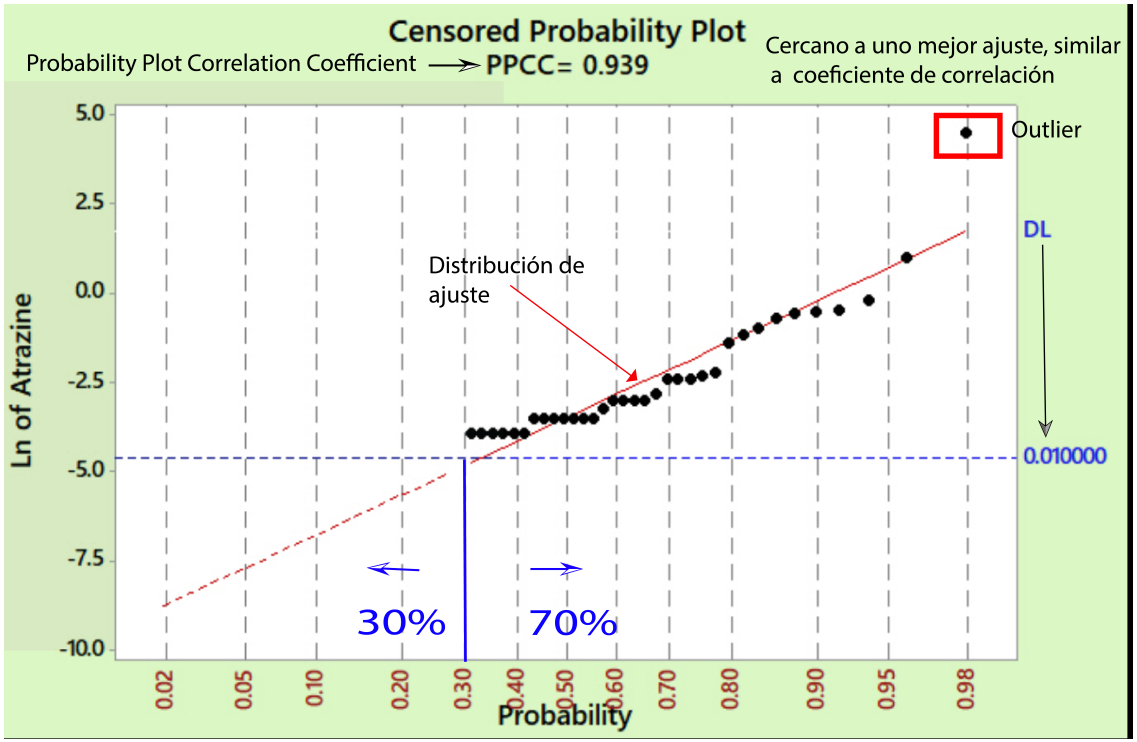
1. Plots de probabilidad <= valores de data para observaciones detectdas en el eje X.
2. No detectados no son ploteados, pero espacios a la izquierda de las observaciones detectadas en percentiles es correcto.



Analytics AoZ

Q-Q plots para data con NDs

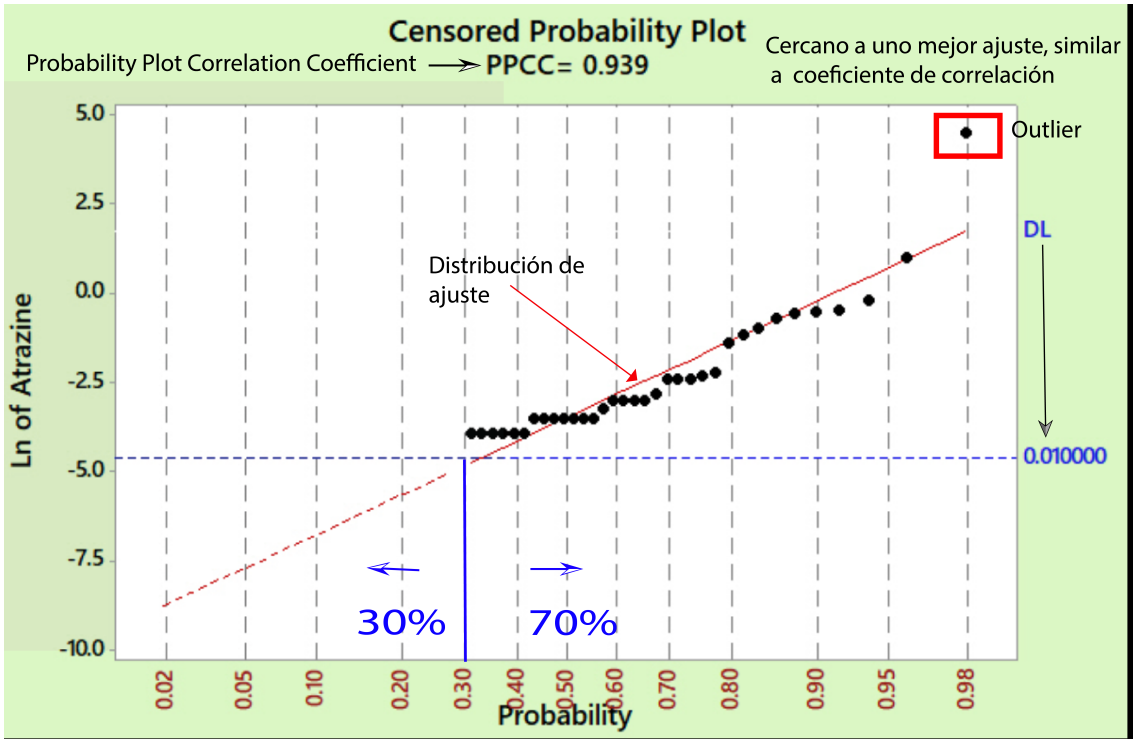
1. Plots de probabilidad <= valores de data para observaciones detectdas en el eje X.
2. No detectados no son ploteados, pero espacios a la izquierda de las observaciones detectadas en percentiles es correcto.
3. Una línea continua representa la disribución tales como la normal, lognormal o gamma.



Analytics AoZ

Q-Q plots para data con NDs

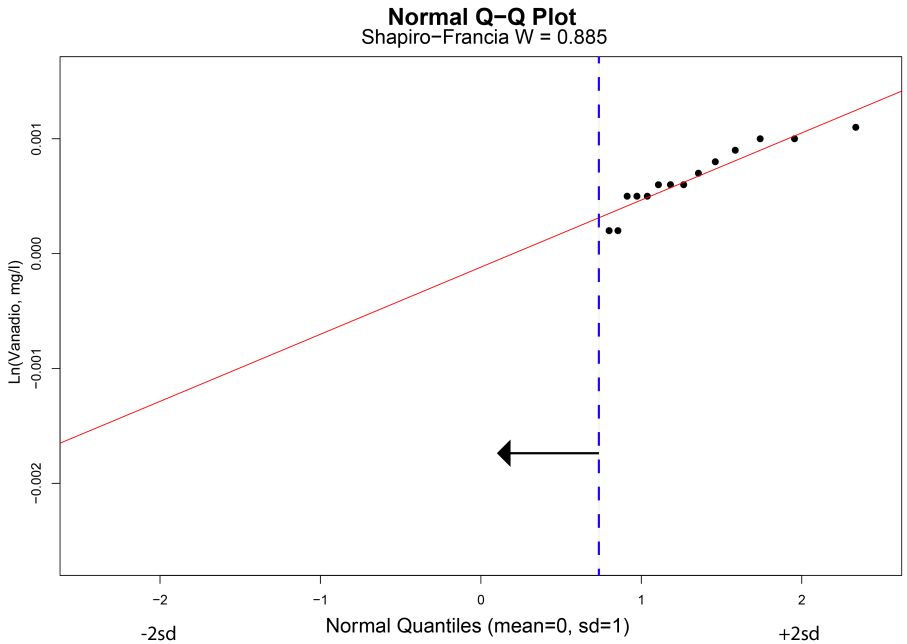
1. Plots de probabilidad <= valores de data para observaciones detectdas en el eje X.
2. No detectados no son ploteados, pero espacios a la izquierda de las observaciones detectadas en percentiles es correcto.
3. Una línea continua representa la disribución tales como la normal, lognormal o gamma.
4. PPCC mide el ajuste. Max PPCC=1. Escoger la distribución con mayor PPCC.



Analytics AoZ

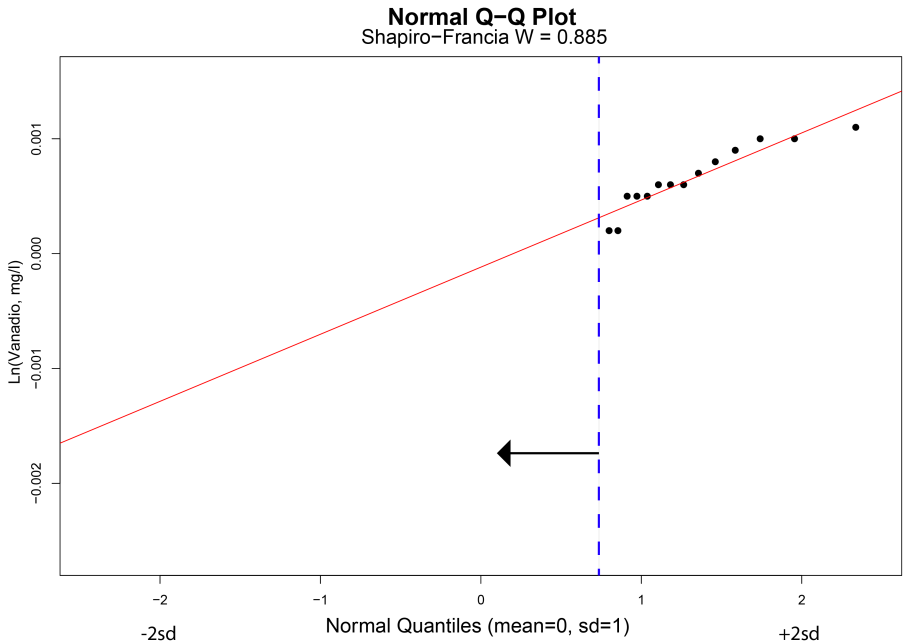
Q-Q plots para data con No Detectados

1. En vez de las probabilidades no lineales $<=$ valor dato, software usualmente plotea una linea escalar. Una es quantiles normales.



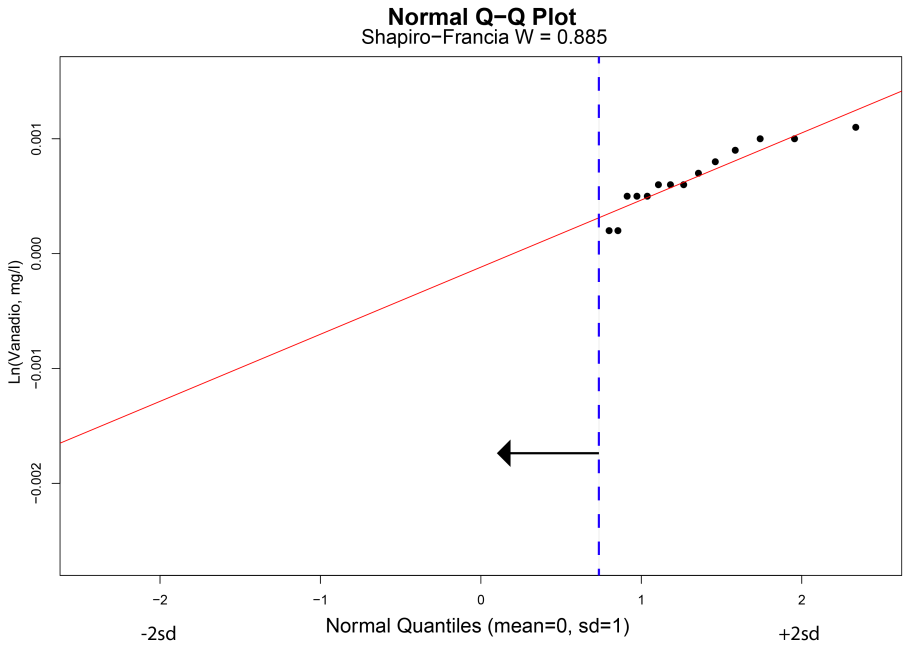
Q-Q plots para data con No Detectados

1. En vez de las probabilidades no lineales $<=$ valor dato, software usualmente plotea una linea escalar. Una es quantiles normales.
2. Quantiles normales son quantiles de la distribución normal con media=0 y sd=1.



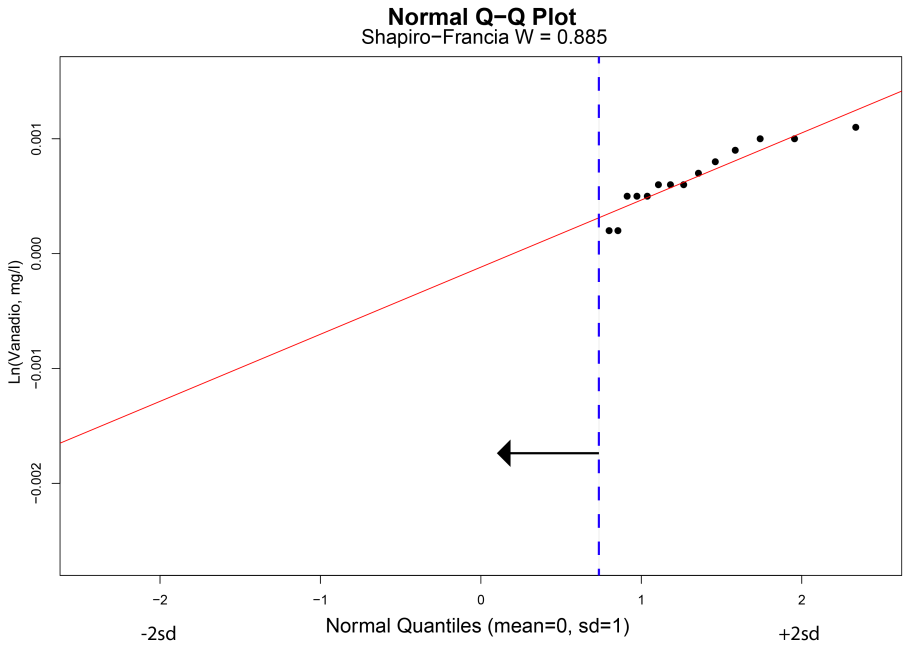
Q-Q plots para data con No Detectados

1. En vez de las probabilidades no lineales <= valor dato, software usualmente plotea una linea escalar. Una es quantiles normales.
2. Quantiles normales son quantiles de la distribución normal con media=0 y sd=1.
3. No detectados influencian en los cuantiles de los detectados.Aquí los menores valores detectados son justo menor que +1, cual esta a 78% del dataset.



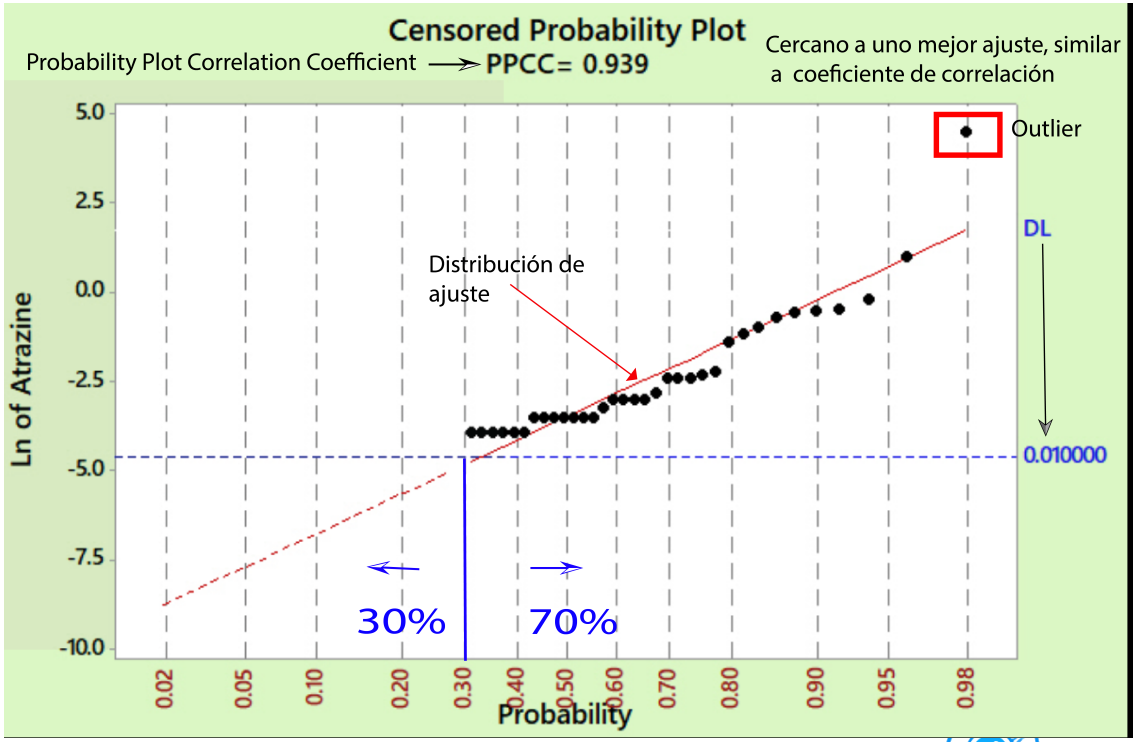
Q-Q plots para data con No Detectados

1. En vez de las probabilidades no lineales <= valor dato, software usualmente plotea una linea escalar. Una es quantiles normales.
2. Quantiles normales son quantiles de la distribución normal con media=0 y sd=1.
3. No detectados influencian en los cuantiles de los detectados.Aquí los menores valores detectados son justo menor que +1, cual esta a 78% del dataset.
4. Aquíe el 78% del área de datos < menor límite de detección. Múltiplos Ld pueden ser incorporados.



Q-Q plots para ajustar distribución de datos con ND

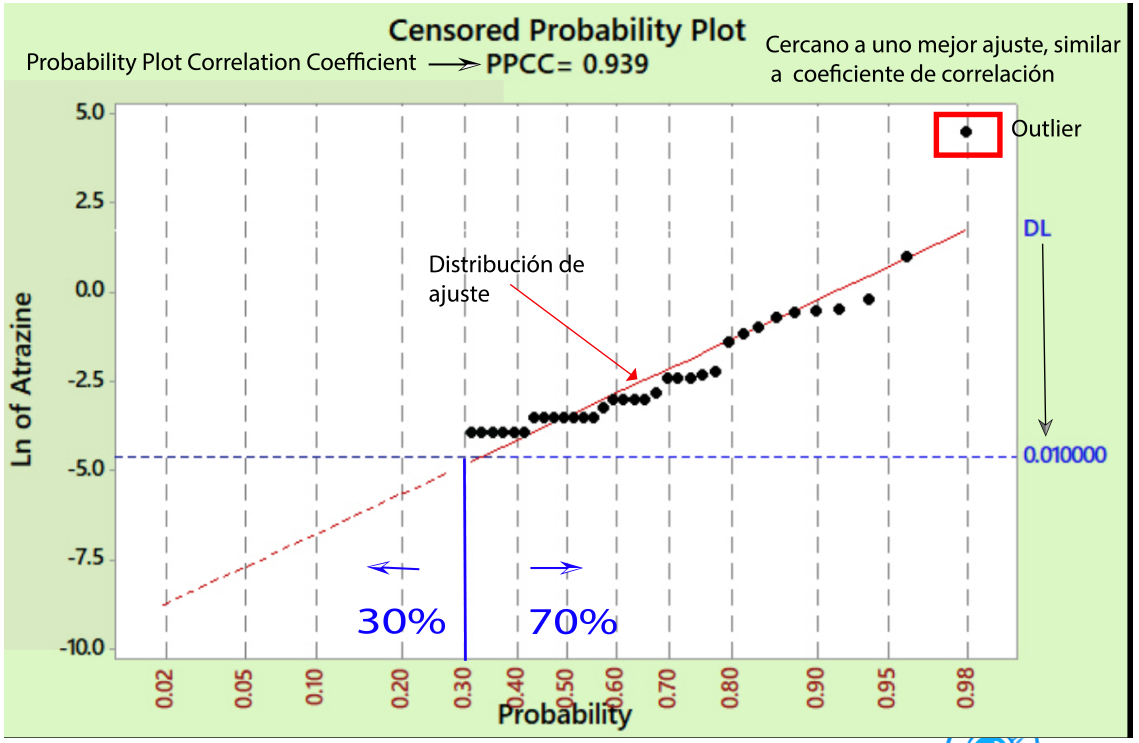
1. Esto es ploteado correctamente.**30%** de no detectados no son mostrado como puntos, pero el espacio es reservado en la parte inferior.



Analytics AoZ

Q-Q plots para ajustar distribución de datos con ND

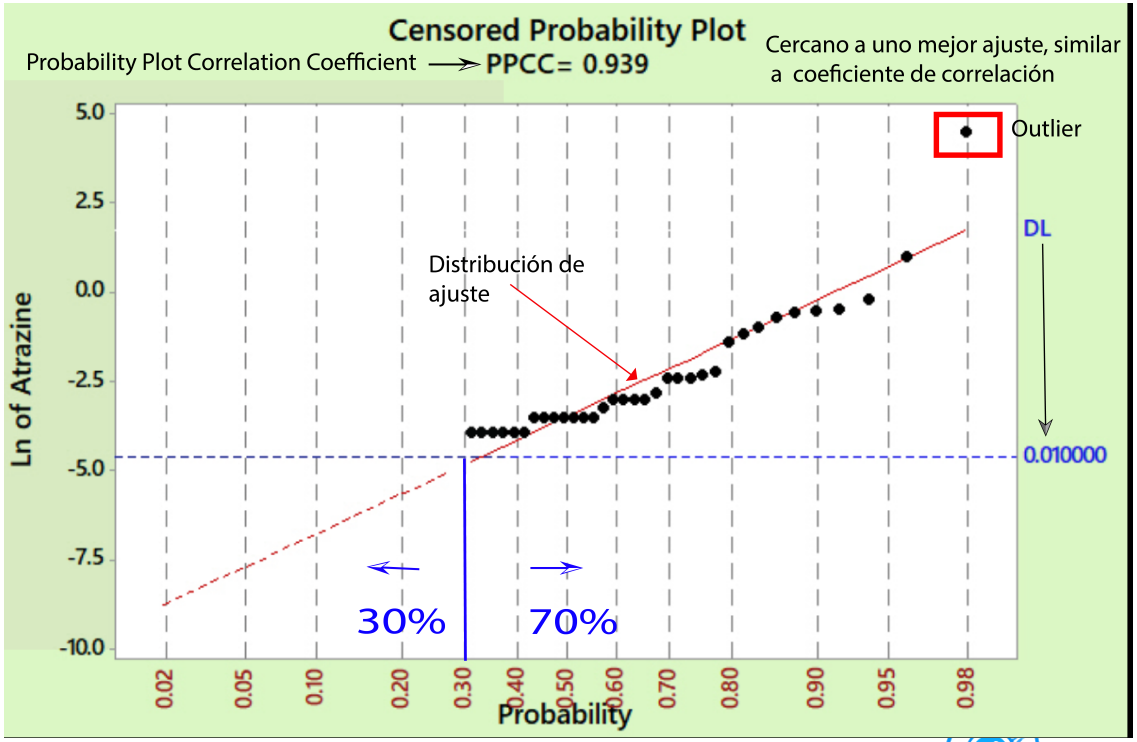
1. Esto es ploteado correctamente.**30%** de no detectados no son mostrado como puntos, pero el espacio es reservado en la parte inferior.
2. Encontraremos la rutina para hacer esto en "Survivial Analysis" o "Censored data" de los software estadísticos.



Analytics AoZ

Q-Q plots para ajustar distribución de datos con ND

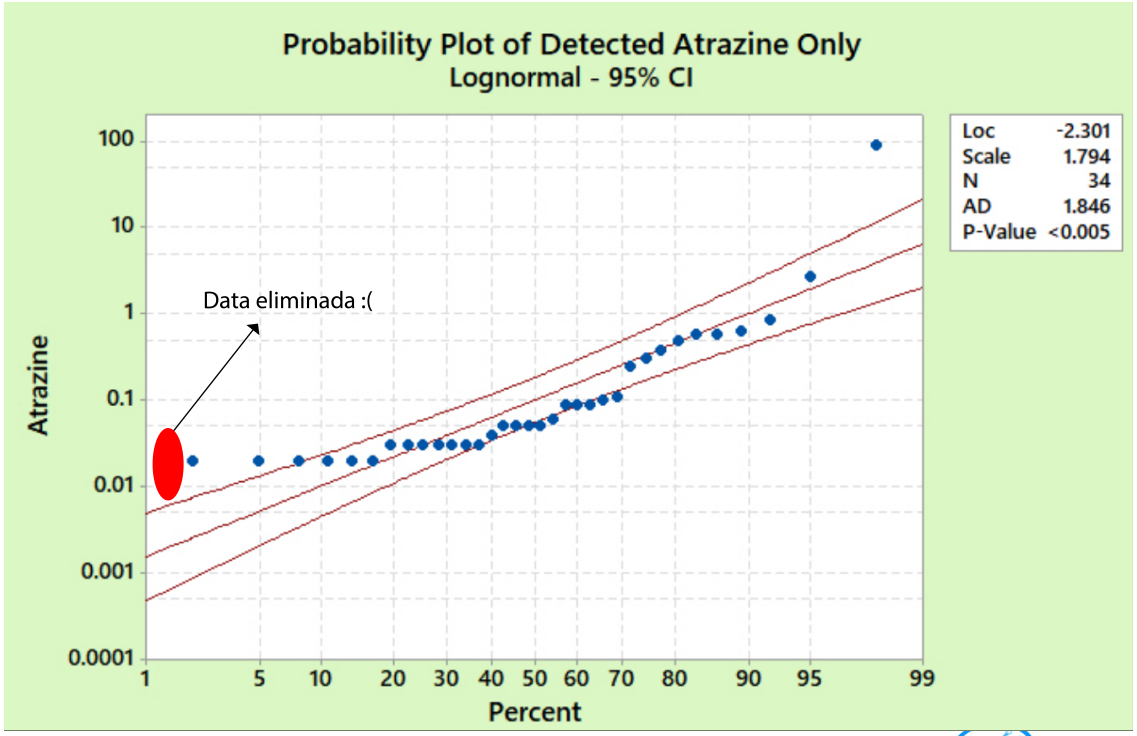
1. Esto es ploteado correctamente.**30%** de no detectados no son mostrado como puntos, pero el espacio es reservado en la parte inferior.
2. Encontraremos la rutina para hacer esto en "Survivial Analysis" o "Censored data" de los software estadísticos.
3. El comando **qqPlotCensored()** en el paquete *EnvStats* es uno de estos.



Analytics AoZ

Q-Q plots data con NDs es eliminada incorrectamente

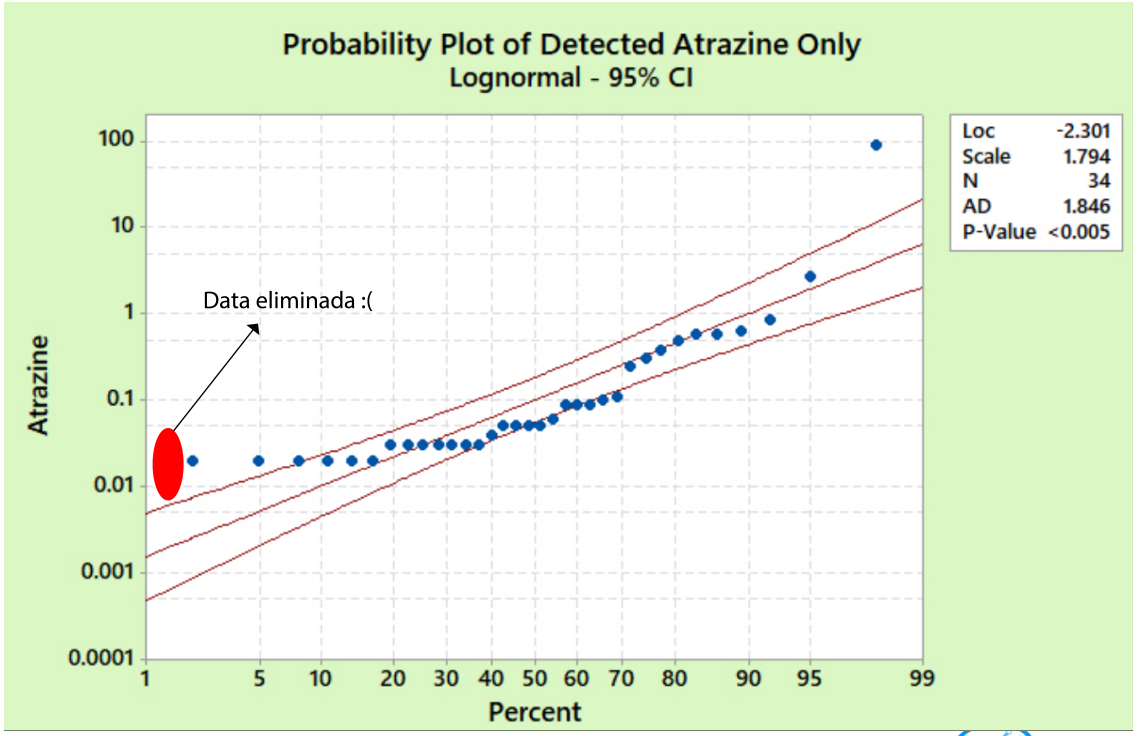
1. No ploteada correctamente. Usar Q-Q plots estándar que no están diseñado para data con no detectados.



Analytics AoZ

Q-Q plots data con NDs es eliminada incorrectamente

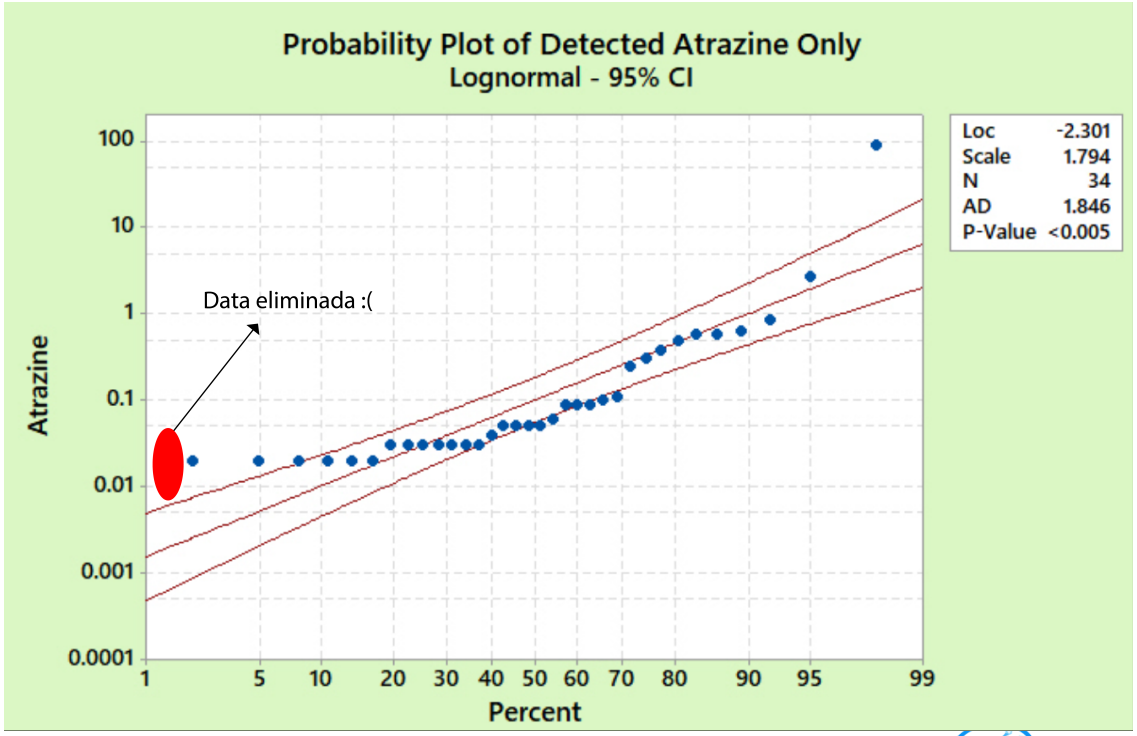
1. No ploteada correctamente. Usar Q-Q plots estándar que no están diseñado para data con no detectados.
2. Eliminar no detectados, así que **todos los percentiles son empujados al inferior** (*desviados hacia la izquierda*).



Analytics AoZ

Q-Q plots data con NDs es eliminada incorrectamente

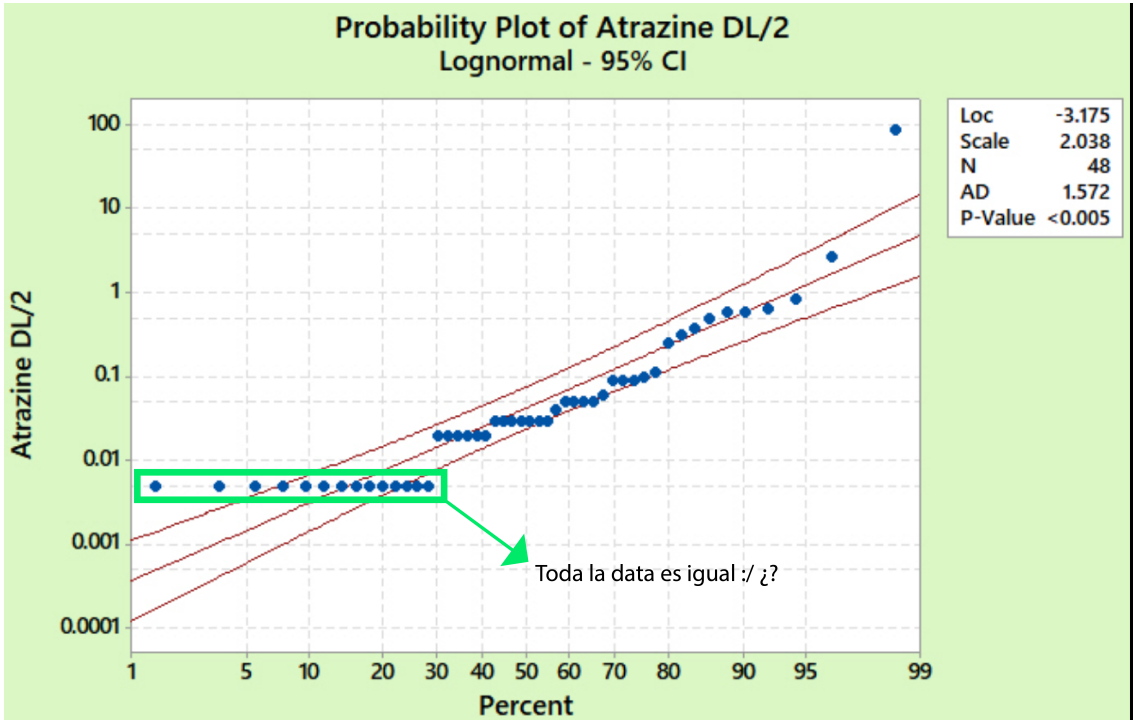
1. No ploteada correctamente. Usar Q-Q plots estándar que no están diseñado para data con no detectados.
2. Eliminar no detectados, así que **todos los percentiles son empujados al inferior** (*desviados hacia la izquierda*).
3. Desajustes de la distribución comparada con la verdadera forma de la data.



Analytics AoZ

Q-Q plots con 1/2 Ld subtituidos por NDs

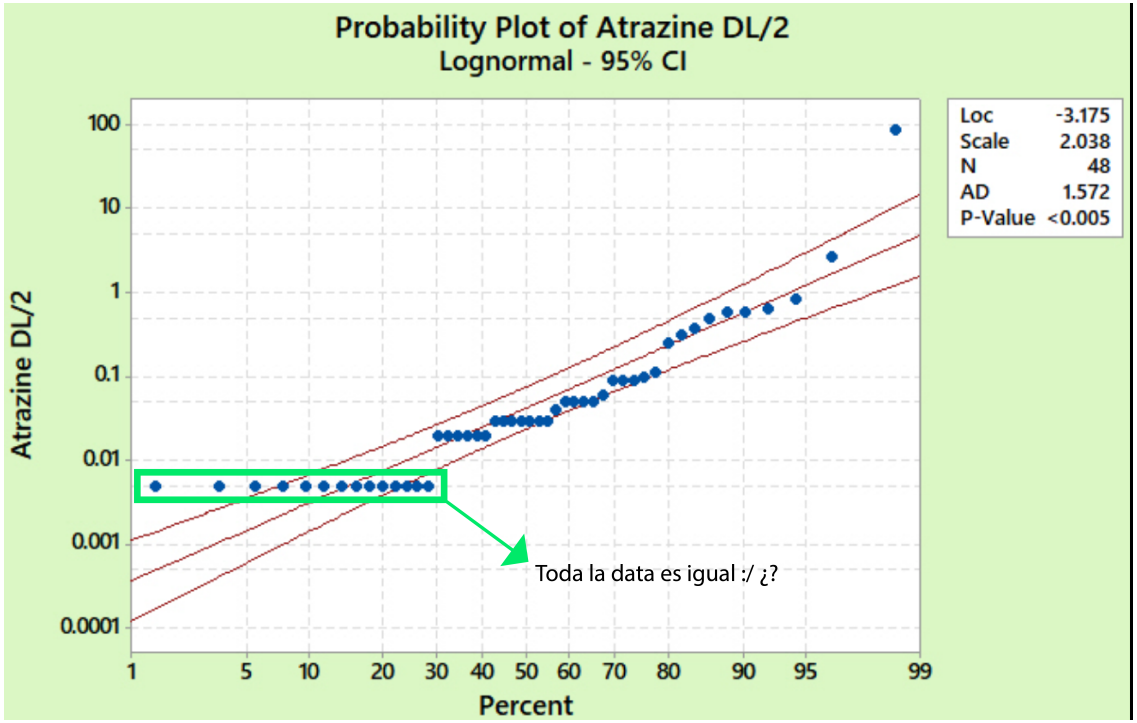
1. Substituir los valores por una linea recta en la parte inferior.



Analytics AoZ

Q-Q plots con 1/2 Ld subtituidos por NDs

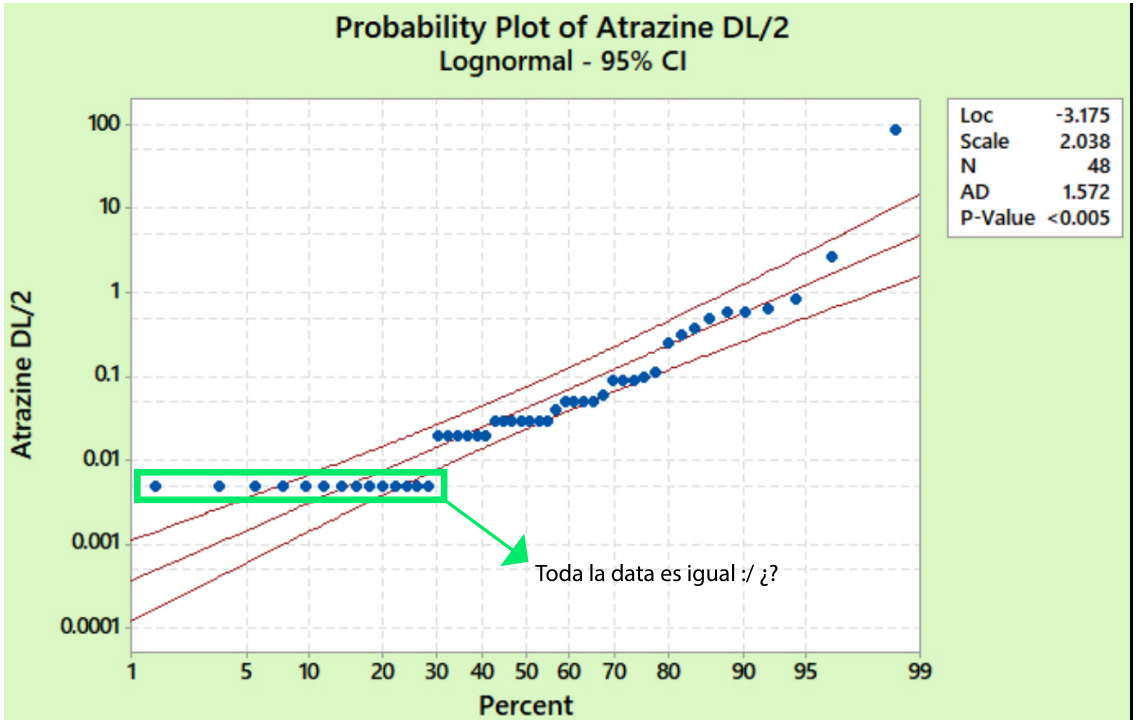
1. Substituir los valores por una linea recta en la parte inferior.
2. Distorsionar la distribución en su parte baja comparada con al verdadera forma de la data.



Analytics AoZ

Q-Q plots con 1/2 Ld subtituidos por NDs

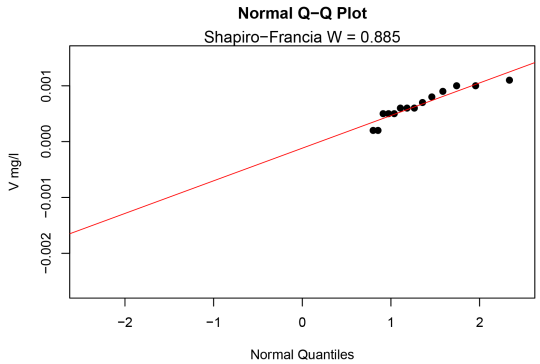
1. Substituir los valores por una linea recta en la parte inferior.
2. Distorsionar la distribución en su parte baja comparada con al verdadera forma de la data.
3. Tendencia a escoger la distribución incorrecta; malos estimados para los percentiles en la parte baja.

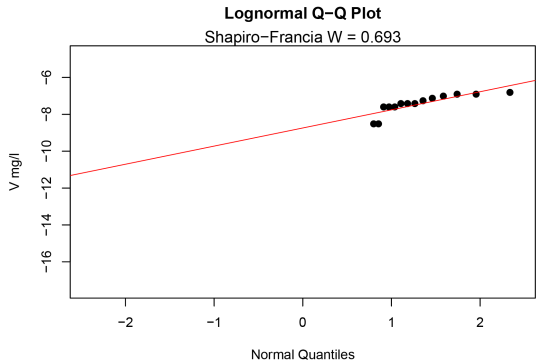


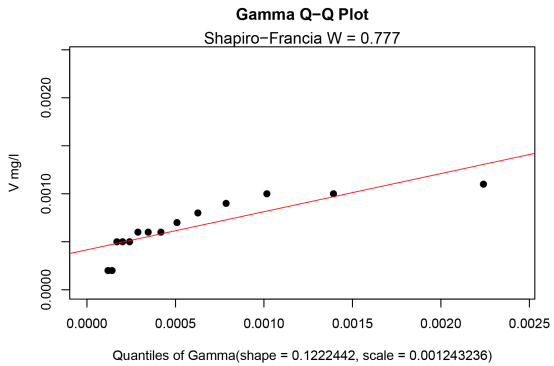
Analytics AoZ

Q-Q plots de posible ajuste distribución para data con NDs

1. Distribución con una data cerrada a una línea recta, o con el más alto BIC (Bayesian Information Criterion) o Shapiro-Francia (coeficiente de correlación), es el mejor ajuste.





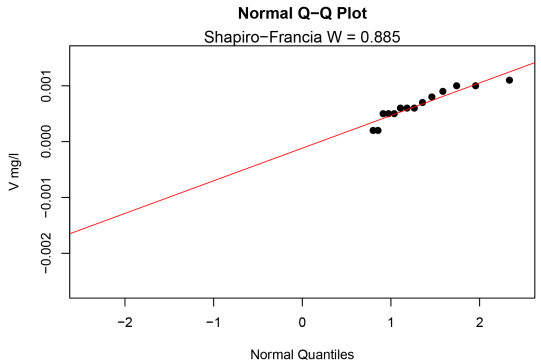


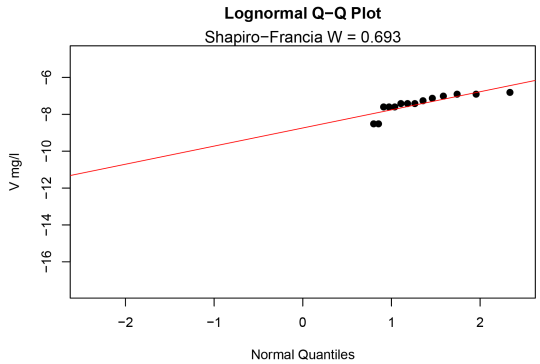
Normal is a good fit

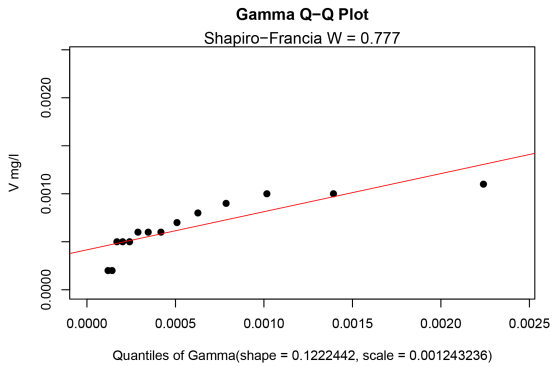
Analytics AoZ

Q-Q plots de posible ajuste distribución para data con NDs

1. Distribución con una data cerrada a una línea recta, o con el más alto BIC (Bayesian Information Criterion) o Shapiro-Francia (coeficiente de correlación), es el mejor ajuste.
2. Aquí la distribución normal es el mejor ajuste comparada con las tres otrs distribuciones.







Normal is a good fit

Analytics AoZ

Section 7

Ajustando Distribuciones en R

Analytics AoZ

Ajuste de Distribución

```
bd1 <- read.csv(file="../ParteIA/Code/Ejemplos/Example1.txt",
attach(bd1)
NADA::censummary(Arsenic, NDisTRUE)
```

```
FALSE all:
FALSE         n    n.cen  pct.cen      min      max
FALSE 21.00000 14.00000 66.66667  0.50000  5.27628
FALSE
FALSE limits:
FALSE   limit  n uncen   pexceed
FALSE 1   0.5  1     3 0.8163265
FALSE 2   2.0  1     0 0.2653061
FALSE 3   3.0  1     1 0.2653061
FALSE 4   4.0 11     3 0.1428571
```

Analytics AoZ

```
# 21 obs. Pequeño para decidir que distribución usar!
```

Calcular PPCC o BIC para mejor distribución.

```
library(EnvStats)
gofTestCensored(Arsenic, NDisTRUE, dist="gamma", test="ppcc")
gofTestCensored(Arsenic, NDisTRUE, dist="lnorm", test="ppcc")
gofTestCensored(Arsenic, NDisTRUE, dist="norm", test="ppcc")
```

```
# gamma - PPCC = r = 0.969  # Mejor Ajuste
# lnorm - PPCC = r = 0.966
# norm  - PPCC = r = 1.968
```

1. Mejor: Maximizar el PPCC, minimizar el BIC para obtener la mejor distribución.

Analytics AoZ

Calcular PPCC o BIC para mejor distribución.

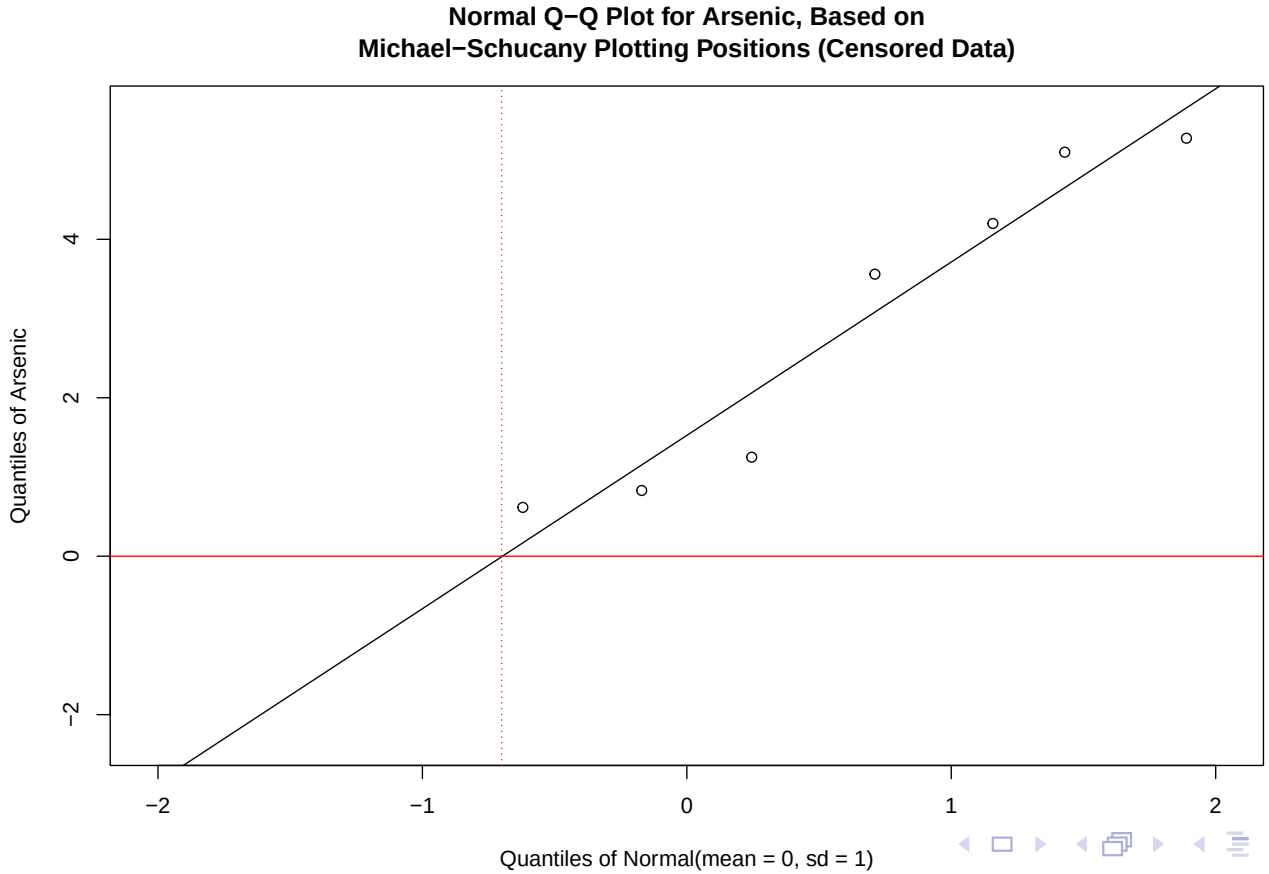
```
library(EnvStats)
gofTestCensored(Arsenic, NDisTRUE, dist="gamma", test="ppcc")
gofTestCensored(Arsenic, NDisTRUE, dist="lnorm", test="ppcc")
gofTestCensored(Arsenic, NDisTRUE, dist="norm", test="ppcc")
```

```
# gamma - PPCC = r = 0.969  # Mejor Ajuste
# lnorm - PPCC = r = 0.966
# norm  - PPCC = r = 1.968
```

1. Mejor: Maximizar el PPCC, minimizar el BIC para obtener la mejor distribución.
2. Mayor PPCC es **0.969** es la distribución gamma. Casi el mismo valor que la distribución normal, podría ser usada la normal? **No! (Recordar como esto esta fuera del plot CDF?)**.

Si la distribución normal es escogida (No!)

```
qqPlotCensored(Arsenic,NDisTRUE,dist="norm",add.line=TRUE)
abline(h=0, col="red")
abline(v=-0.70, col="red", lty=3)
```



Si la distribución normal es escogida (No!) 2

1. Distribución normal produce aproximadamente 15% de números negativos.

Analytics AoZ

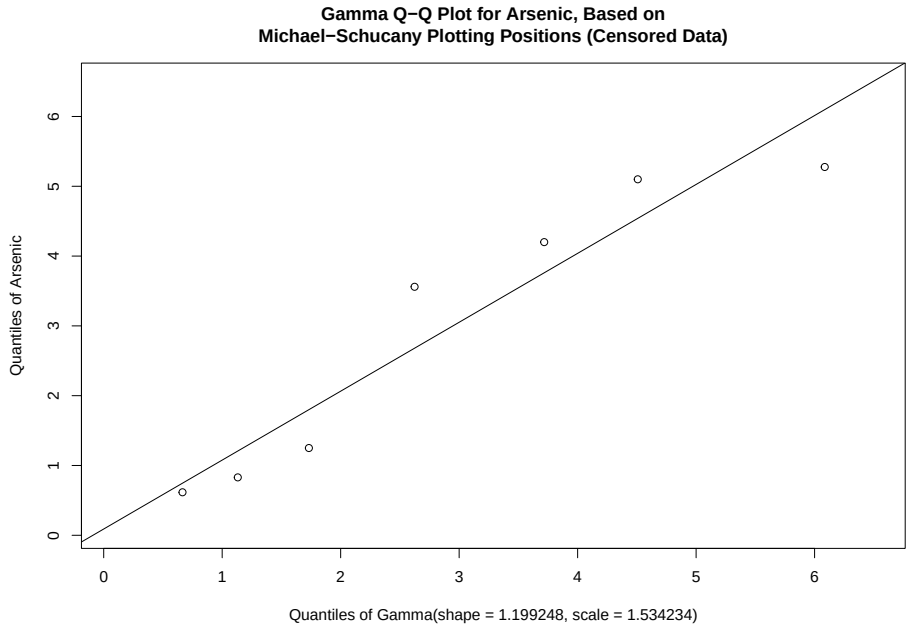
Si la distribución normal es escogida (No!) 2

1. Distribución normal produce aproximadamente 15% de números negativos.
2. Inaceptable! Rechazar incluso si tiene mayor PPCC. Las estimaciones de la media y UCL serán incorrectas.

Analytics AoZ

Usar el siguiente mayor PPCC: distribución gamma

```
qqPlotCensored(Arsenic,NDisTRUE,dist="gamma",add.line=TRUE,
               estimate.params = TRUE)
```

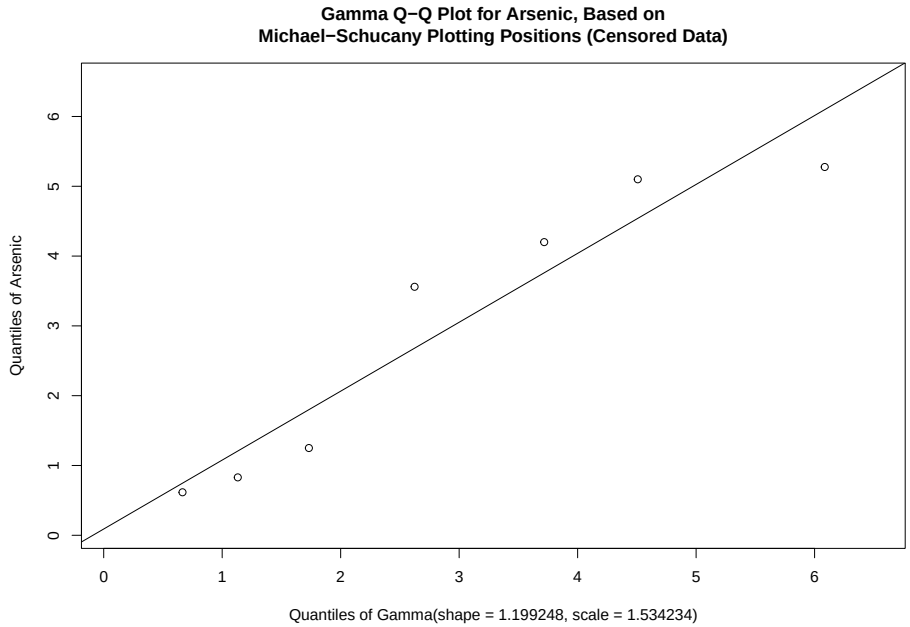


1. La distribución gamm tiene el *mayor PPCC = 0.969*

Analytics AoZ

Usar el siguiente mayor PPCC: distribución gamma

```
qqPlotCensored(Arsenic,NDisTRUE,dist="gamma",add.line=TRUE,
               estimate.params = TRUE)
```



1. La distribución gamm tiene el *mayor PPCC = 0.969*
2. BIC de las 3 distribuciones usar el paquete *fitdist()* escoger el menor es mejor. *gamma = 43.9, lognormal = 44.6 & normal = 50.6*.

Sumario

1. Los mejores métodos son los que usan probabilidades y cuantiles.

Analytics AoZ

Sumario

1. Los mejores métodos son los que usan probabilidades y cuantiles.
2. Esto es debido a que no detectados contienen en probabilidades de ser $<$Ld.

Analytics AoZ

Sumario

1. Los mejores métodos son los que usan probabilidades y cuantiles.
2. Esto es debido a que no detectados contienen en probabilidades de ser $<$Ld.
3. Boxplots, plots de probabilidad, PDFs y CDFs todos proveen información valiosa.

Analytics AoZ

Sumario

1. Los mejores métodos son los que usan probabilidades y cuantiles.
2. Esto es debido a que no detectados contienen en probabilidades de ser $<$Ld.
3. Boxplots, plots de probabilidad, PDFs y CDFs todos proveen información valiosa.
4. Scatterplots puede ser usados para plotear no detectados como lineas punteadas o barras de intervalos en vez de como puntos.

Analytics AoZ