Análisis Exploratorio de los Datos

Análisis Exploratorio de los Datos (EDA)

- El estudio ha sido diseñado y los datos han sido colectados, y ahora se tiene un montón de datos para analizar, presentar y publicar.

- El análisis exploratorio de los datos [en inglés EDA (Exploratory Data Analysis)] es una colección de técnicas para desplegar los datos con el cometido de que los datos “hablen por si mismos” previo a un análisis formal de éstos.
- EDA es una mezcla de técnicas viejas y nuevas implementado de manera formal por John Tukey (1977).

Análisis Exploratorio de los Datos (EDA)

- Antes de comenzar un análisis estadístico es necesario examinar los datos por las siguientes razones:
  - Para asegurarse de que los datos están aportando algo importante
  - Para detectar errores en la entrada de datos
  - Para detectar patrones en los datos que pueden no ser detectados por el análisis estadístico que se va a usar
  - Para asegurarse de que se cumplan los supuestos de los análisis estadísticos que se realizarán
  - Para interpretar desviaciones de los supuestos
  - Para detectar valores inusuales (outliers)

Estadísticos Básicos

- Un primer paso puede ser ver un reporte de los estadísticos básicos.

Ejemplo

Largo de ala para una especie de ave en 2 localidades A y B

Output de R

Output Window

	A	B
Mean	24.117	25.235
CI	+/-0.875	+/-1.045
TrimmedMean	24.054	25.531
Variance	9.97	19.884
SD	3.158	4.459
SEM	0.447	0.533
CV	0.413	0.788
Min	16.981	14.761
1st Qu.	22.302	22.781
Median	23.626	25.604
3rd Qu.	25.926	28.781
Max	32.916	34.196
QRange	3.62414	5.99971
Sum	1205.874	1766.453
Length	50	70
Missing	0	0
Zeros	0	0

Gráficos en EDA

- Uno de las técnicas más útiles en EDA son los gráficos.
- Los gráficos tienen 2 funciones en el contexto del análisis de los datos:

  - Los gráficos son una herramienta usada para explorar patrones de los datos previo a un análisis estadístico formal
  - Los gráficos son usados para comunicar una gran cantidad de información en forma clara, concisa y rápida, además de que pueden esclarecer relaciones complejas en los datos

Gráficos Univariados

Gráficos para representar una variable

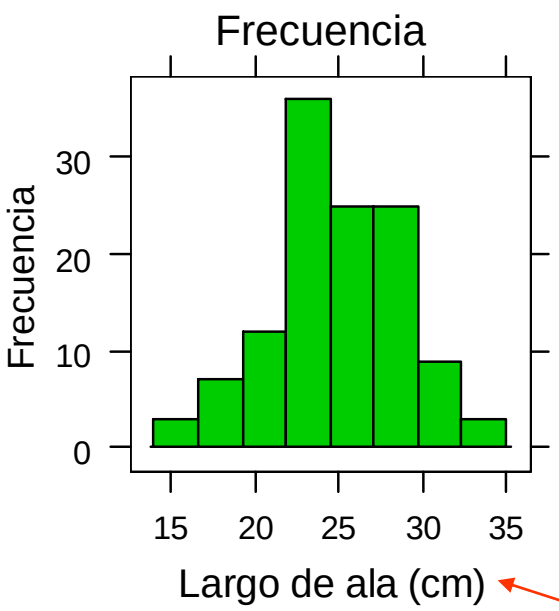
- Histograma
- Árbol de tallos y hojas (Stem-and-leaf)
- Gráfico de puntos (Dotplot)
- Gráficos de caja (Boxplot)

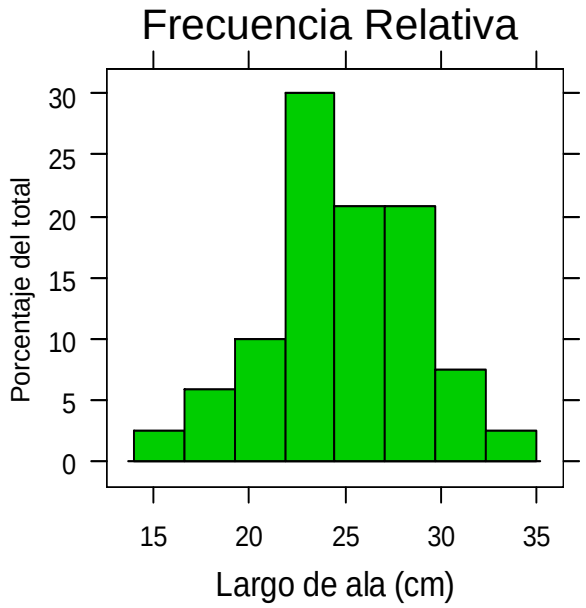
Histograma

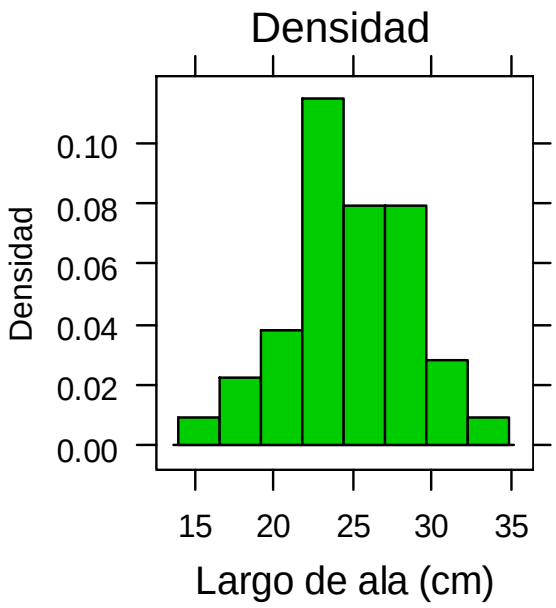
- Un histograma es un ejemplo de un gráfico de densidad, es decir cada barra describe la frecuencia, porcentaje o densidad de los valores que se incluyen en los datos entre el límite inferior y el límite superior de cada barra.

Ejemplo

Largo de ala para una especie de ave







Variable continua

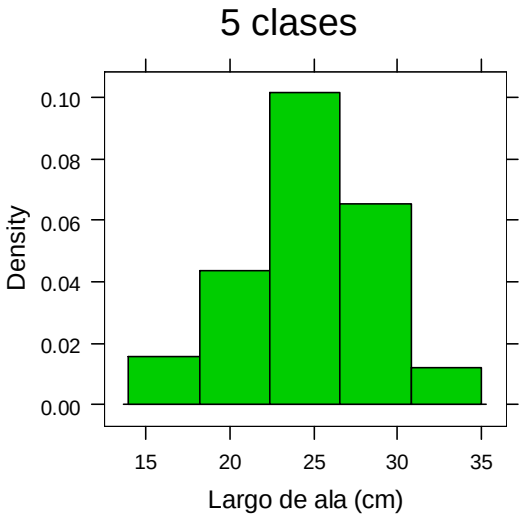
Histograma

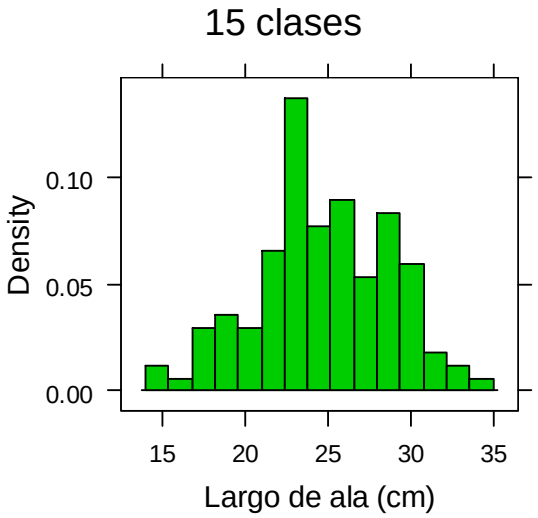
Ventajas

- Es útil para apreciar la forma de la distribución de los datos, si se escoge adecuadamente el número de clases y su amplitud.

Desventajas

- Los observaciones individuales se pierden.
- La división en clases o intervalos es arbitraria. El cambiar el número de intervalos cambia la forma de la distribución.
- No pueden derivarse estadísticos básicos.





Número de intervalos

Regla de Sturge:

$1 + 3{,}3 \times \log(n)$

log = logaritmo en base 10

Histogramas para visualizar la distribución de la variable

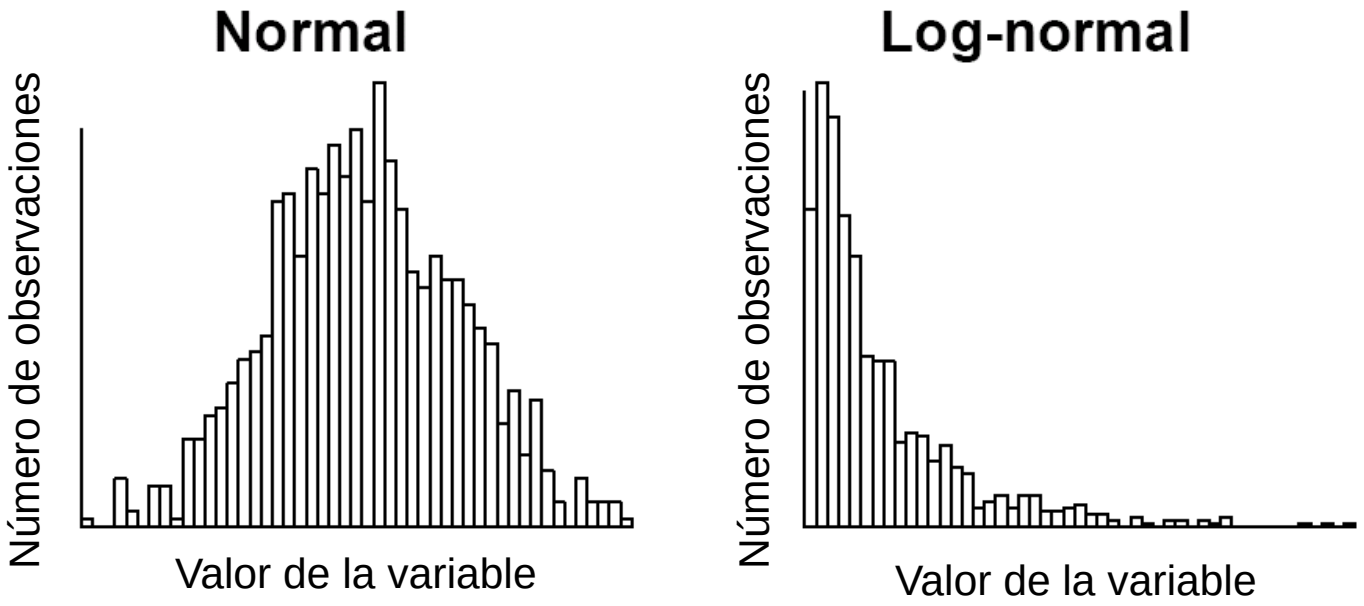
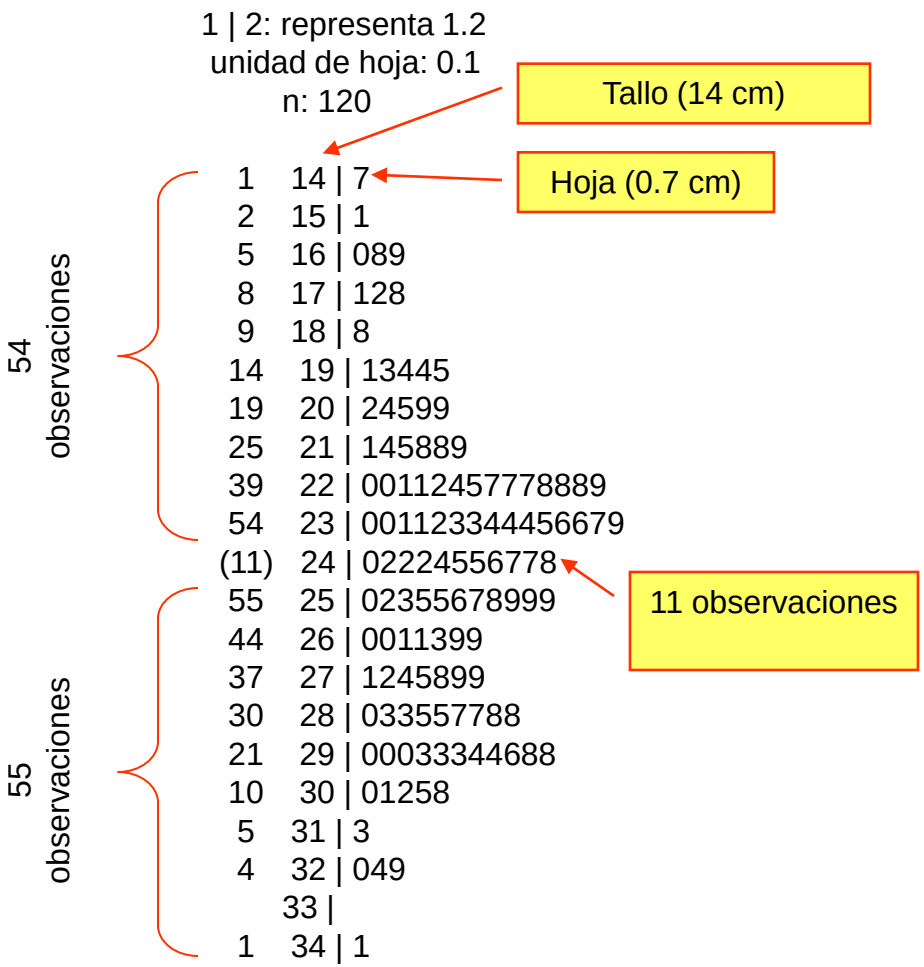


Gráfico de Tallo y Hojas

- Es una simple alternativa al histograma.
- A diferencia del histograma, todos los valores pueden ser visualizados.



¿Que nos muestra?

1. El centro de la distribución.
2. La forma general de la distribución
   - Simétrica: Si las porciones a cada lado del centro son imágenes espejos de las otras.
   - Sesgada a la izquierda: Si la cola izquierda (los valores menores) es mucho más larga que la de la derecha (los valores mayores)
   - Sesgada a la derecha: Opuesto a la sesgada a la izquierda.
3. Desviaciones marcadas de la forma global de la distribución.
4. Outliers: Observaciones individuales que caen muy por fuera del patrón general de los datos.
5. Gaps: Huecos en la distribución.

Gráfico de Tallo y Hojas

Ventajas

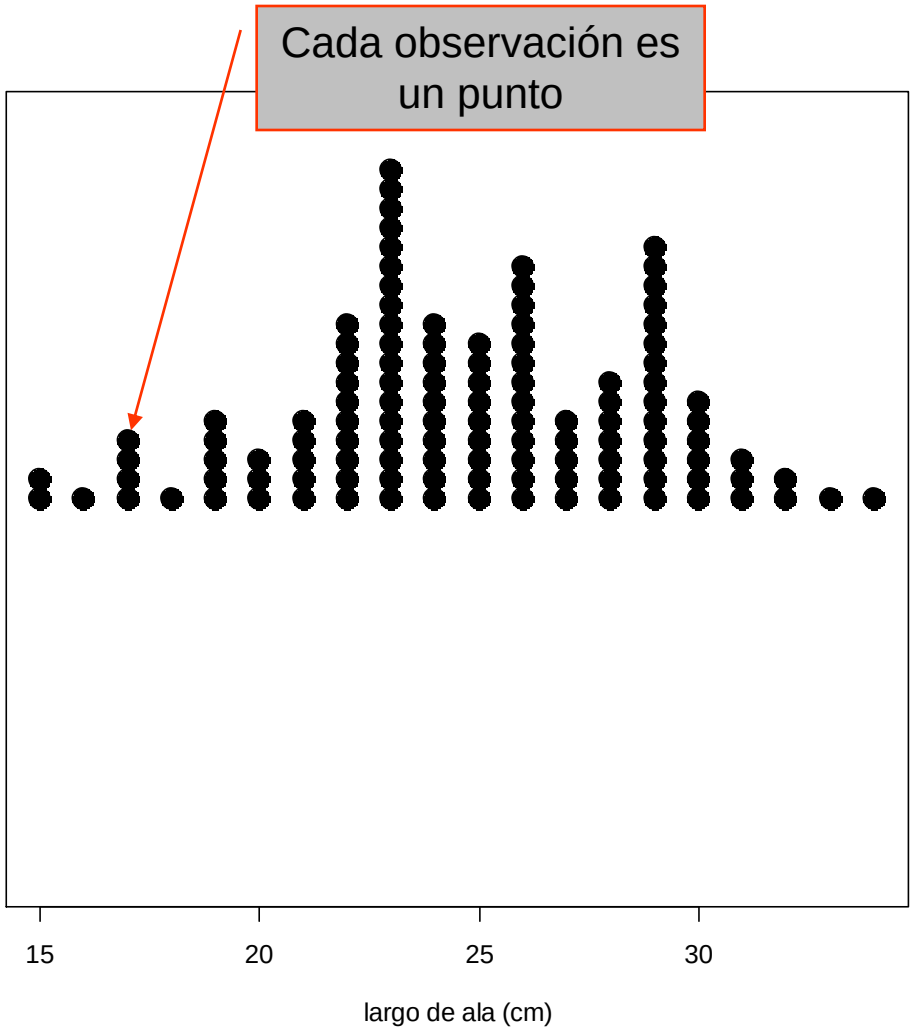
- Muy fácil de realizar y puede hacerse a mano.

Desventajas

- El gráfico es tosco y no sirve para presentaciones definitivas.
- Funciona cuando el número de observaciones no es muy grande.
- No permite comparar claramente diferentes poblaciones.

Gráfico de Puntos Univariado

- Es un diagrama donde cada observación es representada por un punto.
- El valor de la variable se representa a lo largo del eje horizontal.



Ventajas

- Todos los valores son presentados.
- Representan de manera efectiva la distribución de las observaciones.
- Detectan fácilmente asimetría y valores inusuales.
- Fáciles de entender.

Gráfico de caja (Boxplot)

- También llamados “gráfico de caja y bigotes” (box-and-whiskers plot) o Caja de Tukey.
- Efectivos para tamaños de muestra ≥ 8.
- Provee más estadísticos básicos que el histograma.

Ventajas

- Como se basan en la mediana son robustos a valores inusuales.
- Indican la variabilidad de lo datos por la distancia entre los bigotes.
- Indican la forma de la distribución, especialmente si es simétrica o asimétrica.
- Permiten la comparación de varios grupos simultáneamente.
- Indican la presencia de outliers (valores extremos).

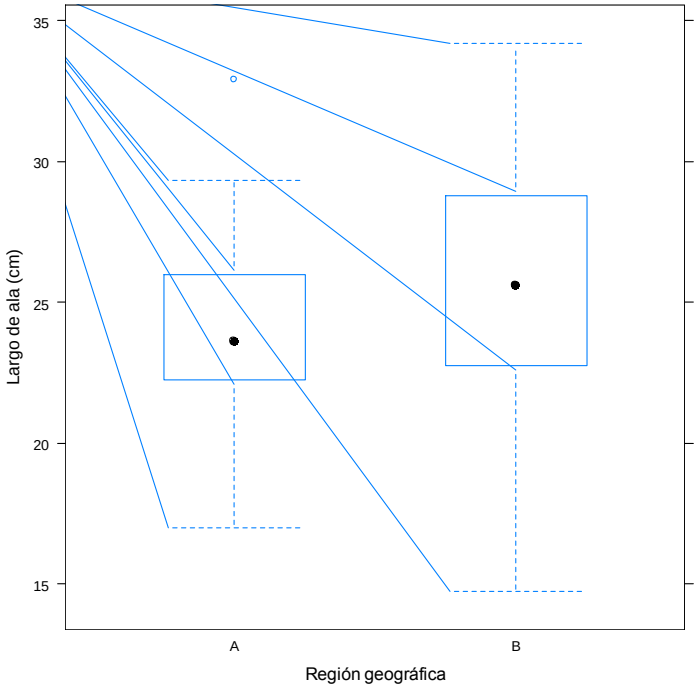


Gráfico de caja (Boxplot)

- La fórmula para identificar outliers varía.
- Generalmente se definen como aquellos valores 1,5 veces el rango intercuartil (spread), es decir, la diferencia entre el cuartil superior e inferior.
- La dispersión esta dada por la altura de la caja, así como por la distancia entre los extremos de los bigotes.

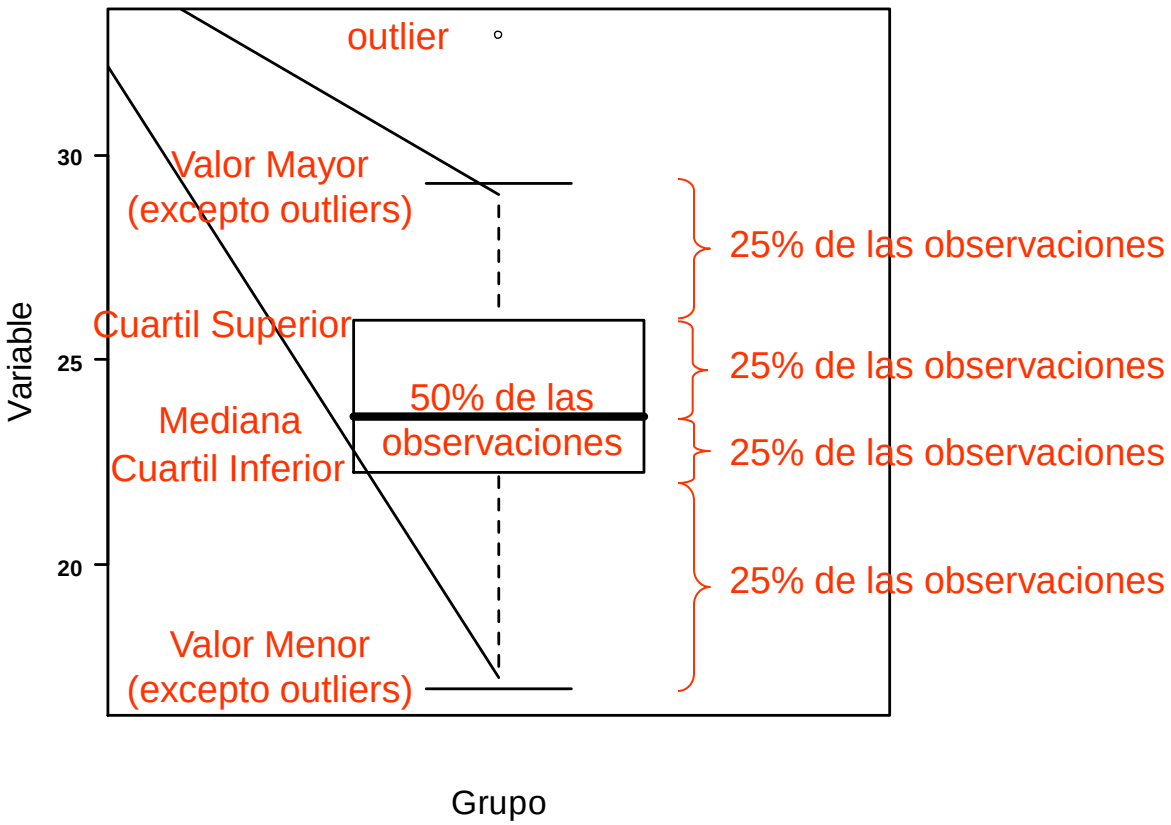
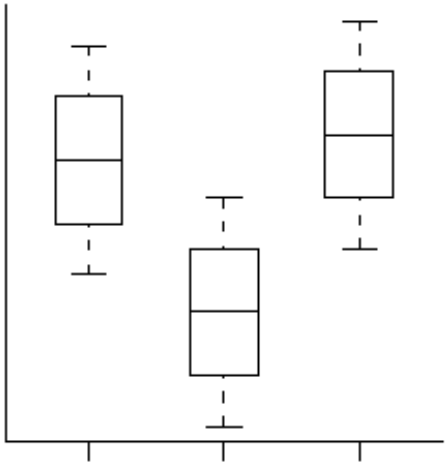
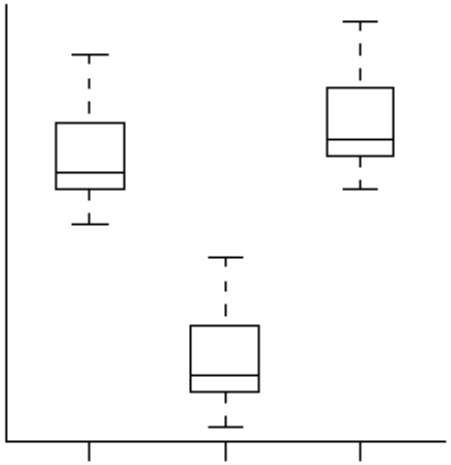


Gráfico de caja (Boxplot)

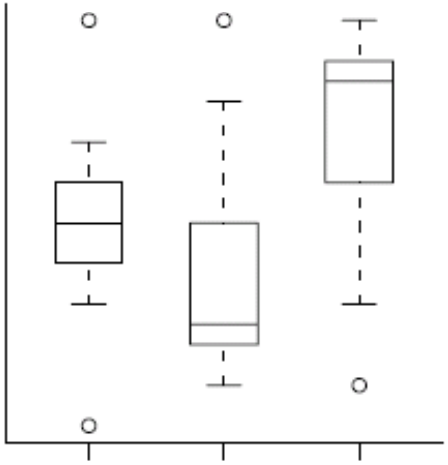
1. Ideal



2. Asimetría



3. Outliers



4. Varianzas heterogéneas

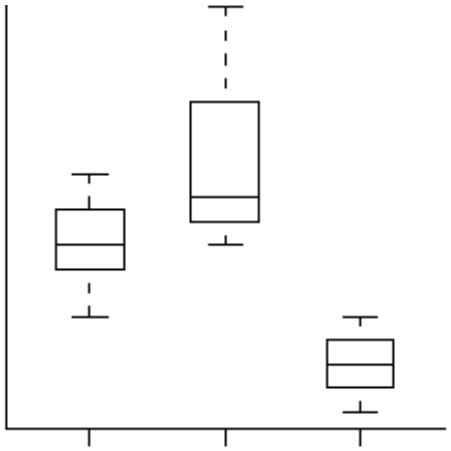
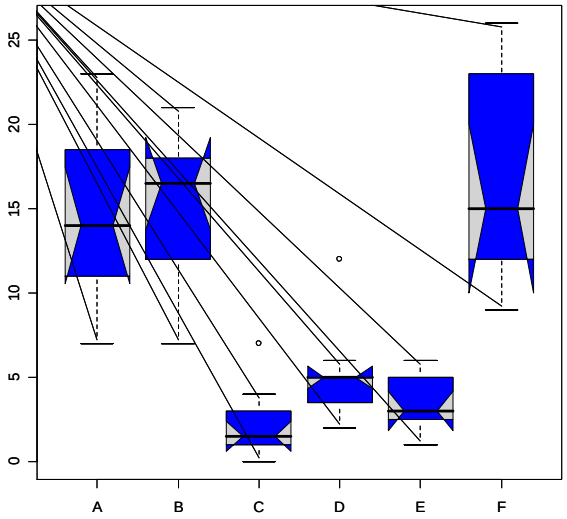
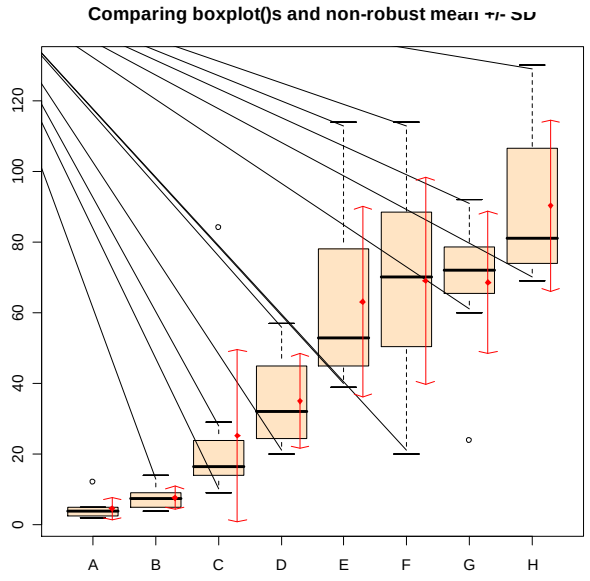


Gráfico de caja en R





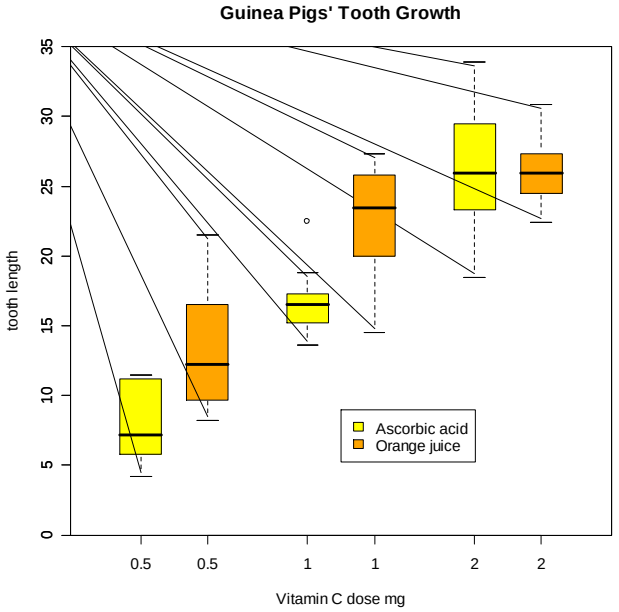


Gráfico Bivariados

- Estudios en recursos naturales generalmente exploran las relaciones entre 2 o más variables continuas.
- Dos preguntas generales pueden ser examinados con métodos gráficos.
  1. ¿Existe alguna relación entre las 2 variables?
  2. ¿Existen atípicos (outliers)?

     Atípicos: puntos que desproporcionadamente afectan la aparente relación entre las 2 variables.

Gráfico de Dispersión (Scatterplot)

- En los gráficos de dispersión el eje de la x está representado por una variable y el eje de las y está representado por otra variable.
- Los puntos en el gráfico son las observaciones.
- Permiten identificar relaciones no-lineales y outliers.
- Muy útiles cuando son acompañados de gráficos de cajas.

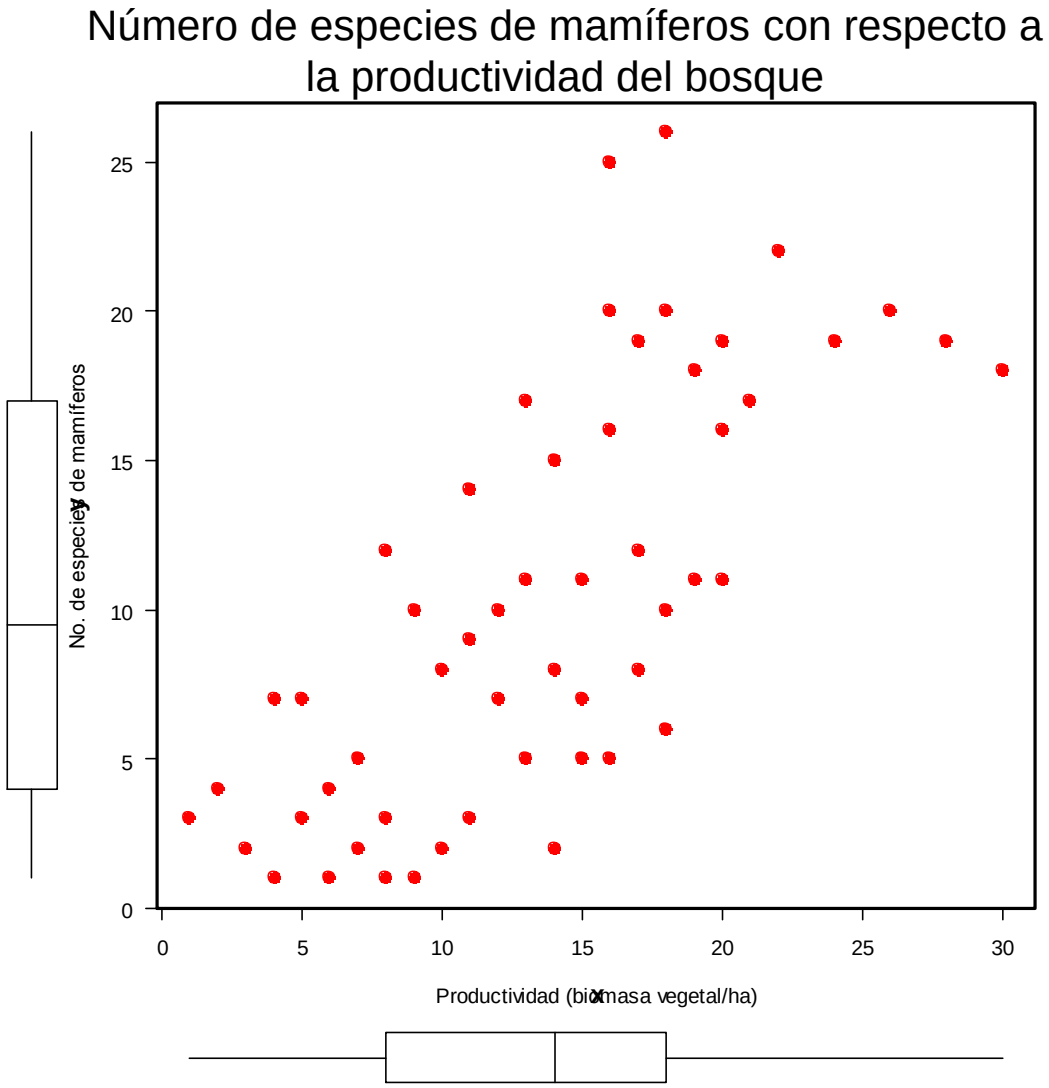
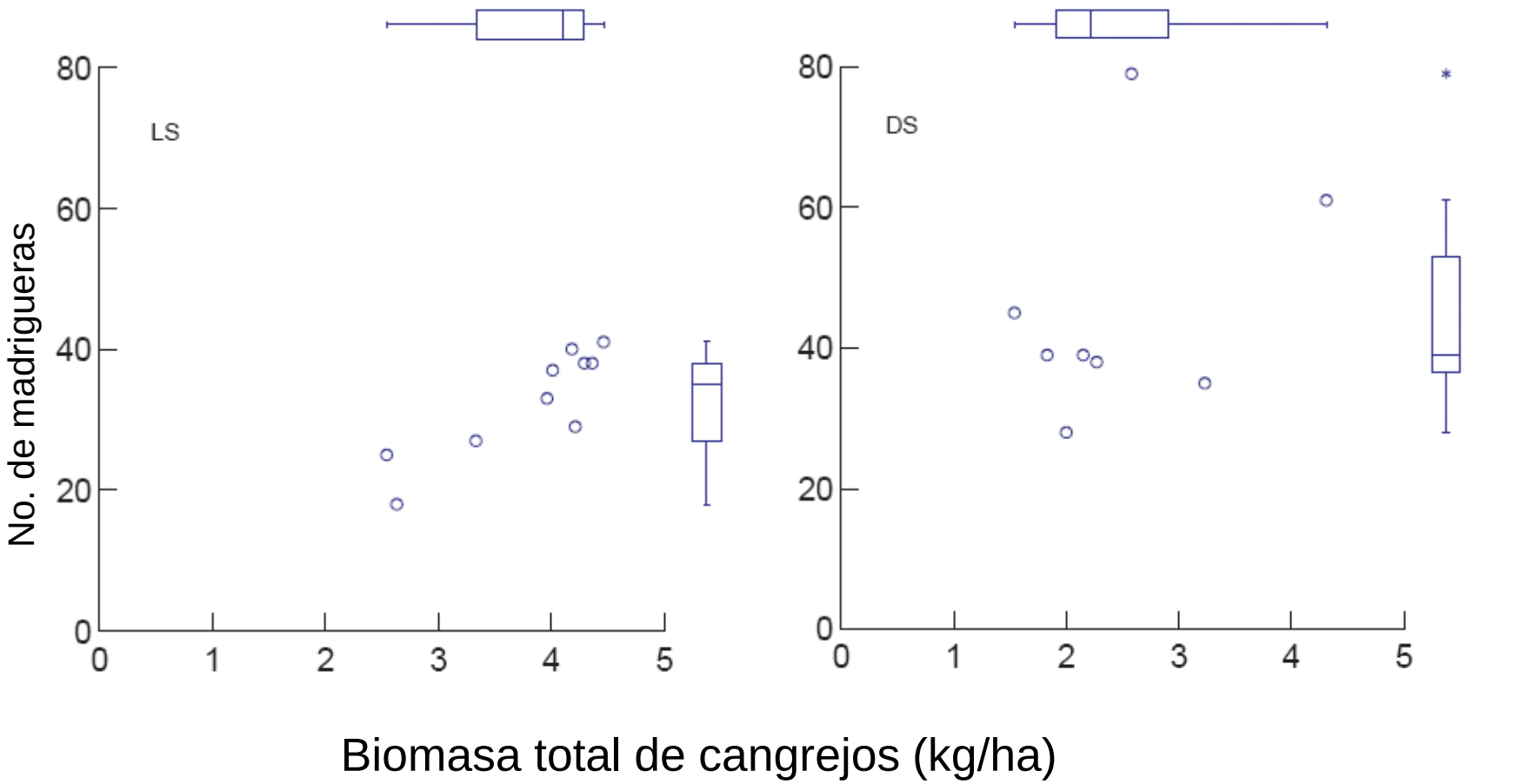


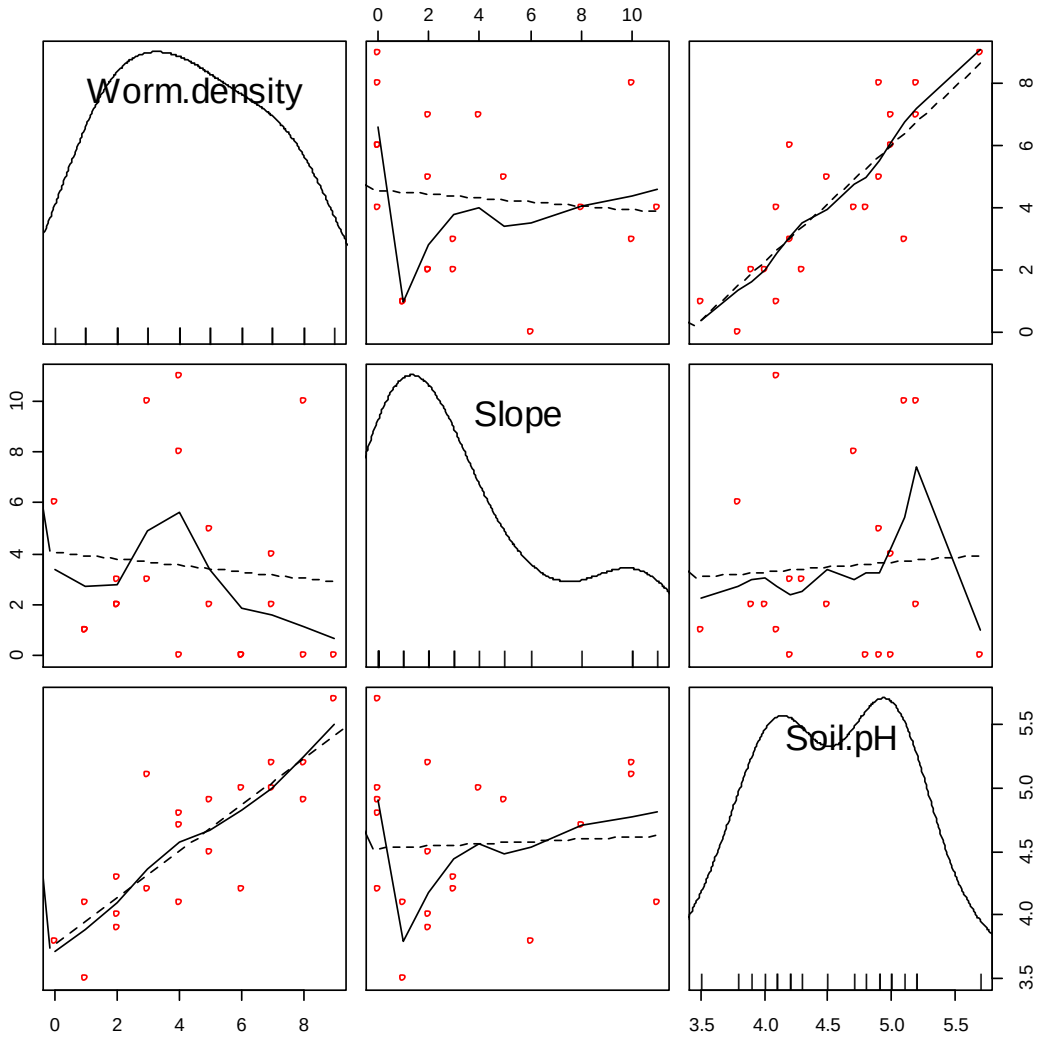
Gráfico de Dispersión (Scatterplot)

Número de madrigueras en relación a la biomasa de cangrejos en 2 islas



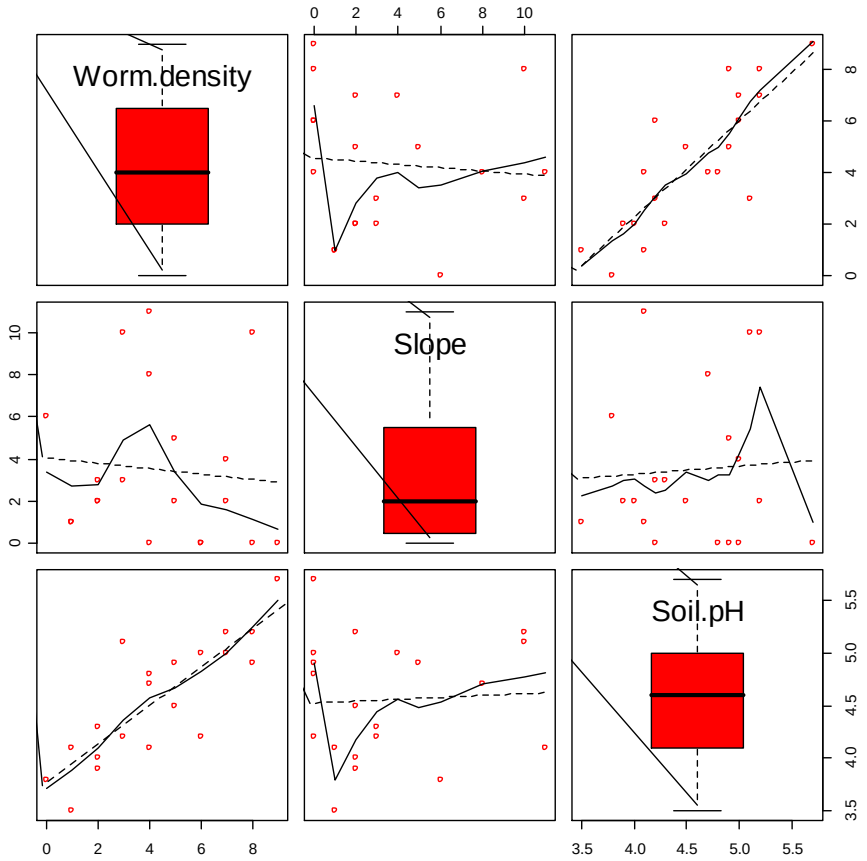
Matriz de Dispersión (SPLOM)

- Es una extensión del gráfico de dispersión para acomodar la comparación de varios pares de variables.
- Consiste en una matriz donde cada entrada presenta un gráfico de dispersión sencillo.
- Se pueden incluir gráficos univariados en los paneles de la diagonal.



Matriz de Dispersión (SPLOM) en R

Con gráficos de caja



Con histogramas

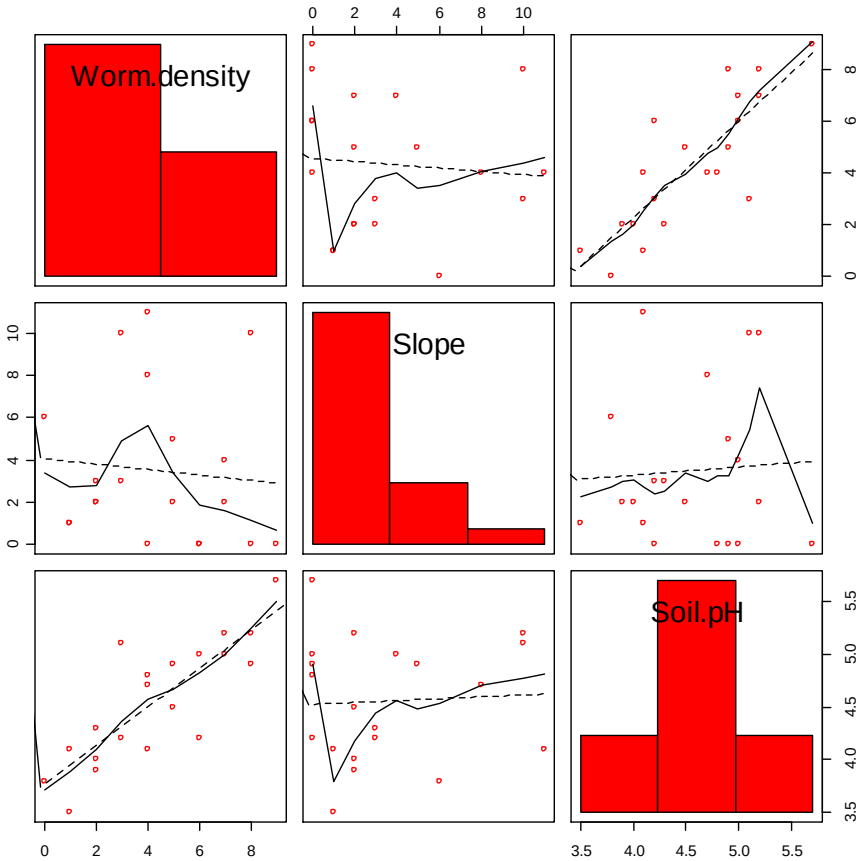
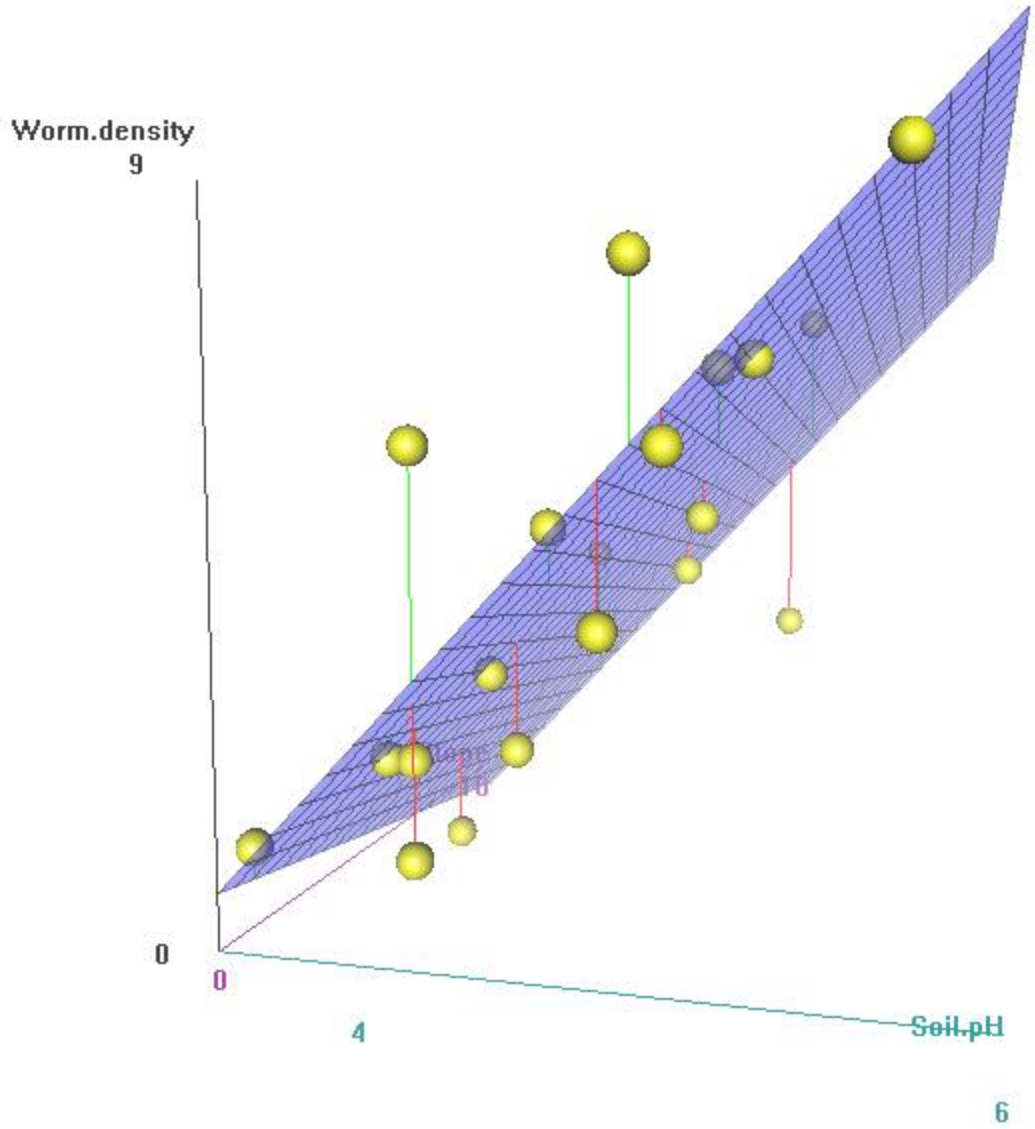


Gráfico de Dispersión de 3 dimensiones en R

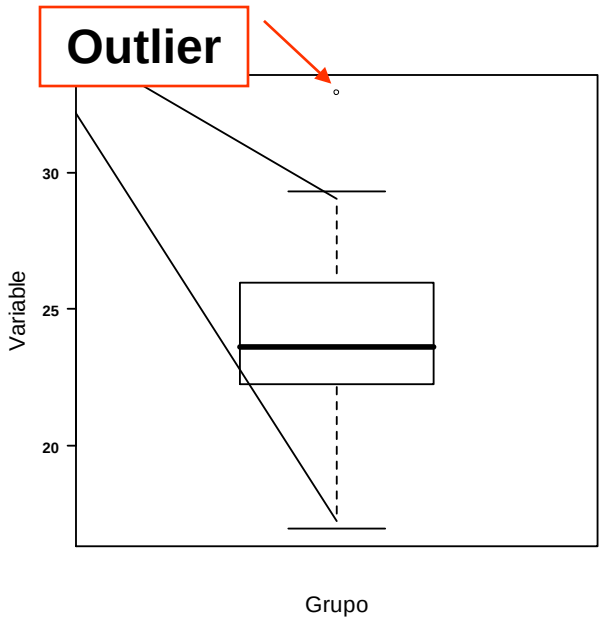


¿Qué buscar en los gráficos?

- Atípicos (Outliers)
- Asimetría de la distribución
- Cambios en variabilidad
- No-linealidad

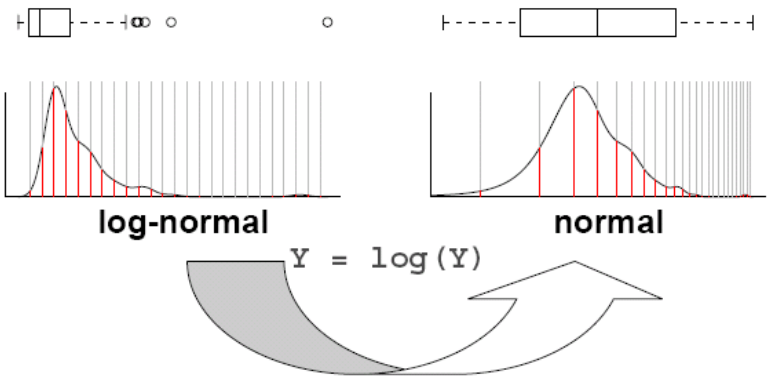
Atípicos (outliers)

- Valores inusuales. Observaciones que caen afuera del rango usual de la variable.
- Estas observaciones pueden ser errores u observaciones genuinas.
- Estos pueden distorsionar cualquier modelo que se quiere ajustar a nuestros datos.
- Para detectar atípicos en una dimensión se pueden usar gráficos de cajas (boxplots).
- Los atípicos también pueden ser detectados con gráficos de probabilidad (pplot o qqplot).
- Para detectar atípicos en dos o tres dimensiones se pueden usar gráficos de dispersión.
- La presencia de atípicos pueden indicar una violación de los supuestos del modelo.



Asimetría de la distribución

- Las distribuciones asimétricas incluyen asimetría positiva y negativa.
- La asimetría positiva es la más común. Distribuciones con una cola larga hacia la derecha, valores cerca del mínimo están agrupados, y los valores grandes están bien dispersos.
- Si todos los valores son mayores que cero, una transformación logarítmica puede normalizar la distribución.
- La normalidad de una distribución se puede ver con gráficos de cajas, o más específicamente, con gráficos de probabilidad (pplots o qqplots).



Asimetría de la distribución

Muchas variables usadas en estudios de recursos naturales presentan una asimetría positiva (se puede decir que en recursos naturales las distribuciones asimétricas son más comunes que distribuciones simétricas).

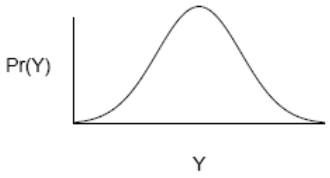
Las variables de medidas (peso, longitud, etc.) generalmente presentan una distribución log-normal

No pueden tomar valores menores que 0 (distribución truncada en 0), pero matemáticamente no tienen limite superior (aunque si tienen límite biológico).

Una transformación puede corregir la asimetría.

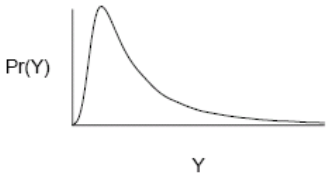
Los conteos generalmente se ajustan a una distribución de Poisson.

Distribución de algunas variables



Distribución simétrica

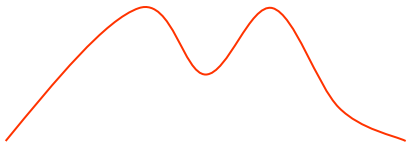
- Distribución normal



Distribución asimétrica

- Log-normal
- Poisson

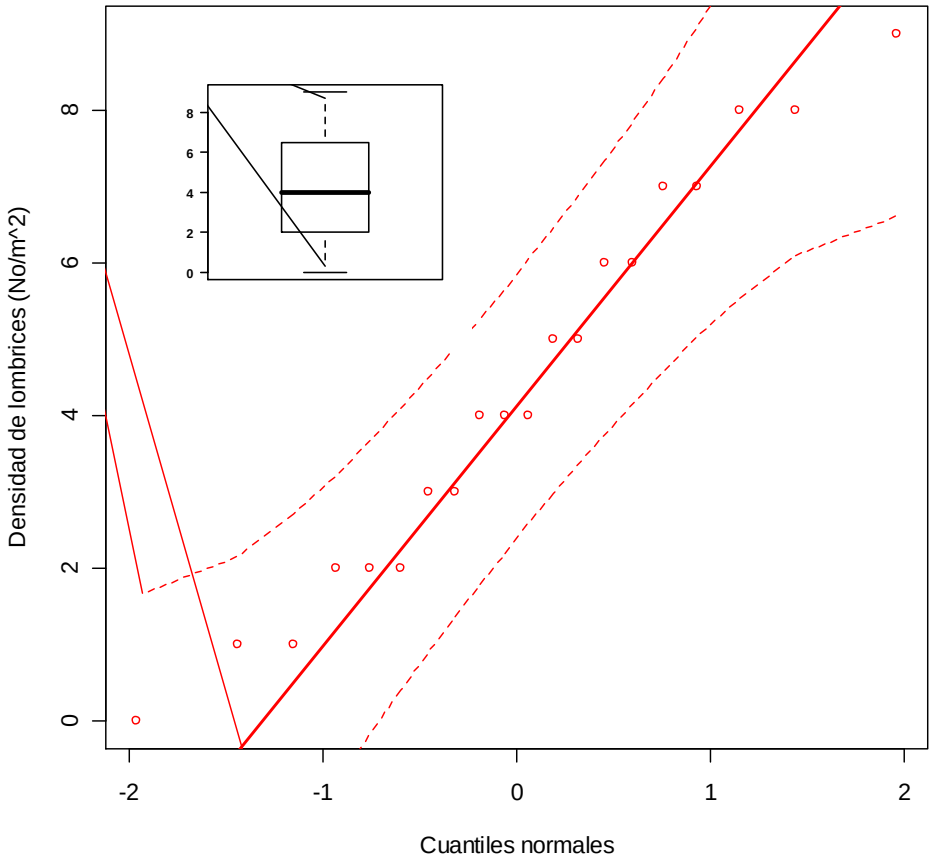
Otra distribución problemática es la multimodal



Evaluando Normalidad

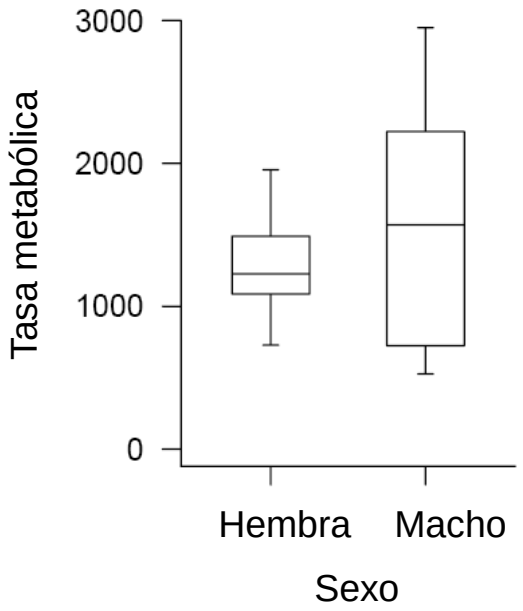
Gráficos de probabilidad (pplots qqplots).

- Examinan una frecuencia acumulada de los datos y la compara la forma de la distribución con la esperada para una distribución normal que tiene la misma media y varianza.
- Si los datos se ajustan a una distribución normal el gráfico de probabilidad se mostrará como una línea recta.
- El gráfico es útil para tamaños de muestra mayores a 25.



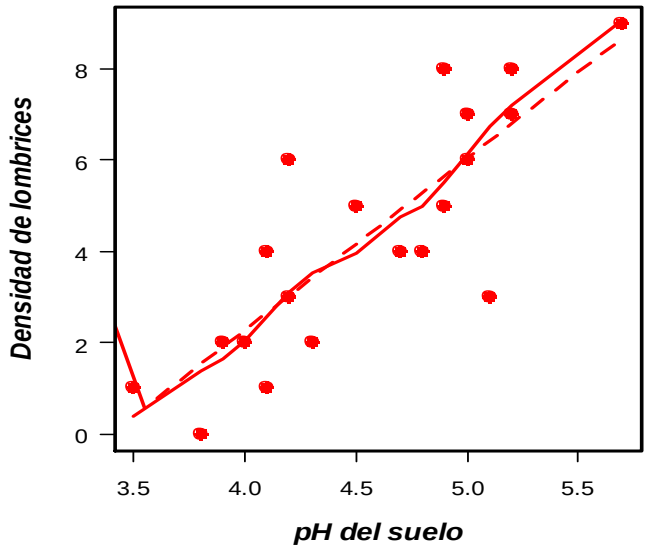
Cambios en Variabilidad

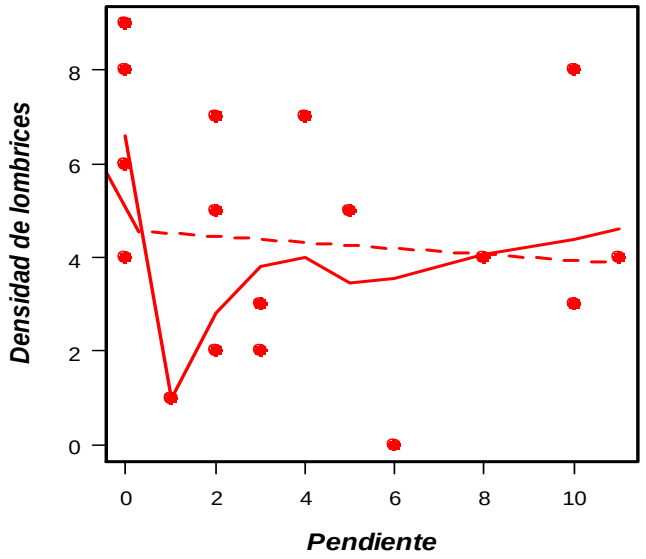
- Gráficos de caja e histogramas dan una idea de la cantidad de variabilidad o dispersión en los datos.
- Muchos modelos estadísticos dependen del supuesto de que la variabilidad es constante a través de los grupos (homogeneidad de varianza u homoscedasticidad).
- Cuando la variabilidad aumenta a medida que los valores de las observaciones aumenta, la transformación logarítmica puede ayudar a corregir la heterogeneidad de varianza.
- Comparación de gráficos de caja son útiles para comparar variabilidad entre grupos.



No-linealidad

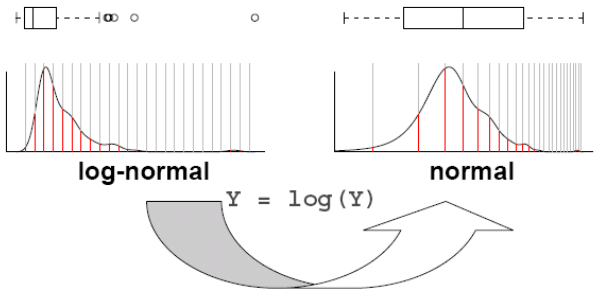
- No se debería ajustar un modelo lineal a una relación entre variables que no es lineal.
- A menudo es posible transformar variables para linealizar la relación. Si esto no resulta hay otras posibilidades.
- Recuerden, en estadística los datos son sagrados (excepto que estos sean defectuosos debido al proceso de colecta). Los modelos se ajustan a los datos y no los datos a los modelos.





Transformaciones

- La transformación de una variable puede solucionar problemas de asimetría, heterogeneidad de varianza, no-linealidad y atípicos (outliers).
- Transformar una variable tiene 5 objetivos:
  1. Hacer que la distribución de la variable se acerque a una distribución normal.
  2. Reducir cualquier relación que pueda existir entre al media y la varianza (mejorar la heterogeneidad de varianza).
  3. Reducir la influencia de atípicos (outliers).
  4. Hacer más lineal la relación entre variables (análisis de regresión).
  5. Hacer que los efectos que son multiplicativos en la escala original se hagan aditivos en la escala transformada, esto es, reduce el tamaño de los efectos de interacción.



Transformaciones

- Una vez que se transformó la variable se debe corroborar que la transformación mejoró la distribución de la variable.
- El pH se mide en escala logarítmica.
- La transformación cambia la variable y por lo tanto la hipótesis nula. Por ejemplo, si se hace una transformación logarítmica de la variable:

H0: La longitud de alas es igual en ambas regiones

debe plantearse como,

H0: La log-longitud de alas es igual en ambas regiones

Transformaciones

- La transformación es una función matemática que es aplicada a todas las observaciones de una variable.

$$Y^* = f(Y)$$

- Transformación logarítmica
- Transformación de raíz (cuadrada u otras)
- Transformación arcoseno o angular
- Transformación recíproca
- Transformación de Box-Cox

Transformaciones

Transformación Logarítmica

Log-normal → Normal

$$Y^* = \log(Y)$$

- Reemplaza el valor de cada observación con su logaritmo.
- Las transformaciones logarítmicas son útiles para datos de medida.
- El logaritmo de 0 no puede ser definido.
- Para variables que presentan valores = 0, se debe usar log(Y+c), donde c es una constante (por ejemplo: 0,05)

Transformaciones

Transformación de Raíz (Potencia)

Poisson → Normal

$Y^* = \sqrt[n]{Y}$ por ejemplo: $Y^* = \sqrt[2]{Y}$ o $Y^* = \sqrt[4]{Y}$

- Reemplaza el valor de cada observación con su raíz (cuadrada, cúbica, etc.)
- Las transformaciones de raíz son útiles para normalizar datos de conteo.
- La raíz de 0 = 0, por lo que valores de 0 no son transformados, algunos autores sugieren agregar 0.5 a cada valor.
- La raíz 4ta se puede usar para corregir variables muy asimétricas.

Transformaciones

Transformación Arcoseno (angular)

$$Y^* = \text{arcoseno}\sqrt{Y}$$

- Reemplaza el valor de cada observación por el arcoseno de la raíz cuadrada del valor.
- Se usa con proporciones y porcentajes.
- Si los datos son porcentajes se deben convertir a proporciones.

Transformaciones

Transformación Recíproca

$$Y^* = \frac{1}{Y}$$

- Reemplaza el valor de cada observación con su recíproco.
- Se usa comúnmente con tasas (ejemplo, número de crías por hembra).

Transformaciones

Transformación de Box-Cox

$$Y^* = (Y^\lambda - 1)/\lambda$$ (para $\lambda \neq 0$)

$$Y^* = \log_e(Y)$$ (para $\lambda = 0$)

- λ es el número que maximiza una función logarítmica de verosimilitud.
- También se la denomina transformación de potencia generalizada.
- Constituye una familia de transformaciones.

Transformaciones

Retransformando los datos transformados

- Si tenemos que la media de un conjunto de observaciones = 421
- La media de los datos log-transformados = 5,999
- Para retransformarlo tenemos que tomar el antilogaritmo de la media de los datos transformados $\rightarrow$ exp(5,999) = $2{,}71828^{(5{,}999)}$ = 403 (se parece a la mediana)
- La media retransformada es siempre menor que la media sin transformar.

$$\bar{Y}_{antilog} \approx \exp(\bar{Y}_{\log} + \frac{s^2_{\log}}{2})$$

$$\bar{Y}_{antilog} \approx \exp(5{,}999 + \frac{0.3^2}{2}) \approx 422$$

Gráficos univariados

Escala de medida de la variable	Gráfico
Nominal u Ordinal	• Gráfico de frecuencias • Gráfico de barras • Gráfico de barras segmentado
Intervalo o Racional	• Gráfico de frecuencia • Gráfico de puntos • Histograma • Gráfico de tallos y hojas • Gráfico de líneas • Gráfico de caja • Gráfico de probabilidades normales

Relaciones entre 2 variables

Tipo de variable	x es intervalo o racional	x es nominal u ordinal
y es intervalo o racional	• Gráfico de dispersión • Regresión • Correlación	• Gráfico de punto comparativo • Gráfico de caja comparativo
y es nominal u ordinal	• Regresión logística	• Gráfico de mosaico • Tabla de contingencia