(1994–2024)
30 años de la Consagración Constitucional de la Autonomía y Autarquía Universitaria en Argentina.

UNIVERSIDAD NACIONAL DEL LITORAL

FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS

ESTADÍSTICA

NOTAS DE INTRODUCCIÓN A LA TEORÍA

UNIDAD 5 – ESTADÍSTICA DESCRIPTIVA

Responsable de cátedra: Prof. Juan Pablo Taulamet

Equipo de cátedra: ***Auxiliares:*** Lic. María José Llop (JTP) - Ing. Ana Lisa Eusebi (JTP) - Prof. Fátima Bolatti (JTP) - Ing. Franco Nardi (Ay. 1°) ***Ayudantes:*** AIA Cristian Bottazzi - Téc. Eliana García

Carreras: Ingenierías en: Recursos Hídricos - Ambiental - Agrimensura

AÑO ACADÉMICO 2024

SOBRE ESTE DOCUMENTO

Las presentes notas, se encuentran destinadas a brindar una orientación para el estudio de la unidad 5, sobre Estadística Descriptiva durante el cursado de la asignatura Estadística, en el dictado para las carreras: Ingeniería en Recursos Hídricos, Ingeniería Ambiental e Ingeniería en Agrimensura.

Este documento fue actualizado por última vez el día 10/10/24 a las 08:20:00 hs.

LECTURAS RECOMENDADAS

Para el abordaje de la presente unidad se sugiere la lectura de algunos recortes disponibles en la biblioteca correspondientes al capítulo 2, del libro (LEVIN et. al, 2004, p.7-p.46) y del libro (WALPOLE et al, 2012, p.21-p.27)

ESTADÍSTICA DESCRIPTIVA

A modo de introducción, se establece que la Estadística Descriptiva (ED), se ocupa de la realización de un análisis que tiene por objetivo describir analítica y gráficamente los datos vinculados a una muestra. Naturalmente para comprender el trabajo que se llevará adelante es necesario definir lo que es una muestra y se toma como base lo planteado por los autores (LEVIN et al, 2004, p.45), que ofrecen conjunto de los términos introducidos en el capítulo 2, cuya lectura se recomienda y de lo que se retoma la siguiente definición de muestra.

MUESTRA

Una muestra es un subconjunto de elementos que pertenecen a una población. Es posible que la muestra sea representativa de la población, siempre que mantenga en las misma proporción las características más importantes de la población. Si bien a lo largo de este cursado no profundizaremos en las diferentes técnicas de muestreo, resulta interesante considerar la diferencia entre el estudio de una VA, habitualmente vinculada con la idea de una población que incluye una totalidad, del estudio de una porción de la misma, que se denomina muestra.

DISTRIBUCIÓN DE PROBABILIDADES VS DISTRIBUCIÓN DE FRECUENCIAS

De manera análoga al estudio de la distribución de probabilidades de una VA aleatoria, que se vincula con las función de cuantía (para el caso de las VA discretas) y de densidad (para el caso de las VA continuas), cuando se cuente con datos correspondientes a una muestra, se efectuará el estudio de la distribución de frecuencias. A partir de lo planteado en la unidad 3, se clasifican las VA entre discretas y continuas. En el caso de las muestras

correspondientes a las VA discretas, se considerará un conteo de la cantidad de veces que aparece en una muestra cada uno de los valores encontrados. De esta forma, cada valor de la VA que se encuentre presente en la muestra podrá relacionarse con un número de veces que aparece en la muestra y este número se denominará frecuencia absoluta y se representará con el símbolo f.

CASO DISCRETO

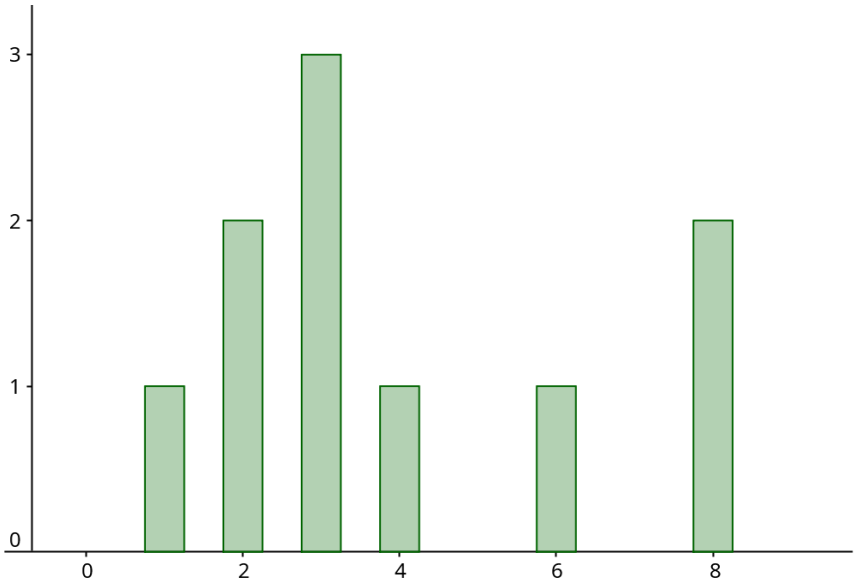
A continuación se ofrecer la siguiente muestra de datos de una VA discreta:

$$m_1 = \{1,2,2,3,3,3,4,6,8,8\}$$

A partir de la muestra de datos m_1 se puede elaborar la siguiente tabla que da cuenta de la distribución de frecuencias absolutas.

X	f
1	1
2	2
3	3
4	1
5	0
6	1
7	0
8	2

Las distribuciones de frecuencias del ejemplo anterior pueden representarse gráficamente en el caso de las VA discretas utilizando un **diagrama de barras** como el que se presenta a continuación (Realizado con el software Geogebra):



Es importante considerar que la separación de las barras simboliza que la VA es discreta.

Como se ha dicho, se puede observar que las magnitudes correspondientes a las frecuencias f, dan cuenta de la cantidad de veces que aparece un determinado valor de la VA en la muestra. En base a dicha distribución de frecuencias f, se pueden considerar las frecuencias acumuladas F, que representarán la cantidad de elementos iguales o menores al valor de la VA que están presentes en la muestra, tal como se muestra en la siguiente tabla:

X	f	F
1	1	1
2	2	3
3	3	6
4	1	7
5	0	7
6	1	8
7	0	8
8	2	10

Además de las frecuencias absolutas, pueden determinarse las frecuencias relativas, denotadas por h y las frecuencias relativas acumuladas H, que representan la cantidad de frecuencias absolutas en relación al total de elementos y que serán los valores más parecidos a lo que en la población se consideraba como función de cuantía y distribución de una VA discreta, que se relacionaban con probabilidades.

Se presentan las frecuencias absolutas y relativas, así como las acumuladas a continuación:

X	f	F	h	H
1	1	1	10%	10%
2	2	3	20%	30%
3	3	6	30%	60%
4	1	7	10%	70%
5	0	7	0%	70%
6	1	8	10%	80%
7	0	8	0%	80%
8	2	10	20%	100%

A partir de la distribución de frecuencias relativas H, se puede identificar las medidas de ubicación de manera similar a lo realizado en la población con la función acumulativa F(x), de tal forma que a partir de la tabla anterior, puede deducirse que el valor 6 de la variable X puede considerarse el percentil 80. También puede observarse que el valor 3 posee mayor frecuencia que el resto de los valores y puede asociarse por tanto al modo.

En el mismo sentido, utilizando la distribución de frecuencias relativas h, -tal como se hizo mediante el uso de la función de cuantía f(x) para el estudio de las VA discretas-, se podrá determinar las diferentes medidas características Tendencia Central, Variabilidad y Forma de la muestra, tal como se verá más adelante.

CASO CONTINUO

Se propone a continuación un ejemplo de muestra de una VA continua:

$$m_2 = \{1.1, 2.1, 2.2, 3.3, 3.3, 3.4, 5.0, 6.1, 8.4, 8.9\}$$

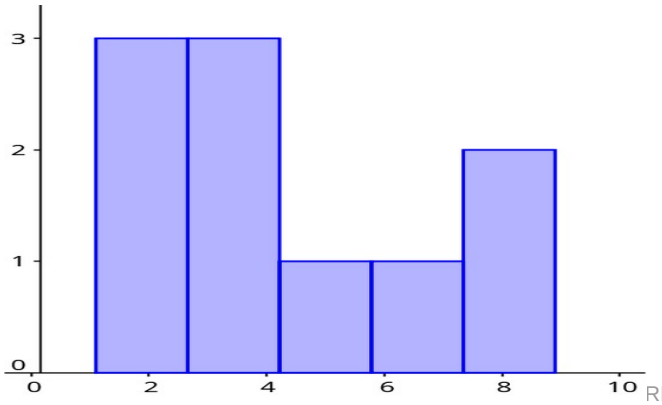
Debido a la naturaleza de la VA, para el estudio de la distribución de las frecuencias absolutas f y relativas h, se propone realizar un agrupamiento de los diferentes valores obtenidos en lo que se denominan clases (que no son otra cosa que intervalos), y se realizará el conteo de la cantidad de valores que aparecen en cada clase.

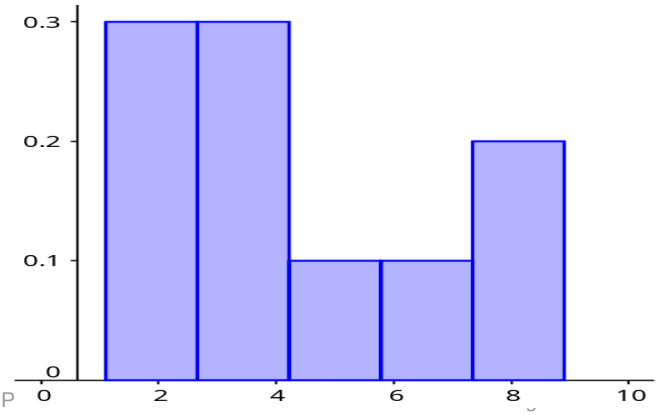
A continuación se presentan dos tablas de frecuencias absolutas (f) y relativas (h) realizadas con el programa Geogebra:

Intervalo	Frecuencia
1.1 - 2.66	3
2.66 - 4.22	3
4.22 - 5.78	1
5.78 - 7.34	1
7.34 - 8.9	2

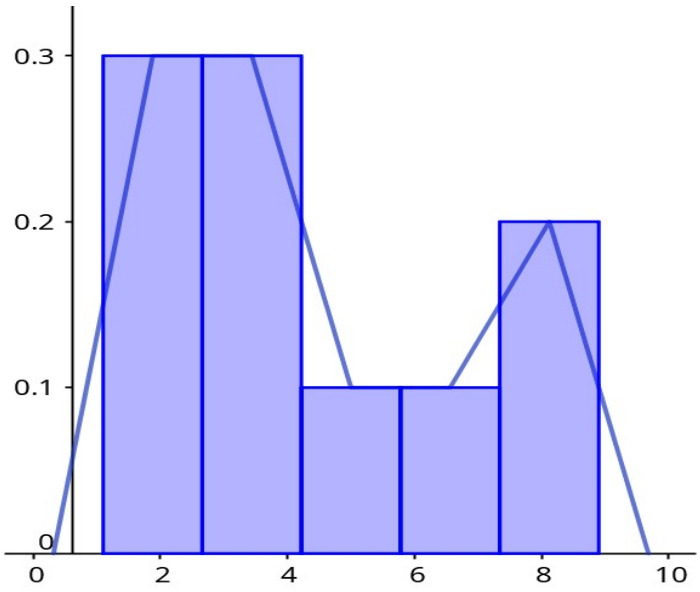
Intervalo	Frecuencia
1.1 - 2.66	0.3
2.66 - 4.22	0.3
4.22 - 5.78	0.1
5.78 - 7.34	0.1
7.34 - 8.9	0.2

La representación de la distribución de las frecuencias para una VA continua, se conoce como **Histograma** y es similar al diagrama de barras, pero eliminando la separación entre las diferentes barras, lo que simboliza la continuidad entre las clases:





A continuación se presenta un un histograma de frecuencias relativas correspondiente al ejemplo anterior, incluyendo un **Polígono de frecuencias**:



A partir de la visualización del gráfico anterior, es posible visualizar la forma de la distribución de frecuencias de la muestra, de una manera análoga a lo nos permitiría visualizar la representación gráfica de la función de densidad de la VA en la población.

Para la realización del Histograma se recomienda tomar un conjunto de clases no menor a 5 y no mayor a 20, tomando como una referencia posible la raíz cuadrada del tamaño de muestra, por ejemplo para 100 datos, 10 clases.

CARACTERÍSTICAS DE UNA MUESTRA

El estudio de las características de una población se encuentra asociado a los temas abordados durante las primeras cuatro unidades, sobre la teoría de probabilidad y particularmente cuando se definieron las principales medidas características en la unidad 3. De manera análoga al estudio de las principales características de una VA asociadas a una población conociendo sus funciones de probabilidad, se pueden estudiar las principales medidas que caracterizarán una muestra de datos.

Presentamos a continuación una tabla con algunas de las principales medidas estudiadas en la población y sus medidas análogas en la muestra, así como una simbología propuesta en cada contexto.

		Contexto	
Categorías		Población (U3)	Muestra (U5)
Tendencia Central	Promedio	μ	$\overline{x}$
	Ubicación	Me	$\tilde{x}$
		Mo	$\hat{x}$

	Contexto	
Categoría	Población (U3)	Muestra (U5)
Variabilidad	σ^2	S^2
	σ	S

En general las medidas características que se encuentran asociadas a la población, utilizaran símbolos del alfabeto griego como por ejemplo μ y σ para referirse a la Esperanza y la Desviación Típica. Cuando el contexto en el que se analizan las medidas tiene que ver con una muestra se suele utilizar símbolos asociados a nuestro alfabeto Latino, tales como x con una línea por encima $\overline{x}$ o S. Debido a que las medidas características pueden hacer referencia a diferentes contextos, suele ser preciso explicitarlo. Si bien la Esperanza suele estar asociada a la población, cuando se utilizan expresiones como “media” o “promedio” será necesario aclarar si se hace referencia a la media poblacional (es decir μ) o si se hace referencia a la media muestral (es decir $\overline{x}$). Lo mismo sucederá con la desviación típica y/o la varianza. A partir de lo anterior se recomienda explicitar el contexto o utilizar una simbología que permita evitar ambigüedades.

RESUMEN DE ESTADÍSTICA DESCRIPTIVA

Si bien en la bibliografía propuesta se encuentran las expresiones correspondientes a las definiciones de las diferentes medidas descriptivas para una muestra, se espera obtener los valores correspondientes a las mismas haciendo uso de los diferentes programas informáticos tal como se verá a continuación. En el caso del software LibreOffice Calc, es posible obtener un resumen de estadística descriptiva que incluye un conjunto de medidas descriptivas, según se presentará a continuación para las dos muestras de ejemplo que se han venido trabajando:

$m_1 = \{1,2,2,3,3,3,4,6,8,8\}$ $m_2 = \{1.1, 2.1, 2.2, 3.3, 3.3, 3.4, 5.0, 6.1, 8.4, 8.9\}$

Media	4
Error estándar	0,788810638
Moda	3
Mediana	3
Primer cuartil	2,25
Tercer cuartil	5,5
Varianza	6,222222222
Desviación típica	2,494438258
Curtosis	-0,7088648
Asimetría	0,805363117
Intervalo	7
Mínimo	1
Máximo	8
Suma	40
Recuento	10

Media	4,38
Error estándar	0,8428523
Moda	3,3
Mediana	3,35
Primer cuartil	2,475
Tercer cuartil	5,825
Varianza	7,104
Desviación típica	2,665333
Curtosis	-0,63537323
Asimetría	0,74581747
Intervalo	7,8
Mínimo	1,1
Máximo	8,9
Suma	43,8
Recuento	10

Es importante aclarar que los resúmenes de ED que pueden brindar los diferentes programas informáticos deben ser siempre interpretados y revisados considerando las limitaciones del procesamiento. En este sentido se puede aclarar que si bien el modo sugerido para la primera muestra es un valor correcto, en el caso de la segunda muestra es un error. Esto se debe a que si bien es cierto que el valor 3,3 es el valor que posee mayor frecuencia, al tratarse de una VA continua, el valor del modo debería interpolarse dentro de clase que posee mayor frecuencia. Si bien no es necesario realizar ese proceso, es interesante desestimar el valor 3.3 y considerar a partir de la distribución de frecuencias si existe una clase que posea mayor frecuencia que las demás y situar el modo dentro de la misma.

Desde luego los resultados ofrecidos pueden variar levemente dependiendo el programa utilizado y podrá complementarse con algunas funciones implementadas que permiten obtener medidas de interés como los Cuartiles, Deciles, Porcentiles o Cuantiles de diferente orden.

A continuación se presenta el resumen que brinda para las mismas muestras el programa Geogebra:

m_1 = {1,2,2,3,3,3,4,6,8,8}

Estadísticas	
n	10
Media	4
σ	2.3664
s	2.4944
Σx	40
Σx^2	216
Mín	1
Q1	2
Mediana	3
Q3	6
Máx	8

m_2 = {1.1, 2.1 ,2.2 ,3.3 ,3.3 ,3.4 ,5.0 ,6.1, 8.4,8.9}

Estadísticas	
n	10
Media	4.38
σ	2.5286
s	2.6653
Σx	43.8
Σx^2	255.78
Mín	1.1
Q1	2.2
Mediana	3.35
Q3	6.1
Máx	8.9

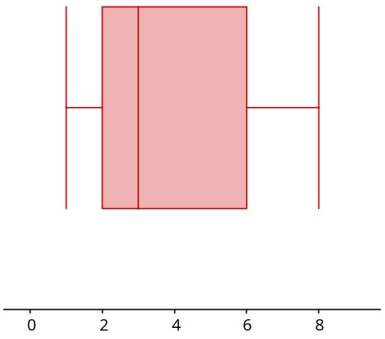
De manera similar a lo dicho con el programa LibreOffice Calc, es preciso revisar lo que se brinda y aclarar que lo que se presenta con el símbolo Sigma no es el Desvío Típico de la población y sugerimos inicialmente desestimarlo.

DIAGRAMA DE CAJA Y BIGOTES

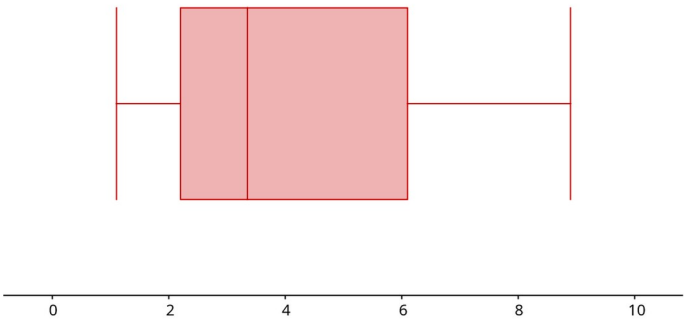
Por último se presenta un un gráfico conocido como Diagrama de Caja y Bigotes o en inglés (Box Plot), que utiliza algunas de las medidas de ubicación más importantes, que suelen aparecer en el resumen de ED (Mínimo, Q1, Q2, Q3 y Máximo), para establecer una escala de proporcionalidad a una gráfica a lo largo de un eje horizontal o vertical. A diferencia de los gráficos anteriores como el Histograma (VA continua) y el Diagrama de Barras (VA discreta), este gráfico permite la representación de cualquier VA. A su vez, permite considerar un señalamiento de los valores atípicos de la muestra (valores que muy grandes o muy pequeños según un criterio que considera el rango intercuartílico, como se verá más adelante).

A continuación se presentan lo diagramas de caja y bigotes correspondientes con las muestras anteriormente propuestas, elaborados con el programa Geogebra.

m_1 = {1,2,2,3,3,3,4,6,8,8}



m_2 = {1.1, 2.1 ,2.2 ,3.3 ,3.3 ,3.4 ,5.0 ,6.1, 8.4,8.9}



En la bibliografía recomendada encontrarán algunas profundizaciones acerca de las interpretaciones que pueden realizarse a partir de éste tipo de diagramas, pero de un modo general se establece que en función de la longitud de la caja, que dependerá del rango intercuartílico (IRQ obtenido como la distancia entre el Q3 y el Q1, es decir Q3-Q1), se podrá pensar en una mayor o menor variabilidad de los datos. Algo similar puede establecerse con respecto a los denominados vigotes. Por otra parte, analizando la distancia entre los cuartiles y la mediana, como así también comparando la longitud de los vigotes, se puede apreciar lo que sucede con la asimetría de la muestra. Tal como se ha dicho, es posible representar los denominados valores atípicos de la muestra, -si es que los hubiera-, en el diagrama de caja. Para determinar cuáles son los valores atípicos se considerará una magnitud correspondiente a una vez y media de la longitud de la caja (es decir 1.5*IRQ), y luego de ignorarán que se encuentren por debajo de esa distancia con respecto al Q1 para el vigote de la izquierda y los que se encuentren por encima de esa

distancia con respecto al Q3 para el vigote de la derecha. Para ilustrar lo anterior, revisemos el resumen ED del programa Geogebra asociado a m_1:

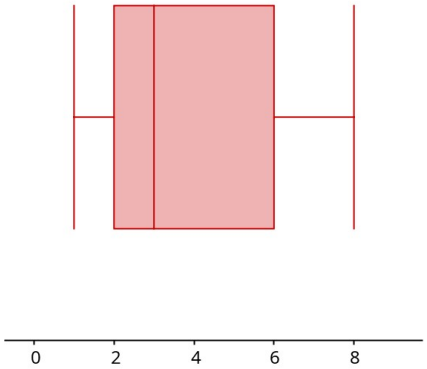
$$m_1 = \{1,2,2,3,3,3,4,6,8,8\}$$

Estadísticas	
n	10
Media	4
σ	2.3664
s	2.4944
Σx	40
Σx^2	216
Mín	1
Q1	2
Mediana	3
Q3	6
Máx	8

En base a lo dicho, el IRQ será Q3-Q1, es decir 6-2 = 4. A su vez, se considerará una magnitud que será una vez y media de esa longitud, es decir 1.5*4 = 6.

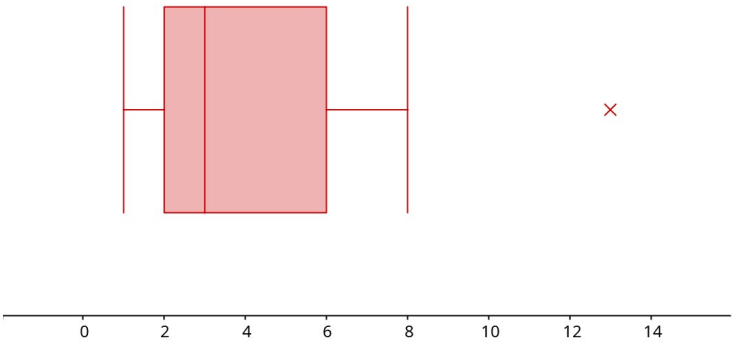
Luego para el vigote izquierdo diremos que aquellos valores que estén por debajo del Q1 en 6 unidades, es decir 2 - 6 = -4, serán atípicos. De esta forma,todos los valores menores a -4 contenidos en la muestra, serán señalados como atípicos y no podrán formar parte del vigote izquierdo. Análogamente para el vigote derecho, diremos que aquellos valores que estén por encima del Q3 en 6 unidades, es decir 6 + 6 = 12, serán considerados atípicos. De esta forma, todos los varloes mayores a 12 contenidos en la muestra, serán señalados como atípicos y no podrán formar parte del vigote de la derecha.

Tal como puede verse en el gráfico, la muestra no posee valores atípicos:



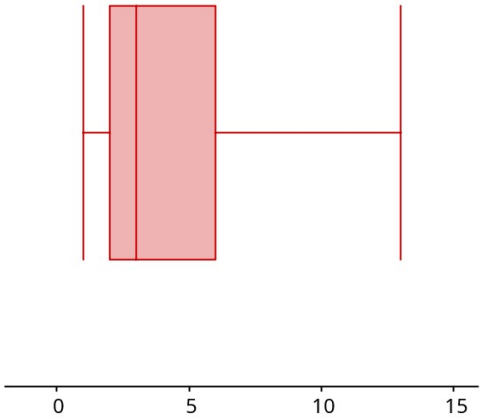
Ofreceremos a continuación una nueva muestra denotada como m_3, la cual será similar a la anterior pero reemplazando el último valor (el número 8), por el 13 (que será un valor atípico para esta muestra que podrá apreciarse en el diagrama):

$$m_3 = \{1,2,2,3,3,3,4,6,8,13\}$$



Tal como se ha dicho, debido a que el valor 13, es mayor a Q3 + 1,5 * IRQ, el mismo es señalado como valor atípico y no forma parte del vigote.

Dependiendo de la configuración del gráfico en Geogebra, es posible omitir este señalamiento:



Desde luego se presenta la versión del diagrama que no señala los valores atípicos lo que permite analizar hasta qué punto puede verse alterada la proporcionalidad del gráfico, no obstante se recomienda utilizar la versión que incorpora dichos señalamientos.

CONCLUSIONES

A partir de lo desarrollado en estas notas de introducción, se establece un recorte para el presente cursado que incluye las representaciones gráficas:

- Diagrama de barras (VA Discreta)
- Histograma (VA Continua)
- Polígono de Frecuencias (VA Continua)
- Box Plot (VA Discreta o VA Continua)

Pueden agregarse a las anteriores otras representaciones similares asociadas a las frecuencias relativas y absolutas en versiones acumulativas.

De manera analítica pueden obtenerse diferentes resúmenes que incluyen algunas medidas descriptivas de muestras que deben analizarse e interpretarse según la clasificación de la Va y es posible que alguno de los valores que brindan los programas no tengan un sentido teórico adecuado. Se espera que a partir del análisis descriptivo de los datos, se pueda obtener una mirada acerca de la distribución de las frecuencias de dicha muestra que permita imaginar una distribución de probabilidades de la población a la que pertenece y que podremos estudiar con herramientas superan a la ED.