Probabilidad y Estadística Fundamental

El gráfico de caja y detección de outliers

Profesor: Nicolás López

Universidad Nacional de Colombia

Contenido

Contenido

Motivación

" Summaries are necessary for vast amounts of data – and often convenient for smaller amounts. They are not supposed to – and cannot be expected to – replace the corresponding details. Often, of course, the details will add little, but it is important to prepare for the occasions when they add much."

John Tukey, *Exploratory Data Analysis*.

Contenido

Gráfico de caja

Interpretación simultánea, numérica y gráfica, de la distribución de los datos

A partir de gráficas/tablas se obtiene una representación general de la distribución de la variable, y a partir de parámetros, logramos describir de manera mas detallada su comportamiento.

1. ¿Cómo mejorar la representación gráfica con estos nuevos resultados numéricos?
2. ¿Cómo detectar valores inusuales de una variable?

Gráfico de caja

Interpretación simultánea, numérica y gráfica, de la distribución de los datos

Cinco números que dan un resumen rápido de la estructura del conjunto de datos son:

- Mediana: medida de centralidad (centro robusto y localización).
- Mínimo: medida de localización (dispersión y localización).
- Máximo: medida de localización (dispersión y localización).
- Q_1: medida de localización (dispersión robusta).
- Q_3 medida de localización (dispersión robusta).

Implícitamente también nos comunican características de forma.

Gráfico de caja

Interpretación simultánea, numérica y gráfica, de la distribución de los datos

Figura 1: Gráfico de caja con 5 medidas de resumen. Tomado de *Exploratory Data Analysis*

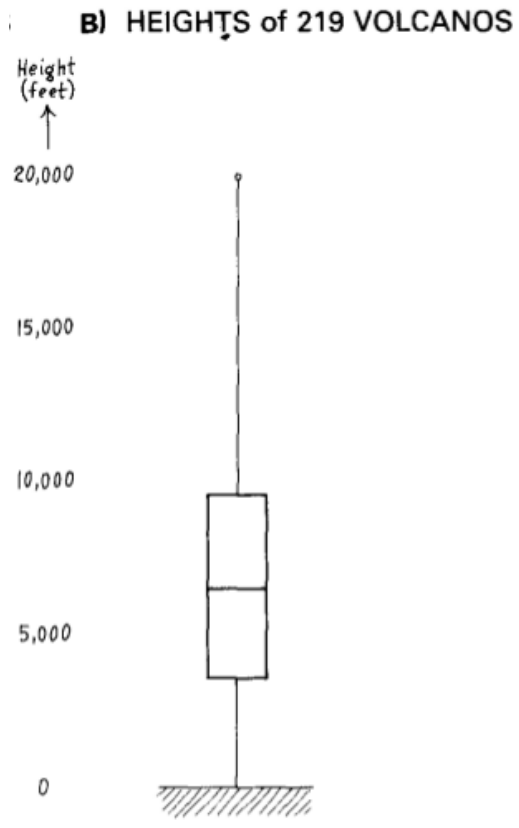


Gráfico de caja

Interpretación simultánea, numérica y gráfica, de la distribución de los datos

Esta representación es útil y permite la identificación de las 5 medidas resumen, sin embargo, no permite determinar valores inusualmente grandes o pequeños, para esto, una **regla de referencia** es necesaria. Se definen:

- Vallas internas:

$$[Q_1 - 1{,}5IQR, Q_3 + 1{,}5IQR]$$

- Vallas externas:

$$[Q_1 - 3{,}0IQR, Q_3 + 3{,}0IQR]$$

Gráfico de caja

Interpretación simultánea, numérica y gráfica, de la distribución de los datos

Con lo cual

- Si una observación se encuentra dentro de las vallas, es una observación usual.
- Si una observación está fuera de las vallas internas, pero dentro de las externas, es llamada *outside* ó *outlier*.
- Si una observación está fuera de las vallas externas, es llamada *far out* ó *far outlier*.

En general se llama **atípico** a un dato *outlier* o *far out*. El gráfico de caja puede adaptarse a esta regla de referencia.

Gráfico de caja

Interpretación simultánea, numérica y gráfica, de la distribución de los datos

Para construir un gráfico de caja con detección de outliers:

- Calcule la mediana, los cuartiles superior e inferior y el *IQR* para el conjunto de datos.
- Trace una recta horizontal que represente la escala de medición. Forme una caja con los extremos derecho e izquierdo en Q_1 y Q_3. Trace una recta vertical que pase por la caja en la ubicación de la mediana.
- Trace los bigotes de la caja hasta los valores *adjacentes* (valores más extremos **no** *outlier*).
- Resalte con un $\circ$ las observaciones *outlier*.
- Resalte con un $\odot$ las observaciones *far outlier*.

Nota: algunos gráficos de caja no distinguen entre *outliers* y *far outliers*, por lo cual usan un mismo símbolo para ambos, generalmente $*$ ó $\circ$.

Gráfico de caja

Interpretación simultánea, numérica y gráfica, de la distribución de los datos

Además de la escolaridad y de la edad, se cuenta con el salario en dólares de la muestra de $n = 3000$ personas. ¿cómo realizar un gráfico de caja con detección de *outliers*?.

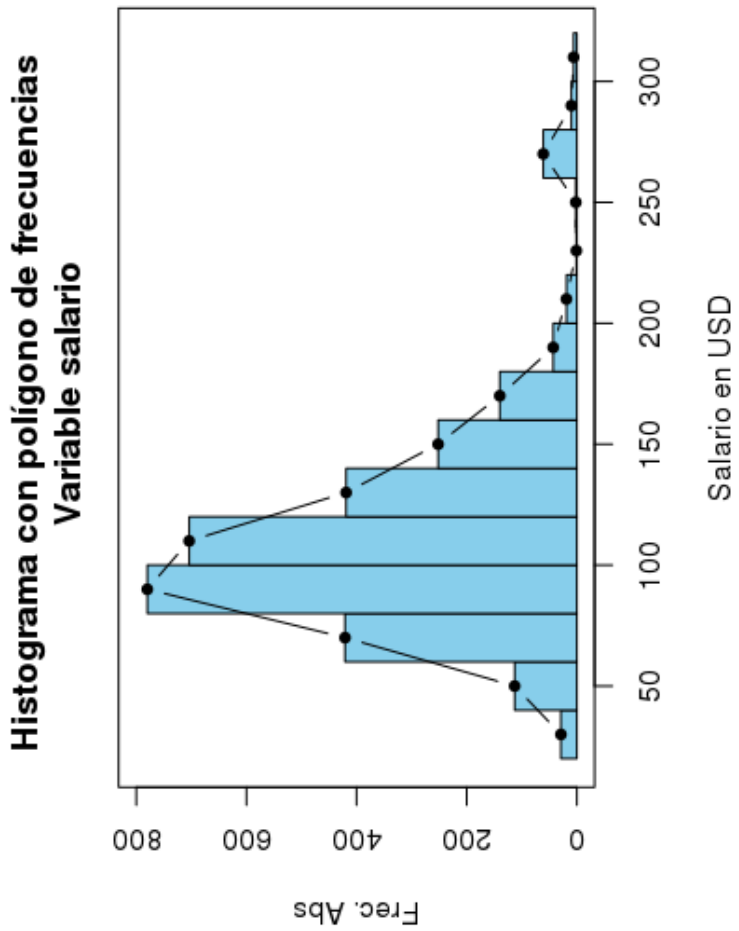


Figura 2: Histograma con polígono de frecuencias para la variable *salario*

Gráfico de caja

Interpretación simultánea, numérica y gráfica, de la distribución de los datos

Figura 3: Gráfico de caja con detección de outliers para la variable *salario*

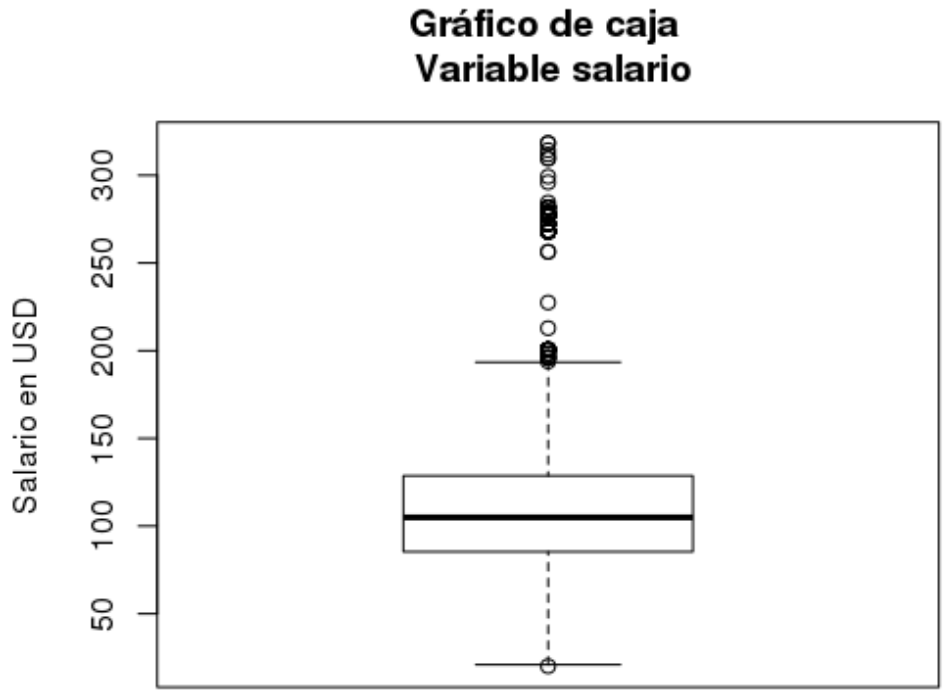


Gráfico de caja

Interpretación simultánea, numérica y gráfica, de la distribución de los datos

Notas

- El resumen de los 5 números en forma de diagrama de caja es menos informativo que gráfico de caja con detección de *outliers*, por lo cual, este último es siempre preferido e implementado por los programas.
- La definición de las vallas con 1.5 y 3.0 es un estándar de referencia generalmente aceptado.
- Algunas veces se presentan **aberraciones** en la caja. Hay que ser cuidadosos y saberlas interpretar.

Ejercicio de aplicación

Elaboración manual de Boxplot

Construya el gráfico boxplot para la variable *Estatura en cms de los estudiantes del curso de Estadística* con la siguiente información:

- La persona más baja mide 100 cm.
- ¡La distribución es completamente simétrica!.
- La segunda persona más alta mide 195 cm, la más alta 220 cm.
- El tercer cuartíl es 170 cm y la mediana 155 cm.

Contenido

Otras formas de detección de atípicos

Desigualdad de Chebyshev

Para un conjunto de n mediciones $x_1,\ldots,x_n$ con media y desviación estándar muestral $\bar{x}$ y s respectivamente, por la teorema de Chebyshev podemos afirmar que

- El "chance" de que un dato esté en el intervalo $\bar{x} \pm 2s$ es **al menos** $3/4{=}75\,\%$

$\rightarrow$ al menos 75 % de los datos están a no más de 2 desviaciones estándar de su media.

$\rightarrow$ al menos 75 % de los puntajes z de los datos están entre -2 y 2.

$\rightarrow$ los datos con puntajes z mayores a 2 o menores a -2 se presentan a lo mas 25 % del tiempo.

$\rightarrow$ Si $|z_i| > 2$, x_i es un tanto improbable de observarse.

Otras formas de detección de atípicos

Desigualdad de Chebyshev

- El "chance" de que un dato esté en el intervalo $\bar{x} \pm 3s$ es **al menos** $8/9 \approx 89\,\%$.

$\rightarrow$ al menos 89 % de los datos están a no más de 3 desviaciones estándar de su media.

$\rightarrow$ al menos 89 % de los puntajes z de los datos están entre -3 y 3.

$\rightarrow$ los datos con puntajes z mayores a 3 o menores a -3 se presentan a lo mas 11 % del tiempo.

$\rightarrow$ Si $|z_i| > 3$, x_i es un muy poco probable de observarse.

El teorema de Chebyshev aplica a cualquier conjunto de mediciones, **independiente de su distribución**.

Otras formas de detección de atípicos

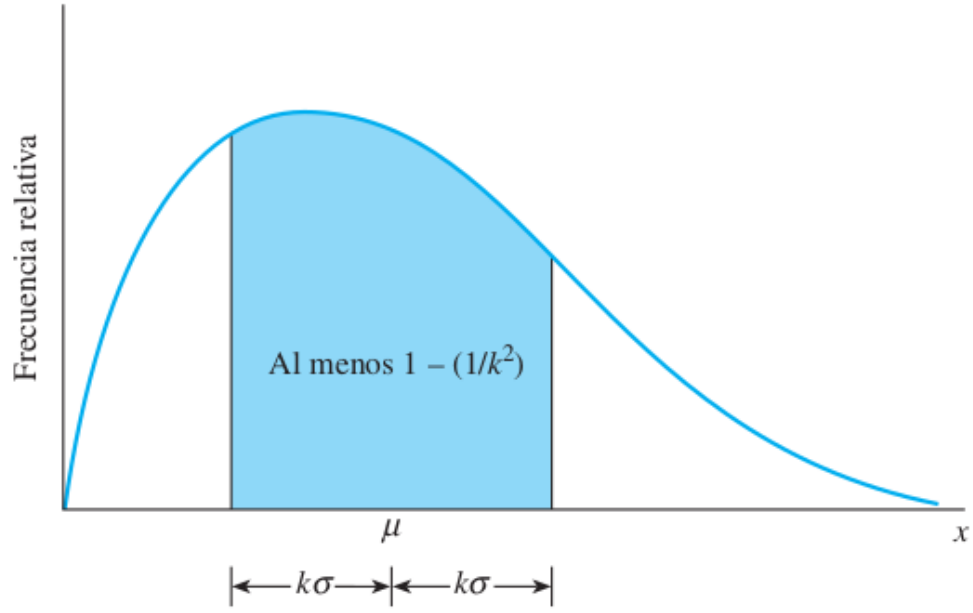
Desigualdad de Chebyshev

En general,

- El "chance" de que un dato esté en el intervalo $\bar{x} \pm ks$ es **al menos** $1 - (1/k^2)$.

$\rightarrow$ (...)

Figura 4: Ilustración del teorema de Chebyshev. Tomado de Mendenhall, Beaver, Beaver.



Otras formas de detección de atípicos

Regla empírica

Para un conjunto de n mediciones $x_1,\ldots,x_n$ con media y desviación estándar muestral $\bar{x}$ y s respectivamente, por la regla empírica podemos afirmar que

- El "chance" de que un dato esté en el intervalo $\bar{x} \pm s$ es de 68 %.

$\rightarrow$ datos con $|z_i| > 1$ se presentan un 32 % del tiempo.

- El "chance" de que un dato esté en el intervalo $\bar{x} \pm 2s$ es de 95 %.

$\rightarrow$ datos con $|z_i| > 2$ se presentan un 5 % del tiempo.

- El "chance" de que un dato esté en el intervalo $\bar{x} \pm 3s$ es de 99 %.

$\rightarrow$ datos con $|z_i| > 3$ se presentan un 1 % del tiempo.

La regla empírica aplica para conjuntos de mediciones con **distribución acampanada**.

Otras formas de detección de atípicos

Conclusión

Notas

- El teorema de Chebyshev da un **límite inferior** a la fracción de mediciones a encontrar en un intervalo construido como $\bar{x} \pm ks$.
- Cuando se usen estas dos herramientas para describir un conjunto de mediciones, el teorema de Chebyshev siempre se satisface pero es una estimación **muy conservadora** de la fracción de mediciones que caen en un intervalo particular.
- Si es apropiado usar la Regla empírica, esta regla dará una estimación **más precisa** de la fracción de mediciones que caen en el intervalo.

Conclusiones

- La detección de atipicidades puede detectar características inusuales en los datos. Dichas características pueden ser errores de digitación o, más interesante, realizaciones anomalas de la variable de interés.
- La detección de atipicidades puede realizarse de múltiples formas: observamos tres en esta sección.

Ejercicio de aplicación

Cuarto laboratorio de programación en R

Realice el cuarto laboratorio de programación en R.