INTRODUCCIÓN A LA ESTADISTICA APLICADA
Clase I - Primera Parte
Introducción a la Estadística Aplicada

María Cristina Martín

Departamento de Matemática
Universidad Nacional de la Pampa

24 y 26 de agosto de 2020

Introducción a la Estadística Aplicada

- Pese a los diferentes nombres que la asignatura que desarrollaremos a lo largo del cuatrimestre lleva para las diferentes carreras de ciencias experimentales, que se dictan en la FCEyN-UNLPam, este es un curso de **Introducción a la Estadistica Aplicada** en el que intentaremos dar respuesta a las siguientes preguntas:

  1. ¿Qué es la Estadistica?
  2. ¿Cómo funciona?
  3. ¿Por qué es tan necesaria en la resolución de ciertos problemas prácticos?
  4. ¿Cómo nos ayuda en esa resolución?

- Para ello, resulta fundamental, que a partir de este curso comencemos a modificar la percepción que de la Estadistica tiene la mayoría de los investigadores y experimentadores.

Introducción a la Estadística Aplicada

- Pese a los diferentes nombres que la asignatura que desarrollaremos a lo largo del cuatrimestre lleva para las diferentes carreras de ciencias experimentales, que se dictan en la FCEyN-UNLPam, este es un curso de **Introducción a la Estadistica Aplicada** en el que intentaremos dar respuesta a las siguientes preguntas:

  1. ¿Qué es la Estadistica?
  2. ¿Cómo funciona?
  3. ¿Por qué es tan necesaria en la resolución de ciertos problemas prácticos?
  4. ¿Cómo nos ayuda en esa resolución?

- Para ello, resulta fundamental, que a partir de este curso comencemos a modificar la percepción que de la Estadistica tiene la mayoría de los investigadores y experimentadores.

Introducción a la Estadística Aplicada

- Pese a los diferentes nombres que la asignatura que desarrollaremos a lo largo del cuatrimestre lleva para las diferentes carreras de ciencias experimentales, que se dictan en la FCEyN-UNLPam, este es un curso de **Introducción a la Estadistica Aplicada** en el que intentaremos dar respuesta a las siguientes preguntas:

  1. ¿Qué es la Estadistica?
  2. ¿Cómo funciona?
  3. ¿Por qué es tan necesaria en la resolución de ciertos problemas prácticos?
  4. ¿Cómo nos ayuda en esa resolución?

- Para ello, resulta fundamental, que a partir de este curso comencemos a modificar la percepción que de la Estadistica tiene la mayoría de los investigadores y experimentadores.

Introducción a la Estadística Aplicada

- Pese a los diferentes nombres que la asignatura que desarrollaremos a lo largo del cuatrimestre lleva para las diferentes carreras de ciencias experimentales, que se dictan en la FCEyN-UNLPam, este es un curso de **Introducción a la Estadistica Aplicada** en el que intentaremos dar respuesta a las siguientes preguntas:

  1. ¿Qué es la Estadistica?
  2. ¿Cómo funciona?
  3. ¿Por qué es tan necesaria en la resolución de ciertos problemas prácticos?
  4. ¿Cómo nos ayuda en esa resolución?

- Para ello, resulta fundamental, que a partir de este curso comencemos a modificar la percepción que de la Estadistica tiene la mayoría de los investigadores y experimentadores.

Introducción a la Estadística Aplicada

- Pese a los diferentes nombres que la asignatura que desarrollaremos a lo largo del cuatrimestre lleva para las diferentes carreras de ciencias experimentales, que se dictan en la FCEyN-UNLPam, este es un curso de **Introducción a la Estadistica Aplicada** en el que intentaremos dar respuesta a las siguientes preguntas:

  1. ¿Qué es la Estadistica?
  2. ¿Cómo funciona?
  3. ¿Por qué es tan necesaria en la resolución de ciertos problemas prácticos?
  4. ¿Cómo nos ayuda en esa resolución?

- Para ello, resulta fundamental, que a partir de este curso comencemos a modificar la percepción que de la Estadistica tiene la mayoría de los investigadores y experimentadores.

Introducción a la Estadística Aplicada

- Pese a los diferentes nombres que la asignatura que desarrollaremos a lo largo del cuatrimestre lleva para las diferentes carreras de ciencias experimentales, que se dictan en la FCEyN-UNLPam, este es un curso de **Introducción a la Estadistica Aplicada** en el que intentaremos dar respuesta a las siguientes preguntas:

  1. ¿Qué es la Estadistica?
  2. ¿Cómo funciona?
  3. ¿Por qué es tan necesaria en la resolución de ciertos problemas prácticos?
  4. ¿Cómo nos ayuda en esa resolución?

- Para ello, resulta fundamental, que a partir de este curso comencemos a modificar la percepción que de la Estadistica tiene la mayoría de los investigadores y experimentadores.

Introducción a la Estadística Aplicada

- ¿Cómo lo hacemos?:
  1. En los Cursos Básicos de Estadística debiéramos **enfatizar** y **trabajar** sobre la definición de objetivos, los problemas de la selección del material, diseño de la muestra, manejo de los datos, inferencia e implicancias prácticas del Modelo Estadístico.
  2. Los asesores en el área de la estadística tendríamos que llevar a cabo una labor permanente de formación, cuyo objetivo es llegar a compartir con el investigador de las diferentes áreas una **filosofía** común de trabajo que implique hacer, en forma conjunta, la planificación, el diseño y selección de la muestra, el manejo de los datos, la inferencia e, inclusive, la redacción de los informes finales.

Introducción a la Estadística Aplicada

- ¿Cómo lo hacemos?:

  1. En los Cursos Básicos de Estadística debiéramos **enfatizar** y **trabajar** sobre la definición de objetivos, los problemas de la selección del material, diseño de la muestra, manejo de los datos, inferencia e implicancias prácticas del Modelo Estadístico.
  2. Los asesores en el área de la estadística tendríamos que llevar a cabo una labor permanente de formación, cuyo objetivo es llegar a compartir con el investigador de las diferentes áreas una **filosofía** común de trabajo que implique hacer, en forma conjunta, la planificación, el diseño y selección de la muestra, el manejo de los datos, la inferencia e, inclusive, la redacción de los informes finales.

- ¿Cómo lo hacemos?:

  1. En los Cursos Básicos de Estadística debiéramos **enfatizar** y **trabajar** sobre la definición de objetivos, los problemas de la selección del material, diseño de la muestra, manejo de los datos, inferencia e implicancias prácticas del Modelo Estadístico.
  2. Los asesores en el área de la estadística tendríamos que llevar a cabo una labor permanente de formación, cuyo objetivo es llegar a compartir con el investigador de las diferentes áreas una **filosofía** común de trabajo que implique hacer, en forma conjunta, la planificación, el diseño y selección de la muestra, el manejo de los datos, la inferencia e, inclusive, la redacción de los informes finales.

Introducción a la Estadística Aplicada

- Teniendo en cuenta los dos puntos anteriores, podemos empezar a revertir la situación que hace que:
  1. los investigadores de las diversas disciplinas realicen desde el diseño hasta el análisis de sus datos, prescindiendo del asesoramiento de un profesional estadístico; y
  2. los estadísticos trabajen sin poder transmitir sus aportes más importantes a las diferentes áreas, o lo que aún peor, desarrollando los temas desde un punto de vista teórico con ejemplos ficticios.
- Según V. SONVICO (1984), un curso de estadística aplicada, en las más diversas áreas, tiene que enfatizar el pensamiento estadistico desde la discusión de los objetivos, de hipótesis, de los conceptos de errores y de independencia, etc., que son fundamentales para el procedimiento estadistico, pero que estén ligados al campo de interés de los investigadores. Esto dejará en los estudiantes e investigadores una correcta valoración de lo que la estadística puede ofrecerles, y a su vez, a que exista una relación más adecuada entre *"experimentador"* y *"estadistico"*.

Introducción a la Estadística Aplicada

- Teniendo en cuenta los dos puntos anteriores, podemos empezar a revertir la situación que hace que:
  1. los investigadores de las diversas disciplinas realicen desde el diseño hasta el análisis de sus datos, prescindiendo del asesoramiento de un profesional estadístico; y
  2. los estadísticos trabajen sin poder transmitir sus aportes más importantes a las diferentes áreas, o lo que aún peor, desarrollando los temas desde un punto de vista teórico con ejemplos ficticios.
- Según V. SONVICO (1984), un curso de estadística aplicada, en las más diversas áreas, tiene que enfatizar el pensamiento estadistico desde la discusión de los objetivos, de hipótesis, de los conceptos de errores y de independencia, etc., que son fundamentales para el procedimiento estadistico, pero que estén ligados al campo de interés de los investigadores. Esto dejará en los estudiantes e investigadores una correcta valoración de lo que la estadística puede ofrecerles, y a su vez, a que exista una relación más adecuada entre *"experimentador"* y *"estadistico"*.

Introducción a la Estadística Aplicada

- Teniendo en cuenta los dos puntos anteriores, podemos empezar a revertir la situación que hace que:
  1. los investigadores de las diversas disciplinas realicen desde el diseño hasta el análisis de sus datos, prescindiendo del asesoramiento de un profesional estadístico; y
  2. los estadísticos trabajen sin poder transmitir sus aportes más importantes a las diferentes áreas, o lo que aún peor, desarrollando los temas desde un punto de vista teórico con ejemplos ficticios.
- Según V. SONVICO (1984), un curso de estadística aplicada, en las más diversas áreas, tiene que enfatizar el pensamiento estadistico desde la discusión de los objetivos, de hipótesis, de los conceptos de errores y de independencia, etc., que son fundamentales para el procedimiento estadistico, pero que estén ligados al campo de interés de los investigadores. Esto dejará en los estudiantes e investigadores una correcta valoración de lo que la estadística puede ofrecerles, y a su vez, a que exista una relación más adecuada entre *"experimentador"* y *"estadistico"*.

Introducción a la Estadística Aplicada

- Teniendo en cuenta los dos puntos anteriores, podemos empezar a revertir la situación que hace que:
  1. los investigadores de las diversas disciplinas realicen desde el diseño hasta el análisis de sus datos, prescindiendo del asesoramiento de un profesional estadístico; y
  2. los estadísticos trabajen sin poder transmitir sus aportes más importantes a las diferentes áreas, o lo que aún peor, desarrollando los temas desde un punto de vista teórico con ejemplos ficticios.
- Según V. SONVICO (1984), un curso de estadística aplicada, en las más diversas áreas, tiene que enfatizar el pensamiento estadistico desde la discusión de los objetivos, de hipótesis, de los conceptos de errores y de independencia, etc., que son fundamentales para el procedimiento estadistico, pero que estén ligados al campo de interés de los investigadores. Esto dejará en los estudiantes e investigadores una correcta valoración de lo que la estadística puede ofrecerles, y a su vez, a que exista una relación más adecuada entre *"experimentador"* y *"estadistico"*.

Introducción a la Estadística Aplicada

- Siguiendo estas premisas, en este curso, vamos a trabajar fundamentalmente con DATOS y en su ANÁLISIS.
- Una **experiencia con datos reales** implica realizar :

  PREVISIÓN + **DECISIÓN**

- Además, cada experiencia implica **MUESTREO**.
- Una **muestra** es una determinada cantidad de unidades (ítemes, individuos, objetos, etc.) de los que se extrae información y que son retirados de un conjunto más grande, al que denomiremos **población**.
- Surgen asi, los dos primeros conceptos (y diferencias) que investigador y estadístico deben identificar perfectamente.

Introducción a la Estadística Aplicada

- Siguiendo estas premisas, en este curso, vamos a trabajar fundamentalmente con DATOS y en su ANÁLISIS.
- Una **experiencia con datos reales** implica realizar :

  PREVISIÓN $+$ **DECISIÓN**

- Además, cada experiencia implica **MUESTREO**.
- Una **muestra** es una determinada cantidad de unidades (ítemes, individuos, objetos, etc.) de los que se extrae información y que son retirados de un conjunto más grande, al que denomiremos **población**.
- Surgen asi, los dos primeros conceptos (y diferencias) que investigador y estadístico deben identificar perfectamente.

Introducción a la Estadística Aplicada

- Siguiendo estas premisas, en este curso, vamos a trabajar fundamentalmente con DATOS y en su ANÁLISIS.
- Una **experiencia con datos reales** implica realizar :

  PREVISIÓN $+$ **DECISIÓN**

- Además, cada experiencia implica **MUESTREO**.
- Una **muestra** es una determinada cantidad de unidades (ítemes, individuos, objetos, etc.) de los que se extrae información y que son retirados de un conjunto más grande, al que denomiremos **población**.
- Surgen asi, los dos primeros conceptos (y diferencias) que investigador y estadístico deben identificar perfectamente.

Introducción a la Estadística Aplicada

- Siguiendo estas premisas, en este curso, vamos a trabajar fundamentalmente con DATOS y en su ANÁLISIS.
- Una **experiencia con datos reales** implica realizar :

  PREVISIÓN $+$ **DECISIÓN**

- Además, cada experiencia implica **MUESTREO**.
- Una **muestra** es una determinada cantidad de unidades (ítemes, individuos, objetos, etc.) de los que se extrae información y que son retirados de un conjunto más grande, al que denomiremos **población**.
- Surgen asi, los dos primeros conceptos (y diferencias) que investigador y estadístico deben identificar perfectamente.

- Siguiendo estas premisas, en este curso, vamos a trabajar fundamentalmente con DATOS y en su ANÁLISIS.
- Una **experiencia con datos reales** implica realizar :

  PREVISIÓN + **DECISIÓN**

- Además, cada experiencia implica **MUESTREO**.
- Una **muestra** es una determinada cantidad de unidades (ítemes, individuos, objetos, etc.) de los que se extrae información y que son retirados de un conjunto más grande, al que denominaremos **población**.
- Surgen asi, los dos primeros conceptos (y diferencias) que investigador y estadístico deben identificar perfectamente.

Introducción a la Estadística Aplicada

- En general:

POBLACIÓN: es el conjunto que representa TODAS las medidas de interés del investigador.

MUESTRA: Es un subconjunto de medidas extraídas de la población de interés.

- Popularmente, el término MUESTRA tiene dos sentidos:
  • puede designar un **conjunto de objetos** acerca de los cuales se pretende determinar ciertas propiedades, y es denominada **"Muestra de Unidades"**,
  • pero también puede designar el **conjunto de las medidas** (valores observados de esas propiedades o característica que serán denominadas **VARIABLES**), denominado **"Muestra Estadística"**.
- Una terminologia análoga es aplicada para **"Población de Unidades"** y **"Población Estadística"**.

Introducción a la Estadística Aplicada

- En general:

POBLACIÓN: es el conjunto que representa TODAS las medidas de interés del investigador.

MUESTRA: Es un subconjunto de medidas extraídas de la población de interés.

- Popularmente, el término MUESTRA tiene dos sentidos:
  - puede designar un **conjunto de objetos** acerca de los cuales se pretende determinar ciertas propiedades, y es denominada **"Muestra de Unidades"**,
  - pero también puede designar el **conjunto de las medidas** (valores observados de esas propiedades o característica que serán denominadas **VARIABLES**), denominado **"Muestra Estadística"**.
- Una terminologia análoga es aplicada para **"Población de Unidades"** y **"Población Estadística"**.

Introducción a la Estadística Aplicada

- En general:

POBLACIÓN: es el conjunto que representa TODAS las medidas de interés del investigador.

MUESTRA: Es un subconjunto de medidas extraídas de la población de interés.

- Popularmente, el término MUESTRA tiene dos sentidos:
  - puede designar un **conjunto de objetos** acerca de los cuales se pretende determinar ciertas propiedades, y es denominada **"Muestra de Unidades"**,
  - pero también puede designar el **conjunto de las medidas** (valores observados de esas propiedades o característica que serán denominadas **VARIABLES**), denominado **"Muestra Estadística"**.
- Una terminologia análoga es aplicada para **"Población de Unidades"** y **"Población Estadística"**.

- En general:

POBLACIÓN: es el conjunto que representa TODAS las medidas de interés del investigador.

MUESTRA: Es un subconjunto de medidas extraídas de la población de interés.

- Popularmente, el término MUESTRA tiene dos sentidos:
  • puede designar un **conjunto de objetos** acerca de los cuales se pretende determinar ciertas propiedades, y es denominada **"Muestra de Unidades"**,
  • pero también puede designar el **conjunto de las medidas** (valores observados de esas propiedades o característica que serán denominadas **VARIABLES**), denominado **"Muestra Estadística"**.
- Una terminologia análoga es aplicada para **"Población de Unidades"** y **"Población Estadística"**.

- En general:

POBLACIÓN: es el conjunto que representa TODAS las medidas de interés del investigador.

MUESTRA: Es un subconjunto de medidas extraídas de la población de interés.

- Popularmente, el término MUESTRA tiene dos sentidos:
  • puede designar un **conjunto de objetos** acerca de los cuales se pretende determinar ciertas propiedades, y es denominada **"Muestra de Unidades"**,
  • pero también puede designar el **conjunto de las medidas** (valores observados de esas propiedades o característica que serán denominadas **VARIABLES**), denominado **"Muestra Estadística"**.
- Una terminologia análoga es aplicada para **"Población de Unidades"** y **"Población Estadística"**.

Introducción a la Estadística Aplicada

- En el estudio de Métodos de Análisis de una Muestra es importante distinguir entre los **objetivos medidos** y las **medidas** propiamente dichas. Así:
- • el objetivo medido es denominado **Unidad Experimental**;
  • **Muestra** (o Muestra Estadística) es el conjunto de medidas (valores de una variable) efectuadas sobre las unidades experimentales (elementos de la muestra o Muestra de Unidades).
- **Observación:**
  ¿Qué es más importante la Población o la Muestra?
  SIEMPRE es más importante la POBLACIÓN, pero...
- Es imposible entrevistar a todos los habitantes de un pais (CENSO), de modo que, es necesario prever su comportamiento, frente a algo, basados en la información de una **muestra** representativa de esa población.
- Similarmente, es imposible vacunar a toda la población contra el COVID−19, a efectos de experimentar la eficacia de una droga.

Introducción a la Estadística Aplicada

- En el estudio de Métodos de Análisis de una Muestra es importante distinguir entre los **objetivos medidos** y las **medidas** propiamente dichas. Así:
- • el objetivo medido es denominado **Unidad Experimental**;
  • **Muestra** (o Muestra Estadística) es el conjunto de medidas (valores de una variable) efectuadas sobre las unidades experimentales (elementos de la muestra o Muestra de Unidades).
- **Observación:**
  ¿Qué es más importante la Población o la Muestra?
  SIEMPRE es más importante la POBLACIÓN, pero...
- Es imposible entrevistar a todos los habitantes de un pais (CENSO), de modo que, es necesario prever su comportamiento, frente a algo, basados en la información de una **muestra** representativa de esa población.
- Similarmente, es imposible vacunar a toda la población contra el COVID−19, a efectos de experimentar la eficacia de una droga.

Introducción a la Estadística Aplicada

- En el estudio de Métodos de Análisis de una Muestra es importante distinguir entre los **objetivos medidos** y las **medidas** propiamente dichas. Así:
- • el objetivo medido es denominado **Unidad Experimental**;
  • **Muestra** (o Muestra Estadística) es el conjunto de medidas (valores de una variable) efectuadas sobre las unidades experimentales (elementos de la muestra o Muestra de Unidades).
- **Observación:**
  ¿Qué es más importante la Población o la Muestra?
  SIEMPRE es más importante la POBLACIÓN, pero...
- Es imposible entrevistar a todos los habitantes de un pais (CENSO), de modo que, es necesario prever su comportamiento, frente a algo, basados en la información de una **muestra** representativa de esa población.
- Similarmente, es imposible vacunar a toda la población contra el COVID−19, a efectos de experimentar la eficacia de una droga.

Introducción a la Estadística Aplicada

- En el estudio de Métodos de Análisis de una Muestra es importante distinguir entre los **objetivos medidos** y las **medidas** propiamente dichas. Así:
- • el objetivo medido es denominado **Unidad Experimental**;
  • **Muestra** (o Muestra Estadística) es el conjunto de medidas (valores de una variable) efectuadas sobre las unidades experimentales (elementos de la muestra o Muestra de Unidades).
- **Observación:**
  ¿Qué es más importante la Población o la Muestra?
  SIEMPRE es más importante la POBLACIÓN, pero...
- Es imposible entrevistar a todos los habitantes de un pais (CENSO), de modo que, es necesario prever su comportamiento, frente a algo, basados en la información de una **muestra** representativa de esa población.
- Similarmente, es imposible vacunar a toda la población contra el COVID−19, a efectos de experimentar la eficacia de una droga.

Introducción a la Estadística Aplicada

- En el estudio de Métodos de Análisis de una Muestra es importante distinguir entre los **objetivos medidos** y las **medidas** propiamente dichas. Así:
- • el objetivo medido es denominado **Unidad Experimental**;
  • **Muestra** (o Muestra Estadística) es el conjunto de medidas (valores de una variable) efectuadas sobre las unidades experimentales (elementos de la muestra o Muestra de Unidades).
- **Observación:**
  ¿Qué es más importante la Población o la Muestra?
  SIEMPRE es más importante la POBLACIÓN, pero...
- Es imposible entrevistar a todos los habitantes de un pais (CENSO), de modo que, es necesario prever su comportamiento, frente a algo, basados en la información de una **muestra** representativa de esa población.
- Similarmente, es imposible vacunar a toda la población contra el COVID−19, a efectos de experimentar la eficacia de una droga.

Introducción a la Estadística Aplicada

- Entonces:

El **Objetivo** de la Estadística es permitir INFERENCIA (Previsión + Decisión) sobre una población, tomando como base las informaciones de la muestra.

- ¿Cómo conseguir ese objetivo?
- A lo largo de todo el curso, veremos que todo problema estadistico contiene cinco elementos:
  1. Una clara definición de OBJETIVOS de la experiencia y de la POBLACIÓN asociada;
  2. La decisión de cómo elegir la muestra, llamado PROCEDIMIENTO MUESTRAL
  3. La recolección y el Análisis Exploratorio de los Datos Muestrales;
  4. Realizar INFERENCIA sobre la población;
  5. Dar una MEDIDA de CONFIABILIDAD de la Inferencia realizada.

Introducción a la Estadística Aplicada

- Entonces:

El **Objetivo** de la Estadística es permitir INFERENCIA (Previsión + Decisión) sobre una población, tomando como base las informaciones de la muestra.

- ¿Cómo conseguir ese objetivo?
- A lo largo de todo el curso, veremos que todo problema estadistico contiene cinco elementos:
  1. Una clara definición de OBJETIVOS de la experiencia y de la POBLACIÓN asociada;
  2. La decisión de cómo elegir la muestra, llamado PROCEDIMIENTO MUESTRAL
  3. La recolección y el Análisis Exploratorio de los Datos Muestrales;
  4. Realizar INFERENCIA sobre la población;
  5. Dar una MEDIDA de CONFIABILIDAD de la Inferencia realizada.

Introducción a la Estadística Aplicada

- Entonces:

El **Objetivo** de la Estadística es permitir INFERENCIA (Previsión + Decisión) sobre una población, tomando como base las informaciones de la muestra.

- ¿Cómo conseguir ese objetivo?
- A lo largo de todo el curso, veremos que todo problema estadistico contiene cinco elementos:
  1. Una clara definición de OBJETIVOS de la experiencia y de la POBLACIÓN asociada;
  2. La decisión de cómo elegir la muestra, llamado PROCEDIMIENTO MUESTRAL
  3. La recolección y el Análisis Exploratorio de los Datos Muestrales;
  4. Realizar INFERENCIA sobre la población;
  5. Dar una MEDIDA de CONFIABILIDAD de la Inferencia realizada.

Introducción a la Estadística Aplicada

- Entonces:

El **Objetivo** de la Estadística es permitir INFERENCIA (Previsión + Decisión) sobre una población, tomando como base las informaciones de la muestra.

- ¿Cómo conseguir ese objetivo?
- A lo largo de todo el curso, veremos que todo problema estadistico contiene cinco elementos:
  1. Una clara definición de OBJETIVOS de la experiencia y de la POBLACIÓN asociada;
  2. La decisión de cómo elegir la muestra, llamado PROCEDIMIENTO MUESTRAL
  3. La recolección y el Análisis Exploratorio de los Datos Muestrales;
  4. Realizar INFERENCIA sobre la población;
  5. Dar una MEDIDA de CONFIABILIDAD de la Inferencia realizada.

- Entonces:

El **Objetivo** de la Estadística es permitir INFERENCIA (Previsión + Decisión) sobre una población, tomando como base las informaciones de la muestra.

- ¿Cómo conseguir ese objetivo?
- A lo largo de todo el curso, veremos que todo problema estadistico contiene cinco elementos:
  1. Una clara definición de OBJETIVOS de la experiencia y de la POBLACIÓN asociada;
  2. La decisión de cómo elegir la muestra, llamado PROCEDIMIENTO MUESTRAL
  3. La recolección y el Análisis Exploratorio de los Datos Muestrales;
  4. Realizar INFERENCIA sobre la población;
  5. Dar una MEDIDA de CONFIABILIDAD de la Inferencia realizada.

Introducción a la Estadística Aplicada

- Entonces:

El **Objetivo** de la Estadística es permitir INFERENCIA (Previsión + Decisión) sobre una población, tomando como base las informaciones de la muestra.

- ¿Cómo conseguir ese objetivo?
- A lo largo de todo el curso, veremos que todo problema estadistico contiene cinco elementos:
  1. Una clara definición de OBJETIVOS de la experiencia y de la POBLACIÓN asociada;
  2. La decisión de cómo elegir la muestra, llamado PROCEDIMIENTO MUESTRAL
  3. La recolección y el Análisis Exploratorio de los Datos Muestrales;
  4. Realizar INFERENCIA sobre la población;
  5. Dar una MEDIDA de CONFIABILIDAD de la Inferencia realizada.

- Entonces:

El **Objetivo** de la Estadística es permitir INFERENCIA (Previsión + Decisión) sobre una población, tomando como base las informaciones de la muestra.

- ¿Cómo conseguir ese objetivo?
- A lo largo de todo el curso, veremos que todo problema estadistico contiene cinco elementos:
  1. Una clara definición de OBJETIVOS de la experiencia y de la POBLACIÓN asociada;
  2. La decisión de cómo elegir la muestra, llamado PROCEDIMIENTO MUESTRAL
  3. La recolección y el Análisis Exploratorio de los Datos Muestrales;
  4. Realizar INFERENCIA sobre la población;
  5. Dar una MEDIDA de CONFIABILIDAD de la Inferencia realizada.

- Entonces:

El **Objetivo** de la Estadística es permitir INFERENCIA (Previsión + Decisión) sobre una población, tomando como base las informaciones de la muestra.

- ¿Cómo conseguir ese objetivo?
- A lo largo de todo el curso, veremos que todo problema estadistico contiene cinco elementos:
  1. Una clara definición de OBJETIVOS de la experiencia y de la POBLACIÓN asociada;
  2. La decisión de cómo elegir la muestra, llamado PROCEDIMIENTO MUESTRAL
  3. La recolección y el Análisis Exploratorio de los Datos Muestrales;
  4. Realizar INFERENCIA sobre la población;
  5. Dar una MEDIDA de CONFIABILIDAD de la Inferencia realizada.

- Entonces:

El **Objetivo** de la Estadística es permitir INFERENCIA (Previsión + Decisión) sobre una población, tomando como base las informaciones de la muestra.

- ¿Cómo conseguir ese objetivo?
- A lo largo de todo el curso, veremos que todo problema estadistico contiene cinco elementos:
  1. Una clara definición de OBJETIVOS de la experiencia y de la POBLACIÓN asociada;
  2. La decisión de cómo elegir la muestra, llamado PROCEDIMIENTO MUESTRAL
  3. La recolección y el Análisis Exploratorio de los Datos Muestrales;
  4. Realizar INFERENCIA sobre la población;
  5. Dar una MEDIDA de CONFIABILIDAD de la Inferencia realizada.

Introducción a la Estadística Aplicada

- Los cinco elementos enumerados anteriormente, llevan a dividir los procedimientos de análisis en DOS:

ESTADÍSTICA DESCRIPTIVA o ANÁLISIS EXPLORATORIO DE DATOS: Utiliza métodos numéricos y gráficos para resumir, o para representar la información, o para descubrir modelos en un conjunto de datos dado.

ESTADÍSTICA INFERENCIAL: Utiliza los datos muestrales para hacer estimaciones o tomar decisiones o hacer previsiones sobre un conjunto más grande de datos (la población de la que se eligió la muestra).

Introducción a la Estadística Aplicada

- Los cinco elementos enumerados anteriormente, llevan a dividir los procedimientos de análisis en DOS:

ESTADÍSTICA DESCRIPTIVA o ANÁLISIS EXPLORATORIO DE DATOS: Utiliza métodos numéricos y gráficos para resumir, o para representar la información, o para descubrir modelos en un conjunto de datos dado.

ESTADÍSTICA INFERENCIAL: Utiliza los datos muestrales para hacer estimaciones o tomar decisiones o hacer previsiones sobre un conjunto más grande de datos (la población de la que se eligió la muestra).

Introducción a la Estadística Aplicada

- Los cinco elementos enumerados anteriormente, llevan a dividir los procedimientos de análisis en DOS:

ESTADÍSTICA DESCRIPTIVA o ANÁLISIS EXPLORATORIO DE DATOS: Utiliza métodos numéricos y gráficos para resumir, o para representar la información, o para descubrir modelos en un conjunto de datos dado.

-

ESTADÍSTICA INFERENCIAL: Utiliza los datos muestrales para hacer estimaciones o tomar decisiones o hacer previsiones sobre un conjunto más grande de datos (la población de la que se eligió la muestra).

-

Introducción a la Estadística Aplicada

- Los cinco elementos enumerados anteriormente, llevan a dividir los procedimientos de análisis en DOS:

ESTADÍSTICA DESCRIPTIVA o ANÁLISIS EXPLORATORIO DE DATOS: Utiliza métodos numéricos y gráficos para resumir, o para representar la información, o para descubrir modelos en un conjunto de datos dado.

ESTADÍSTICA INFERENCIAL: Utiliza los datos muestrales para hacer estimaciones o tomar decisiones o hacer previsiones sobre un conjunto más grande de datos (la población de la que se eligió la muestra).

Introducción a la Estadística Aplicada

- Los cinco elementos enumerados anteriormente, llevan a dividir los procedimientos de análisis en DOS:

ESTADÍSTICA DESCRIPTIVA o ANÁLISIS EXPLORATORIO DE DATOS: Utiliza métodos numéricos y gráficos para resumir, o para representar la información, o para descubrir modelos en un conjunto de datos dado.

ESTADÍSTICA INFERENCIAL: Utiliza los datos muestrales para hacer estimaciones o tomar decisiones o hacer previsiones sobre un conjunto más grande de datos (la población de la que se eligió la muestra).

Análisis Exploratorio de Datos Unidimensionales

- Vamos a comenzar el **Análisis Estadistico Descriptivo** o **Exploratorio** de una muestra, utilizando los datos proporcionados por la Dirección Provincial del Agua de La Pampa, con el objetivo de dar respuesta a la pregunta ¿es posible el asentamiento de seres vivos en la zona de la Laguna La Colorado Grande?.

- **Ejemplo:** Los datos, suministrado por la Dirección Provincial del Agua, de la provincia de La Pampa, corresponden a 111 pozos y/o perforaciones (de la Hoja Laguna Colorada Grande -HLCG-, sudeste de La Pampa) y se han seleccionado variables a Cloruros, Sulfatos, Calcio, Magnesio, Sodio, Potasio, Fluor, Carbonato y Arsénico, medidos en mg/l. (p.p.m.).

Análisis Exploratorio de Datos Unidimensionales

- Vamos a comenzar el **Análisis Estadistico Descriptivo** o **Exploratorio** de una muestra, utilizando los datos proporcionados por la Dirección Provincial del Agua de La Pampa, con el objetivo de dar respuesta a la pregunta ¿es posible el asentamiento de seres vivos en la zona de la Laguna La Colorado Grande?.

- **Ejemplo:** Los datos, suministrado por la Dirección Provincial del Agua, de la provincia de La Pampa, corresponden a 111 pozos y/o perforaciones (de la Hoja Laguna Colorada Grande -HLCG-, sudeste de La Pampa) y se han seleccionado variables a Cloruros, Sulfatos, Calcio, Magnesio, Sodio, Potasio, Fluor, Carbonato y Arsénico, medidos en mg/l. (p.p.m.).

Análisis Exploratorio de Datos Unidimensionales

Pozo	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
1	136	342	380	29	12	5	368	6	0.1
2	100	279	348	24	19	4	322	5.5	0.05
3	168	461	304	40	17	3.4	396	8	0.05
4	224	457	460	24	9.7	7	525	11	0.1
5	172	334	560	16	23	17	441	8	0.2
6	136	348	324	27	21	3.4	331	8	0.1
7	188	444	276	35	44	2.2	368	8.6	0
8	100	250	320	21	16	5	285	7	0.1
9	88	291	344	26	21	5	257	6	0.1
10	168	232	348	19	32	1.8	267	6.6	0.03
11	524	698	880	26	60	19	975	11	0.3
12	676	944	788	13	17	17	1141	10	0.3
13	136	205	376	22	23	7	322	5.5	0.1
14	176	217	280	34	31	1.8	251	5	0.05
15	612	480	260	75	112	2.3	589	10	0.03
16	892	1598	344	62	99	5	1086	9.5	0.05
17	944	1919	236	77	208	1.8	1012	16	0
18	1124	1331	228	115	173	1.8	865	12	0.015
19	256	511	308	42	24	2.2	414	8	0.05
20	220	297	344	24	17	3.8	370	8	0.05
21	112	47	356	19	15	7	193	4	0.05
22	140	245	388	29	23	5	216	5	0.1
23	132	299	356	27	14	5	202	5	0.03
24	336	592	352	37	32	7	359	8	0.03
25	176	365	372	29	21	5	285	8	0
26	276	572	324	38	30	3	331	10	0.015
27	76	144	332	18	12	4	165	6	0.03
28	284	625	448	26	23	9	386	7	0.15
29	460	591	348	45	37	3	368	8	0.1
30	1448	1553	152	242	137	0.6	918	34	0.1
108	790	2183	100	408	163	0.6	923	75	0.100
109	1780	1722	272	124	231	1.8	1619	70	0.050
110	2140	1600	204	120	313	2.2	1987	96	0.050
111	890	1593	220	156	170	1.6	938	30	0.030

Análisis Exploratorio de Datos Unidimensionales

- En un primer paso, debemos distinguir:

1. **Unidad Experimental**: 1 pozo (o una muestra de agua de 1-pozo) de la HLCG.
2. **Muestra de Unidades**: 111 pozos de la HLCG.
3. **Población de Unidades**: TODOS los pozos de la HLCG.

4. **Variable**: *"Concentraciones de Ca de 1-pozo de la HLCG"*
5. **Muestra Estadística**: Concentraciones de Ca de 111-pozos de la HLCG.
6. **Población Estadística**: Concentraciones de Ca de los pozos de la HLCG.

Análisis Exploratorio de Datos Unidimensionales

- En un primer paso, debemos distinguir:

1. **Unidad Experimental**: 1 pozo (o una muestra de agua de 1-pozo) de la HLCG.
2. **Muestra de Unidades**: 111 pozos de la HLCG.
3. **Población de Unidades**: TODOS los pozos de la HLCG.

4. **Variable**: *"Concentraciones de Ca de 1-pozo de la HLCG"*
5. **Muestra Estadística**: Concentraciones de Ca de 111-pozos de la HLCG.
6. **Población Estadística**: Concentraciones de Ca de los pozos de la HLCG.

Análisis Exploratorio de Datos Unidimensionales

- En un primer paso, debemos distinguir:

1. **Unidad Experimental**: 1 pozo (o una muestra de agua de 1-pozo) de la HLCG.
2. **Muestra de Unidades**: 111 pozos de la HLCG.
3. **Población de Unidades**: TODOS los pozos de la HLCG.

4. **Variable**: *"Concentraciones de Ca de 1-pozo de la HLCG"*
5. **Muestra Estadística**: Concentraciones de Ca de 111-pozos de la HLCG.
6. **Población Estadística**: Concentraciones de Ca de los pozos de la HLCG.

Análisis Exploratorio de Datos Unidimensionales

- En un primer paso, debemos distinguir:

1. **Unidad Experimental**: 1 pozo (o una muestra de agua de 1-pozo) de la HLCG.
2. **Muestra de Unidades**: 111 pozos de la HLCG.
3. **Población de Unidades**: TODOS los pozos de la HLCG.

4. **Variable**: *"Concentraciones de Ca de 1-pozo de la HLCG"*
5. **Muestra Estadística**: Concentraciones de Ca de 111-pozos de la HLCG.
6. **Población Estadística**: Concentraciones de Ca de los pozos de la HLCG.

Análisis Exploratorio de Datos Unidimensionales

- En un primer paso, debemos distinguir:

1. **Unidad Experimental**: 1 pozo (o una muestra de agua de 1-pozo) de la HLCG.
2. **Muestra de Unidades**: 111 pozos de la HLCG.
3. **Población de Unidades**: TODOS los pozos de la HLCG.

4. **Variable**: *"Concentraciones de Ca de 1-pozo de la HLCG"*
5. **Muestra Estadística**: Concentraciones de Ca de 111-pozos de la HLCG.
6. **Población Estadística**: Concentraciones de Ca de los pozos de la HLCG.

Análisis Exploratorio de Datos Unidimensionales

- En un primer paso, debemos distinguir:

1. **Unidad Experimental**: 1 pozo (o una muestra de agua de 1-pozo) de la HLCG.
2. **Muestra de Unidades**: 111 pozos de la HLCG.
3. **Población de Unidades**: TODOS los pozos de la HLCG.

4. **Variable**: *"Concentraciones de Ca de 1-pozo de la HLCG"*
5. **Muestra Estadística**: Concentraciones de Ca de 111-pozos de la HLCG.
6. **Población Estadística**: Concentraciones de Ca de los pozos de la HLCG.

Análisis Exploratorio de Datos Unidimensionales

- En un primer paso, debemos distinguir:

  1. **Unidad Experimental**: 1 pozo (o una muestra de agua de 1-pozo) de la HLCG.
  2. **Muestra de Unidades**: 111 pozos de la HLCG.
  3. **Población de Unidades**: TODOS los pozos de la HLCG.

  4. **Variable**: *"Concentraciones de Ca de 1-pozo de la HLCG"*
  5. **Muestra Estadística**: Concentraciones de Ca de 111-pozos de la HLCG.
  6. **Población Estadística**: Concentraciones de Ca de los pozos de la HLCG.

Análisis Exploratorio de Datos Unidimensionales

- A seguir, debemos identificar el tipo de variable medida.

$$\text{Variables} \begin{cases} \text{Cualitativas} \begin{cases} \text{Nominal} & \text{(Ej: Sexo)} \\ \text{Ordinal} & \text{(Ej: Educación)} \end{cases} \\ \text{Cuantitativas} \begin{cases} \text{Discretas} & \text{(Ej: N}^\circ \text{ de hijos)} \\ \text{Continuas} & \text{(Ej: Altura)} \end{cases} \end{cases}$$

- **Observación 1:** Las variables de tipo cuantitativo son las más importante en la estadística experimental.
- **Observación 2:** Las variables cuantitativas discretas provienen de un **conteo**, mientras que las continuas provienen de una **medición**.
- Para cada tipo de variable exiten técnicas apropiadas para resumir la información.

Análisis Exploratorio de Datos Unidimensionales

- A seguir, debemos identificar el tipo de variable medida.

$$\text{Variables} \begin{cases} \text{Cualitativas} \begin{cases} \text{Nominal} & \text{(Ej: Sexo)} \\ \text{Ordinal} & \text{(Ej: Educación)} \end{cases} \\ \text{Cuantitativas} \begin{cases} \text{Discretas} & \text{(Ej: N° de hijos)} \\ \text{Continuas} & \text{(Ej: Altura)} \end{cases} \end{cases}$$

- **Observación 1:** Las variables de tipo cuantitativo son las más importante en la estadística experimental.
- **Observación 2:** Las variables cuantitativas discretas provienen de un **conteo**, mientras que las continuas provienen de una **medición**.
- Para cada tipo de variable exiten técnicas apropiadas para resumir la información.

Análisis Exploratorio de Datos Unidimensionales

- A seguir, debemos identificar el tipo de variable medida.

$$\text{Variables} \begin{cases} \text{Cualitativas} \begin{cases} \text{Nominal} & \text{(Ej: Sexo)} \\ \text{Ordinal} & \text{(Ej: Educación)} \end{cases} \\ \text{Cuantitativas} \begin{cases} \text{Discretas} & \text{(Ej: N}^\circ \text{ de hijos)} \\ \text{Continuas} & \text{(Ej: Altura)} \end{cases} \end{cases}$$

- **Observación 1:** Las variables de tipo cuantitativo son las más importante en la estadística experimental.
- **Observación 2:** Las variables cuantitativas discretas provienen de un **conteo**, mientras que las continuas provienen de una **medición**.
- Para cada tipo de variable exiten técnicas apropiadas para resumir la información.

Análisis Exploratorio de Datos Unidimensionales

- A seguir, debemos identificar el tipo de variable medida.

$$\text{Variables} \begin{cases} \text{Cualitativas} \begin{cases} \text{Nominal} & \text{(Ej: Sexo)} \\ \text{Ordinal} & \text{(Ej: Educación)} \end{cases} \\ \text{Cuantitativas} \begin{cases} \text{Discretas} & \text{(Ej: N}^\circ \text{ de hijos)} \\ \text{Continuas} & \text{(Ej: Altura)} \end{cases} \end{cases}$$

- **Observación 1:** Las variables de tipo cuantitativo son las más importante en la estadística experimental.
- **Observación 2:** Las variables cuantitativas discretas provienen de un **conteo**, mientras que las continuas provienen de una **medición**.
- Para cada tipo de variable exiten técnicas apropiadas para resumir la información.

- **Distribución de Frecuencias para Variables Cualitativas**

- En el ejemplo de la Laguna Colorada Grande, todas las variables son de tipo cuantitativo continuas.
- Sin embargo, podemos "**categorizar**´´ cualquiera de ellas y transformarla en una variable de tipo cualitativo.
- Por ejemplo, si consideramos la variable Arsénico (As) categorizada de la siguiente manera:

$$\text{As} \begin{cases} \text{Negativo} & = \text{Neg.} \longrightarrow \text{si toma el valor } 0 \\ \text{Leves Vestigios} & = \text{L.V.} \longrightarrow \text{si toma valores en } (0; 0{,}05] \\ \text{Vestigios} & = \text{V.} \longrightarrow \text{si toma valores } > 0{,}05 \end{cases}$$

- **Distribución de Frecuencias para Variables Cualitativas**
- En el ejemplo de la Laguna Colorada Grande, todas las variables son de tipo cuantitativo continuas.
- Sin embargo, podemos "**categorizar**´´ cualquiera de ellas y transformarla en una variable de tipo cualitativo.
- Por ejemplo, si consideramos la variable Arsénico (As) categorizada de la siguiente manera:

$$\text{As} \begin{cases} \text{Negativo} & = \text{Neg.} \longrightarrow \text{si toma el valor } 0 \\ \text{Leves Vestigios} & = \text{L.V.} \longrightarrow \text{si toma valores en } (0; 0{,}05] \\ \text{Vestigios} & = \text{V.} \longrightarrow \text{si toma valores } > 0{,}05 \end{cases}$$

- **Distribución de Frecuencias para Variables Cualitativas**

- En el ejemplo de la Laguna Colorada Grande, todas las variables son de tipo cuantitativo continuas.
- Sin embargo, podemos **"categorizar´´** cualquiera de ellas y transformarla en una variable de tipo cualitativo.
- Por ejemplo, si consideramos la variable Arsénico (As) categorizada de la siguiente manera:

$$\text{As} \begin{cases} \text{Negativo} & = \text{Neg.} \longrightarrow \text{si toma el valor } 0 \\ \text{Leves Vestigios} & = \text{L.V.} \longrightarrow \text{si toma valores en } (0; 0{,}05] \\ \text{Vestigios} & = \text{V.} \longrightarrow \text{si toma valores } > 0{,}05 \end{cases}$$

- **Distribución de Frecuencias para Variables Cualitativas**

- En el ejemplo de la Laguna Colorada Grande, todas las variables son de tipo cuantitativo continuas.
- Sin embargo, podemos "**categorizar**´´ cualquiera de ellas y transformarla en una variable de tipo cualitativo.
- Por ejemplo, si consideramos la variable Arsénico (As) categorizada de la siguiente manera:

$$\text{As} \begin{cases} \text{Negativo} & = \text{Neg.} \longrightarrow \text{si toma el valor } 0 \\ \text{Leves Vestigios} & = \text{L.V.} \longrightarrow \text{si toma valores en } (0; 0{,}05] \\ \text{Vestigios} & = \text{V.} \longrightarrow \text{si toma valores } > 0{,}05 \end{cases}$$

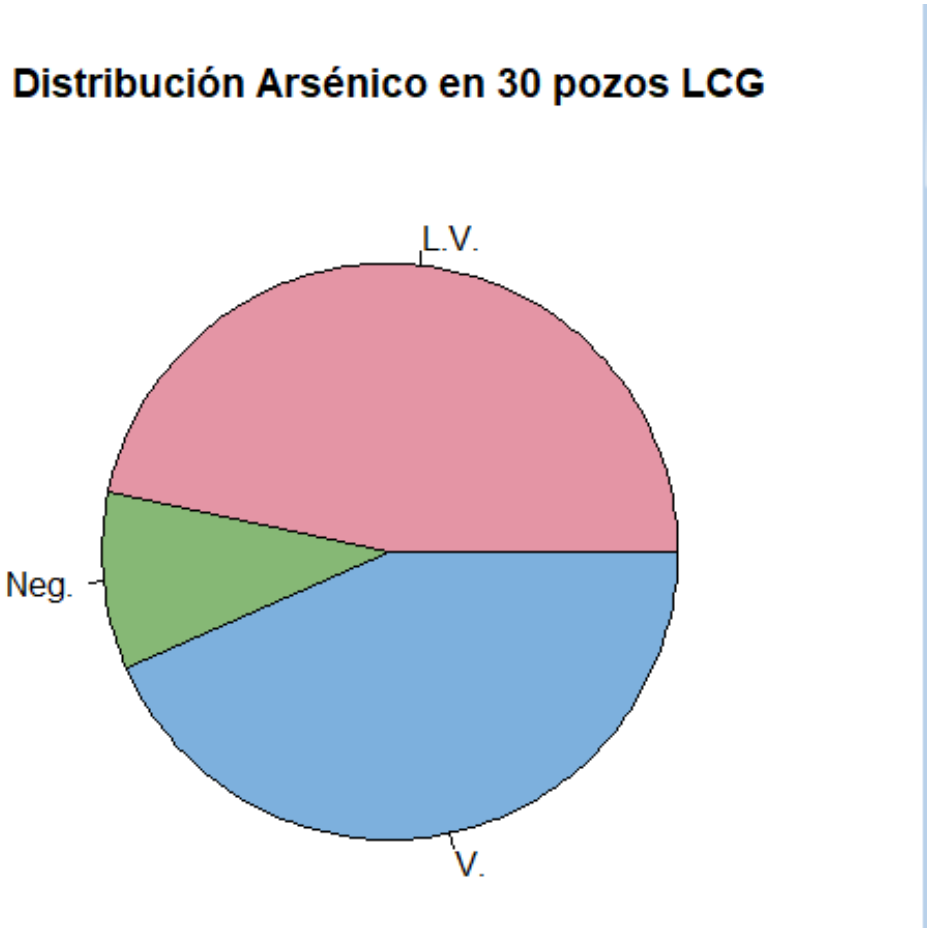
- Una tabla resumen de los datos observados para una muestra compuesta por las mediciones de los primeros 30 pozos es:

As	Frecuencia Observada: n_i	Frecuencia Relativa: f_i	Porcentaje (%)	Porcentaje Acumulado
Neg.	3	$\frac{3}{30} = 0,10$	10 %	10 %
L.V.	14	$\frac{14}{30} = 0,47$	47 %	57 %
V.	13	$\frac{13}{30} = 0,43$	43 %	100 %
Total	30	1, 00	100 %	$--$

Cuadro: Distribución de Frecuencias del Arsénico en 30 pozos de la Laguna Colorada Grande

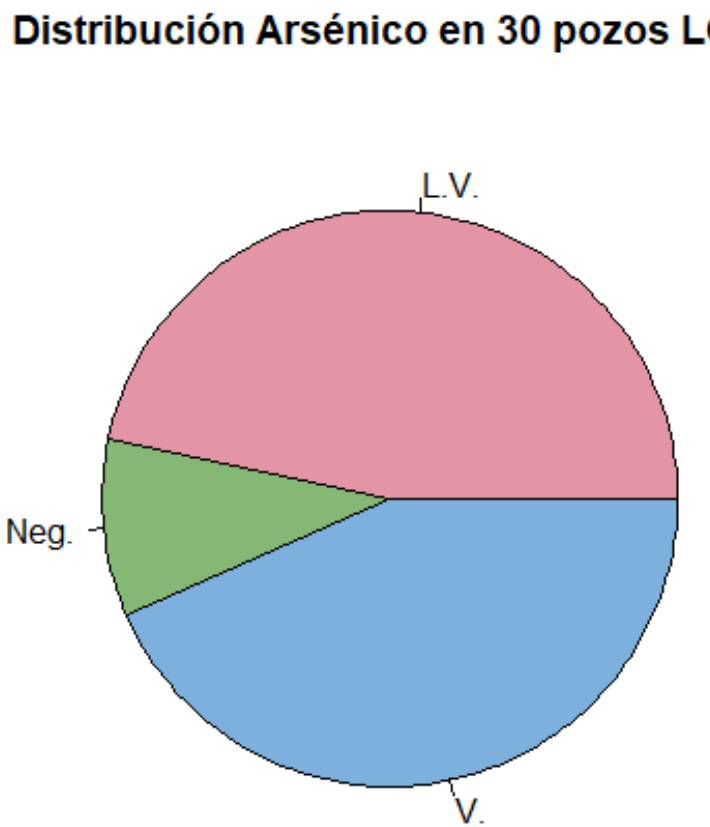
Análisis Exploratorio de Datos Unidimensionales

- La representación gráfica apropiada para Variables Cualitativas es conocida como *"Gráfico de Sectores"* o más conmunmente *"Pizza"* o "Torta".
- Para la variable Arsénico es:



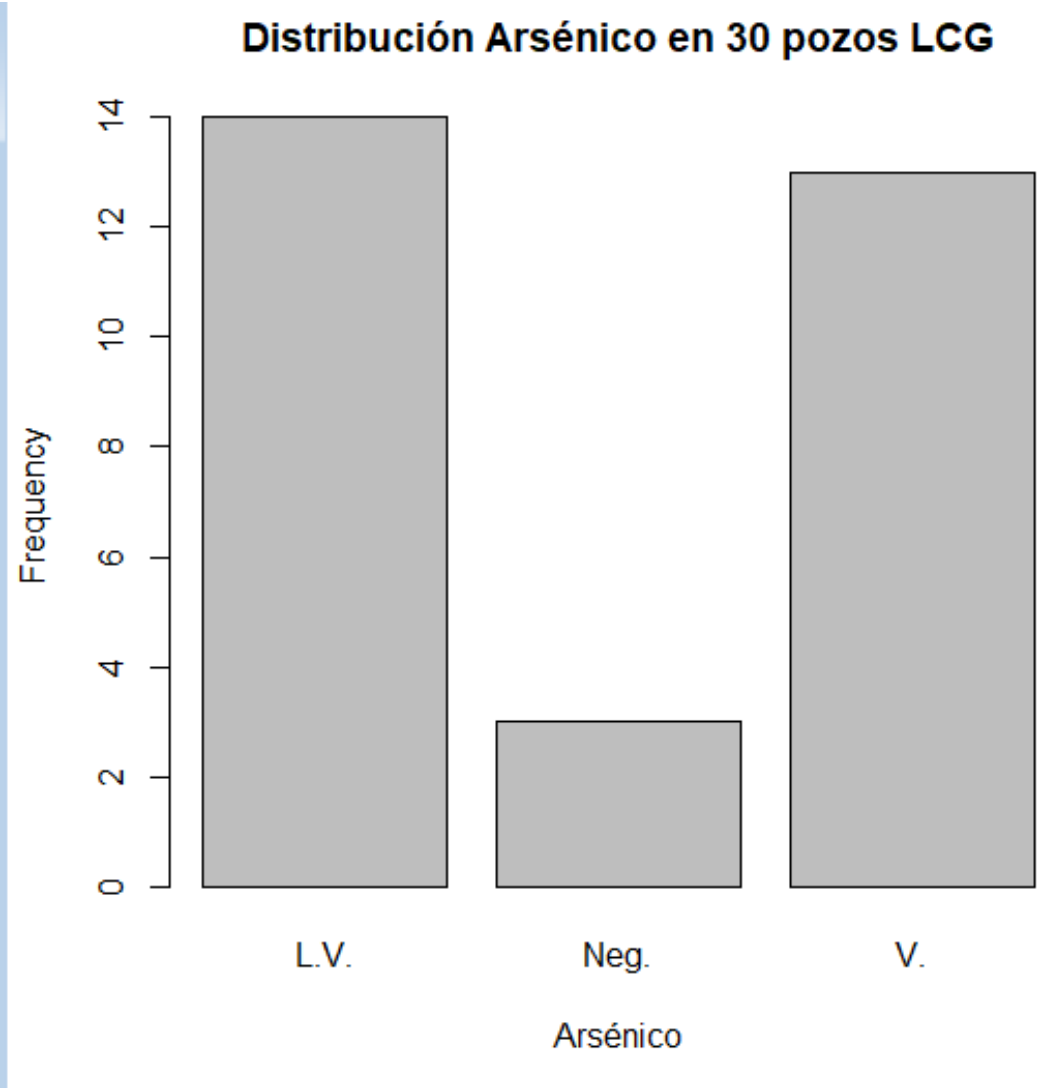
Análisis Exploratorio de Datos Unidimensionales

- La representación gráfica apropiada para Variables Cualitativas es conocida como *"Gráfico de Sectores"* o más conmunmente *"Pizza"* o "Torta".
- Para la variable Arsénico es:



Análisis Exploratorio de Datos Unidimensionales

- Otro gráfico que suele usarse para este tipo de variables es el *"Diagrama de Barras"*. Sin embargo, lo considero menos apropiado.



- **Distribución de Frecuencias para Variables Cuantitativas**

- Consideremos las mediciones de las concentraciones del ***"Calcio"*** de los 30-pozos de la Hoja La Laguna Colorada Grande.
- Observamos que el valor mínimo $=$ 152 y el valor máximo $=$ 880, mostrando una gran variabilidad en los datos de esta variable.
- Un método de conteo para resumir estos datos es considerar el procedimiento (gráfico, y en el que no se pierde información de los datos) denominado "**Tallo y Hoja**" (Stem and Leaf):

Análisis Exploratorio de Datos Unidimensionales

- **Distribución de Frecuencias para Variables Cuantitativas**
- Consideremos las mediciones de las concentraciones del ***"Calcio"*** de los 30-pozos de la Hoja La Laguna Colorada Grande.
- Observamos que el valor mínimo $=$ 152 y el valor máximo $=$ 880, mostrando una gran variabilidad en los datos de esta variable.
- Un método de conteo para resumir estos datos es considerar el procedimiento (gráfico, y en el que no se pierde información de los datos) denominado "**Tallo y Hoja**" (Stem and Leaf):

- **Distribución de Frecuencias para Variables Cuantitativas**
- Consideremos las mediciones de las concentraciones del ***"Calcio"*** de los 30-pozos de la Hoja La Laguna Colorada Grande.
- Observamos que el valor mínimo $=$ 152 y el valor máximo $=$ 880, mostrando una gran variabilidad en los datos de esta variable.
- Un método de conteo para resumir estos datos es considerar el procedimiento (gráfico, y en el que no se pierde información de los datos) denominado "**Tallo y Hoja**" (Stem and Leaf):

- **Distribución de Frecuencias para Variables Cuantitativas**
- Consideremos las mediciones de las concentraciones del ***"Calcio"*** de los 30-pozos de la Hoja La Laguna Colorada Grande.
- Observamos que el valor mínimo $=$ 152 y el valor máximo $=$ 880, mostrando una gran variabilidad en los datos de esta variable.
- Un método de conteo para resumir estos datos es considerar el procedimiento (gráfico, y en el que no se pierde información de los datos) denominado "**Tallo y Hoja**" (Stem and Leaf):

Análisis Exploratorio de Datos Unidimensionales

$[100 - 200)|52$

$[200 - 300)|76\ 80\ 60\ 36\ 28$

$[300 - 400)|80\ 48\ 04\ 24\ 20\ 44\ 48\ 76\ 44\ 08\ 44\ 56\ 88\ 56\ 52\ 72\ 24\ 32\ 48$

$[400 - 500)|60\ 48$

$[500 - 600)|60$

$[600 - 700)|$

$[700 - 800)|88$

$[800 - 900)|80$

Se observa:

1. Un destaque importante de los valores 788 y 880 (outliers).
2. Los demás valores están concentrados entre 152 y 560.
3. Un valor “más o menos” típico del calcio es 350.
4. Hay una leve-asimetría hacia valores grandes.

Análisis Exploratorio de Datos Unidimensionales

$[100 - 200)|52$

$[200 - 300)|76\ 80\ 60\ 36\ 28$

$[300 - 400)|80\ 48\ 04\ 24\ 20\ 44\ 48\ 76\ 44\ 08\ 44\ 56\ 88\ 56\ 52\ 72\ 24\ 32\ 48$

$[400 - 500)|60\ 48$

$[500 - 600)|60$

$[600 - 700)|$

$[700 - 800)|88$

$[800 - 900)|80$

Se observa:

1. Un destaque importante de los valores 788 y 880 (outliers).
2. Los demás valores están concentrados entre 152 y 560.
3. Un valor “más o menos” típico del calcio es 350.
4. Hay una leve-asimetría hacia valores grandes.

Análisis Exploratorio de Datos Unidimensionales

$[100 - 200)|52$

$[200 - 300)|76\ 80\ 60\ 36\ 28$

$[300 - 400)|80\ 48\ 04\ 24\ 20\ 44\ 48\ 76\ 44\ 08\ 44\ 56\ 88\ 56\ 52\ 72\ 24\ 32\ 48$

$[400 - 500)|60\ 48$

$[500 - 600)|60$

$[600 - 700)|$

$[700 - 800)|88$

$[800 - 900)|80$

Se observa:

1. Un destaque importante de los valores 788 y 880 (outliers).
2. Los demás valores están concentrados entre 152 y 560.
3. Un valor “más o menos” típico del calcio es 350.
4. Hay una leve-asimetría hacia valores grandes.

Análisis Exploratorio de Datos Unidimensionales

$[100 - 200)|52$

$[200 - 300)|76\ 80\ 60\ 36\ 28$

$[300 - 400)|80\ 48\ 04\ 24\ 20\ 44\ 48\ 76\ 44\ 08\ 44\ 56\ 88\ 56\ 52\ 72\ 24\ 32\ 48$

$[400 - 500)|60\ 48$

$[500 - 600)|60$

$[600 - 700)|$

$[700 - 800)|88$

$[800 - 900)|80$

Se observa:

1. Un destaque importante de los valores 788 y 880 (outliers).
2. Los demás valores están concentrados entre 152 y 560.
3. Un valor “más o menos” típico del calcio es 350.
4. Hay una leve-asimetría hacia valores grandes.

Análisis Exploratorio de Datos Unidimensionales

$[100 - 200)|52$

$[200 - 300)|76\ 80\ 60\ 36\ 28$

$[300 - 400)|80\ 48\ 04\ 24\ 20\ 44\ 48\ 76\ 44\ 08\ 44\ 56\ 88\ 56\ 52\ 72\ 24\ 32\ 48$

$[400 - 500)|60\ 48$

$[500 - 600)|60$

$[600 - 700)|$

$[700 - 800)|88$

$[800 - 900)|80$

Se observa:

1. Un destaque importante de los valores 788 y 880 (outliers).
2. Los demás valores están concentrados entre 152 y 560.
3. Un valor “más o menos” típico del calcio es 350.
4. Hay una leve-asimetría hacia valores grandes.

- Resumiendo en una Tabla de Distribución de Frecuencias tenemos:

Calcio	Frecuencia Observada: n_i	Frecuencia Relativa: f_i	Porcentaje (%)	Porcentaje Acumulado
$[100 - 200)$	1	$0,0333$	$3,33\,\%$	$3,33\,\%$
$[200 - 300)$	5	$0,1667$	$16,67\,\%$	$20,00\,\%$
$[300 - 400)$	19	$0,6333$	$63,33\,\%$	$83,33\,\%$
$[400 - 500)$	2	$0,0667$	$6,67\,\%$	$90,00\,\%$
$[500 - 600)$	1	$0,0333$	$3,33\,\%$	$96,67\,\%$
$[600 - 900)$	2	$0,0667$	$6,67\,\%$	$100\,\%$
Total	30	$1,00$	$100\,\%$	$--$

Cuadro: Distribución de Frecuencias del Calcio en 30 pozos de la Laguna Colorada Grande

Análisis Exploratorio de Datos Unidimensionales

- Utilizando el librería **RCommander** del paquete **R** se obtuvo la siguiente gráfica de “Tallo y Hoja”:

```
> with(Calcio, stem.leaf(Calcio, m=1, na.rm=TRUE))
1 | 2: represents 120
 leaf unit: 10
            n: 30
LO: 152
     6     2 | 23678
   (19)    3 | 0022234444445557788
     5     4 | 46
     3     5 | 6
HI: 788 880
```

- Hay diferentes algoritmos en RCommader para formar este gráfico (se usó el tipo 1 coincidente con el razonamiento manual).
- **Ejercicio:** Investigue las otras opciones y realice comparaciones.

Análisis Exploratorio de Datos Unidimensionales

- Utilizando el librería **RCommander** del paquete **R** se obtuvo la siguiente gráfica de “Tallo y Hoja”:

```
> with(Calcio, stem.leaf(Calcio, m=1, na.rm=TRUE))
1 | 2: represents 120
 leaf unit: 10
            n: 30
LO: 152
    6     2 | 23678
  (19)    3 | 0022234444445557788
    5     4 | 46
    3     5 | 6
HI: 788 880
```

- Hay diferentes algoritmos en RCommader para formar este gráfico (se usó el tipo 1 coincidente con el razonamiento manual).
- **Ejercicio:** Investigue las otras opciones y realice comparaciones.

Análisis Exploratorio de Datos Unidimensionales

- Utilizando el librería **RCommander** del paquete **R** se obtuvo la siguiente gráfica de “Tallo y Hoja”:

```
> with(Calcio, stem.leaf(Calcio, m=1, na.rm=TRUE))
1 | 2: represents 120
 leaf unit: 10
            n: 30
LO: 152
     6     2 | 23678
   (19)    3 | 0022234444445557788
     5     4 | 46
     3     5 | 6
HI: 788 880
```

- Hay diferentes algoritmos en RCommader para formar este gráfico (se usó el tipo 1 coincidente con el razonamiento manual).
- **Ejercicio:** Investigue las otras opciones y realice comparaciones.

Análisis Exploratorio de Datos Unidimensionales

- ¿En cuántos **intervalos** o **clases** puede resumirse un conjunto de datos?
- En general, la elección del número de intervalos es arbitraria y la familiaridad del investigador con los datos es lo que indicará **cuántos** y **cuáles** clases (o intervalos) deben ser usados.
- Pero...
  - con un número pequeño de intervalos se pierde información!
  - con un número grande de intervalos el objetivo de "**resumir**" los datos se ve perjudicado!
- Los investigadores estadisticos aplicados, normalmente, sugieren el uso de **5 a 15 clases**, con la **misma amplitud**.
- **Observación:** Si los intervalos son de diferente amplitud debe tenerse cuidado en la elaboración de un gráfico que resuma y explique la distribución de la variable de interés!

Análisis Exploratorio de Datos Unidimensionales

- ¿En cuántos **intervalos** o **clases** puede resumirse un conjunto de datos?
- En general, la elección del número de intervalos es arbitraria y la familiaridad del investigador con los datos es lo que indicará **cuántos** y **cuáles** clases (o intervalos) deben ser usados.
- Pero...
  - con un número pequeño de intervalos se pierde información!
  - con un número grande de intervalos el objetivo de "**resumir**" los datos se ve perjudicado!
- Los investigadores estadisticos aplicados, normalmente, sugieren el uso de **5 a 15 clases**, con la **misma amplitud**.
- **Observación:** Si los intervalos son de diferente amplitud debe tenerse cuidado en la elaboración de un gráfico que resuma y explique la distribución de la variable de interés!

Análisis Exploratorio de Datos Unidimensionales

- ¿En cuántos **intervalos** o **clases** puede resumirse un conjunto de datos?
- En general, la elección del número de intervalos es arbitraria y la familiaridad del investigador con los datos es lo que indicará **cuántos** y **cuáles** clases (o intervalos) deben ser usados.
- Pero...
  - con un número pequeño de intervalos se pierde información!
  - con un número grande de intervalos el objetivo de **"resumir"** los datos se ve perjudicado!
- Los investigadores estadisticos aplicados, normalmente, sugieren el uso de **5 a 15 clases**, con la **misma amplitud**.
- **Observación:** Si los intervalos son de diferente amplitud debe tenerse cuidado en la elaboración de un gráfico que resuma y explique la distribución de la variable de interés!

Análisis Exploratorio de Datos Unidimensionales

- ¿En cuántos **intervalos** o **clases** puede resumirse un conjunto de datos?
- En general, la elección del número de intervalos es arbitraria y la familiaridad del investigador con los datos es lo que indicará **cuántos** y **cuáles** clases (o intervalos) deben ser usados.
- Pero...
  - con un número pequeño de intervalos se pierde información!
  - con un número grande de intervalos el objetivo de **"resumir"** los datos se ve perjudicado!
- Los investigadores estadisticos aplicados, normalmente, sugieren el uso de **5 a 15 clases**, con la **misma amplitud**.
- **Observación:** Si los intervalos son de diferente amplitud debe tenerse cuidado en la elaboración de un gráfico que resuma y explique la distribución de la variable de interés!

Análisis Exploratorio de Datos Unidimensionales

- ¿En cuántos **intervalos** o **clases** puede resumirse un conjunto de datos?
- En general, la elección del número de intervalos es arbitraria y la familiaridad del investigador con los datos es lo que indicará **cuántos** y **cuáles** clases (o intervalos) deben ser usados.
- Pero...
  - con un número pequeño de intervalos se pierde información!
  - con un número grande de intervalos el objetivo de **"resumir"** los datos se ve perjudicado!
- Los investigadores estadisticos aplicados, normalmente, sugieren el uso de **5 a 15 clases**, con la **misma amplitud**.
- **Observación:** Si los intervalos son de diferente amplitud debe tenerse cuidado en la elaboración de un gráfico que resuma y explique la distribución de la variable de interés!

Análisis Exploratorio de Datos Unidimensionales

- Una forma de determinar el número de intervalos en los que agrupar la variable es utilizando la **Fórmula de Sturges**:

$$\#\text{ Intervalos} = 1 + 3,3log(n) = 1 + 3,3(1,4771) = 5,87 \approx 6$$

o equivalentemente:

$$\#\text{ Intervalos} = 1 + 3,3ln(n) = 1 + 3,3(3,40) = 12,22 \approx 12$$

- **Observe** que por un cálculo se obtendría el número mínimo de intervalos en los que agrupar la variable, y por el otro, el número máximo de intervalos (tenga en cuenta la solución empírica antes señalada).
- Para determinar la amplitud de estos intervalos se procede calculando:

$$\Delta_i = \frac{\text{Rango}}{\text{N° intervalos}} = \frac{880 - 152}{6} = 121,33 \approx 120.$$

Análisis Exploratorio de Datos Unidimensionales

- Una forma de determinar el número de intervalos en los que agrupar la variable es utilizando la **Fórmula de Sturges**:

$$\#\text{ Intervalos} = 1 + 3,3log(n) = 1 + 3,3(1,4771) = 5,87 \approx 6$$

o equivalentemente:

$$\#\text{ Intervalos} = 1 + 3,3ln(n) = 1 + 3,3(3,40) = 12,22 \approx 12$$

- **Observe** que por un cálculo se obtendría el número mínimo de intervalos en los que agrupar la variable, y por el otro, el número máximo de intervalos (tenga en cuenta la solución empírica antes señalada).
- Para determinar la amplitud de estos intervalos se procede calculando:

$$\Delta_i = \frac{\text{Rango}}{\text{N° intervalos}} = \frac{880 - 152}{6} = 121,33 \approx 120.$$

Análisis Exploratorio de Datos Unidimensionales

- Una forma de determinar el número de intervalos en los que agrupar la variable es utilizando la **Fórmula de Sturges**:

$$\#\text{ Intervalos} = 1 + 3,3log(n) = 1 + 3,3(1,4771) = 5,87 \approx 6$$

o equivalentemente:

$$\#\text{ Intervalos} = 1 + 3,3ln(n) = 1 + 3,3(3,40) = 12,22 \approx 12$$

- **Observe** que por un cálculo se obtendría el número mínimo de intervalos en los que agrupar la variable, y por el otro, el número máximo de intervalos (tenga en cuenta la solución empírica antes señalada).
- Para determinar la amplitud de estos intervalos se procede calculando:

$$\Delta_i = \frac{\text{Rango}}{\text{N° intervalos}} = \frac{880 - 152}{6} = 121,33 \approx 120.$$

Análisis Exploratorio de Datos Unidimensionales

- Un gráfico apropiado para Variables Cuantitativas Continuas es el **HISTOGRAMA**.
- Un histograma es un gráfico de sectores contiguos, donde la altura es proporcional a la frecuencia relativa f_i y la base está constituida por un segmento cuyos extremos representan los extremos de la i-ésima clase.
- El único cuidado que debe tomarse es que **el area total sea igual a 1**, correspondiendo con la suma total de proporciones.
- Si los intervalos o clases tienen igual amplitud, se puede construir el histograma a partir de los pares (clases; n_i) o bien (clases; f_i) (los gráficos serán proporcionales).
- Si los intervalos o clases tuvieran distinta amplitud deberán construirse los pares (clases; *densidad*) donde la densidad se define por $\frac{n_i}{\Delta_i}$ o $\frac{f_i}{\Delta_i}$.

Análisis Exploratorio de Datos Unidimensionales

- Un gráfico apropiado para Variables Cuantitativas Continuas es el **HISTOGRAMA**.
- Un histograma es un gráfico de sectores contiguos, donde la altura es proporcional a la frecuencia relativa f_i y la base está constituida por un segmento cuyos extremos representan los extremos de la i-ésima clase.
- El único cuidado que debe tomarse es que **el area total sea igual a 1**, correspondiendo con la suma total de proporciones.
- Si los intervalos o clases tienen igual amplitud, se puede construir el histograma a partir de los pares (clases; n_i) o bien (clases; f_i) (los gráficos serán proporcionales).
- Si los intervalos o clases tuvieran distinta amplitud deberán construirse los pares (clases; *densidad*) donde la densidad se define por $\frac{n_i}{\Delta_i}$ o $\frac{f_i}{\Delta_i}$.

Análisis Exploratorio de Datos Unidimensionales

- Un gráfico apropiado para Variables Cuantitativas Continuas es el **HISTOGRAMA**.
- Un histograma es un gráfico de sectores contiguos, donde la altura es proporcional a la frecuencia relativa f_i y la base está constituida por un segmento cuyos extremos representan los extremos de la i-ésima clase.
- El único cuidado que debe tomarse es que **el area total sea igual a 1**, correspondiendo con la suma total de proporciones.
- Si los intervalos o clases tienen igual amplitud, se puede construir el histograma a partir de los pares (clases; n_i) o bien (clases; f_i) (los gráficos serán proporcionales).
- Si los intervalos o clases tuvieran distinta amplitud deberán construirse los pares (clases; *densidad*) donde la densidad se define por $\frac{n_i}{\Delta_i}$ o $\frac{f_i}{\Delta_i}$.

Análisis Exploratorio de Datos Unidimensionales

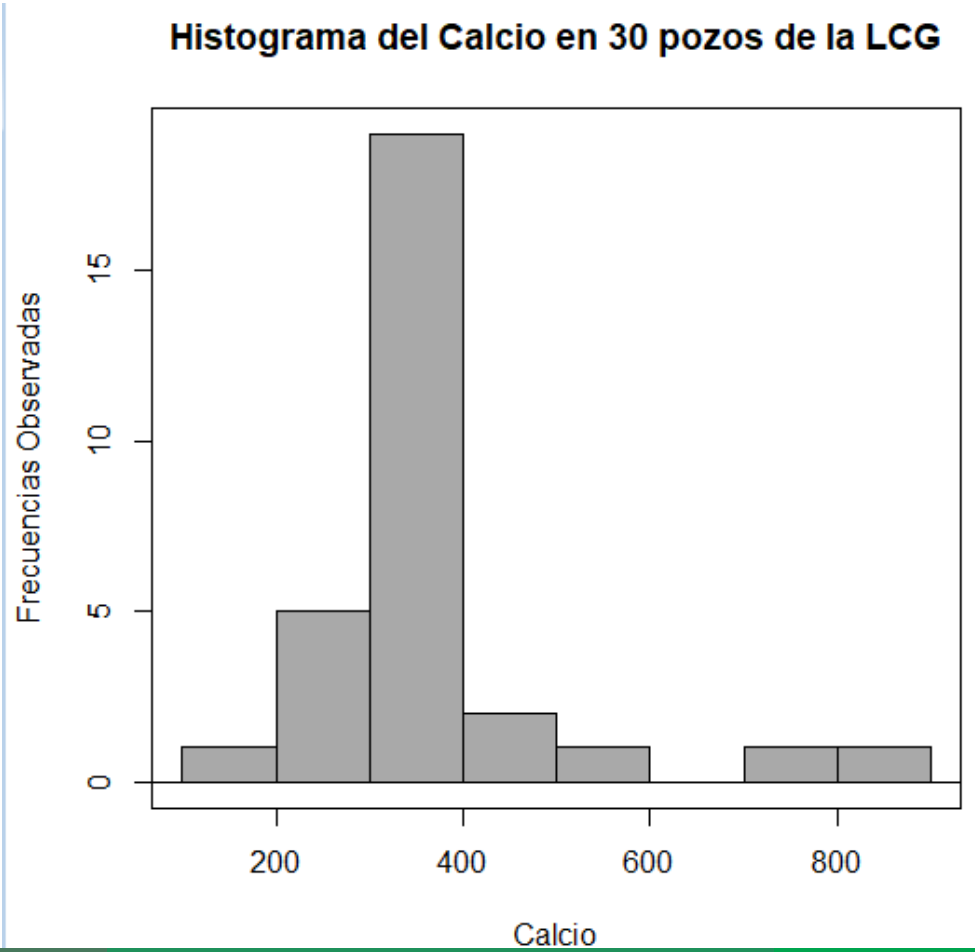
- Un gráfico apropiado para Variables Cuantitativas Continuas es el **HISTOGRAMA**.
- Un histograma es un gráfico de sectores contiguos, donde la altura es proporcional a la frecuencia relativa f_i y la base está constituida por un segmento cuyos extremos representan los extremos de la i-ésima clase.
- El único cuidado que debe tomarse es que **el area total sea igual a 1**, correspondiendo con la suma total de proporciones.
- Si los intervalos o clases tienen igual amplitud, se puede construir el histograma a partir de los pares (clases; n_i) o bien (clases; f_i) (los gráficos serán proporcionales).
- Si los intervalos o clases tuvieran distinta amplitud deberán construirse los pares (clases; *densidad*) donde la densidad se define por $\frac{n_i}{\Delta_i}$ o $\frac{f_i}{\Delta_i}$.

Análisis Exploratorio de Datos Unidimensionales

- Un gráfico apropiado para Variables Cuantitativas Continuas es el **HISTOGRAMA**.
- Un histograma es un gráfico de sectores contiguos, donde la altura es proporcional a la frecuencia relativa f_i y la base está constituida por un segmento cuyos extremos representan los extremos de la i-ésima clase.
- El único cuidado que debe tomarse es que **el area total sea igual a 1**, correspondiendo con la suma total de proporciones.
- Si los intervalos o clases tienen igual amplitud, se puede construir el histograma a partir de los pares (clases; n_i) o bien (clases; f_i) (los gráficos serán proporcionales).
- Si los intervalos o clases tuvieran distinta amplitud deberán construirse los pares (clases; *densidad*) donde la densidad se define por $\frac{n_i}{\Delta_i}$ o $\frac{f_i}{\Delta_i}$.

Análisis Exploratorio de Datos Unidimensionales

- Así, para las *"concentraciones de Calcio"*, de interés en nuestro conjunto de datos de la Laguna Colorada Grande, el histograma a partir de las clases construidas por el Tallo y Hoja (clases de igual amplitud) tiene la forma:



Análisis Exploratorio de Datos Unidimensionales

- En el caso de contar con Variables Cuantitativas **Discretas** el procedimiento de resumir los datos en una tabla de frecuencias es similar a aquel anteriormente realizado para variables cualitativas o cuantitativas continuas.
- La diferencia básicamente radica en los valores que la variable asume, siendo su recorrido los Números Enteros, incluido el cero.
- Por **Ejemplo:** La Distribución del N° de Asignaturas aprobadas por una muestra de 20-estudiantes del curso de Estadistica es:

Materias aprobadas	Frecuencia Observada: n_i	Frecuencia Relativa: f_i	Porcentaje (%)	Porcentaje Acumulado
0	4	0, 20	20 %	20 %
1	5	0, 25	25 %	45 %
2	7	0, 35	35 %	80 %
3	3	0, 15	15 %	95 %
4	0	0, 00	0 %	95 %
5	1	0, 05	5 %	100 %
Total	20	1, 00	100 %	– –

Cuadro: Distribución de Frecuencias del N° de Materias aprobadas por una muestra de estudiantes del Curso de Estadística

Análisis Exploratorio de Datos Unidimensionales

- En el caso de contar con Variables Cuantitativas **Discretas** el procedimiento de resumir los datos en una tabla de frecuencias es similar a aquel anteriormente realizado para variables cualitativas o cuantitativas continuas.
- La diferencia básicamente radica en los valores que la variable asume, siendo su recorrido los Números Enteros, incluido el cero.
- Por **Ejemplo:** La Distribución del N° de Asignaturas aprobadas por una muestra de 20-estudiantes del curso de Estadistica es:

Materias aprobadas	Frecuencia Observada: n_i	Frecuencia Relativa: f_i	Porcentaje (%)	Porcentaje Acumulado
0	4	0, 20	20 %	20 %
1	5	0, 25	25 %	45 %
2	7	0, 35	35 %	80 %
3	3	0, 15	15 %	95 %
4	0	0, 00	0 %	95 %
5	1	0, 05	5 %	100 %
Total	20	1, 00	100 %	– –

Cuadro: Distribución de Frecuencias del N° de Materias aprobadas por una muestra de estudiantes del Curso de Estadística

Análisis Exploratorio de Datos Unidimensionales

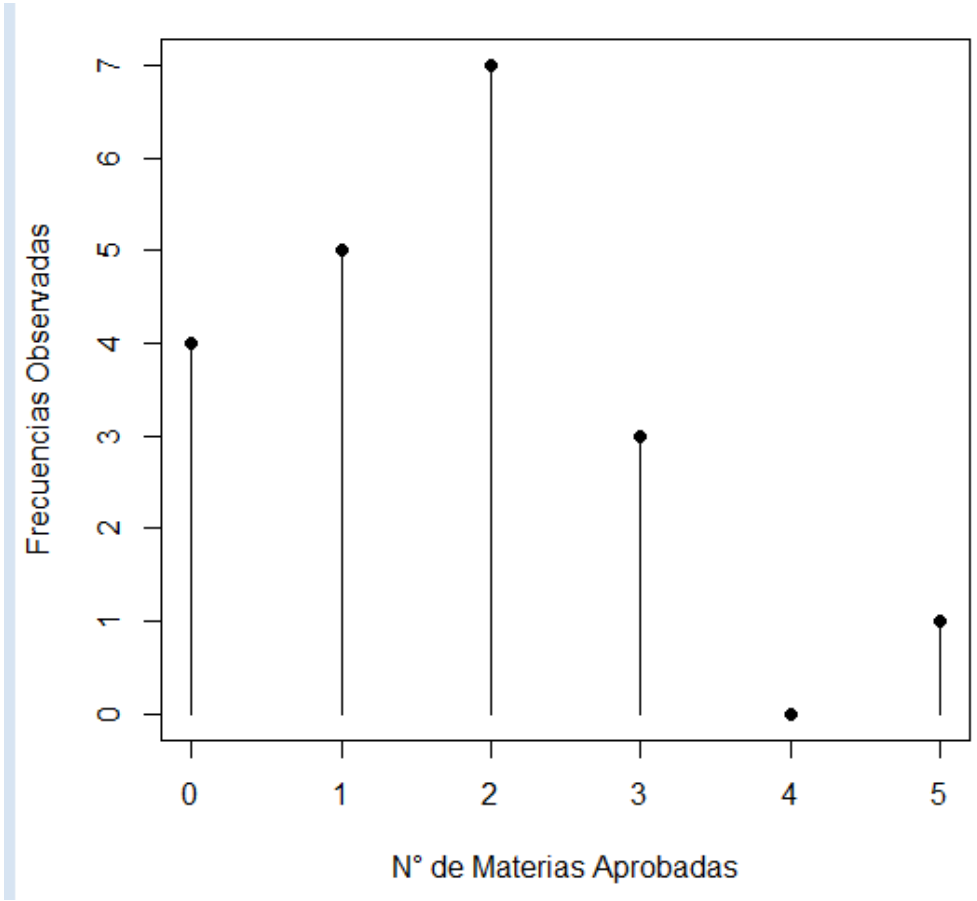
- En el caso de contar con Variables Cuantitativas **Discretas** el procedimiento de resumir los datos en una tabla de frecuencias es similar a aquel anteriormente realizado para variables cualitativas o cuantitativas continuas.
- La diferencia básicamente radica en los valores que la variable asume, siendo su recorrido los Números Enteros, incluido el cero.
- Por **Ejemplo:** La Distribución del N° de Asignaturas aprobadas por una muestra de 20-estudiantes del curso de Estadistica es:

Materias aprobadas	Frecuencia Observada: n_i	Frecuencia Relativa: f_i	Porcentaje (%)	Porcentaje Acumulado
0	4	0,20	20 %	20 %
1	5	0,25	25 %	45 %
2	7	0,35	35 %	80 %
3	3	0,15	15 %	95 %
4	0	0,00	0 %	95 %
5	1	0,05	5 %	100 %
Total	20	1,00	100 %	−−

Cuadro: Distribución de Frecuencias del N° de Materias aprobadas por una muestra de estudiantes del Curso de Estadística

Análisis Exploratorio de Datos Unidimensionales

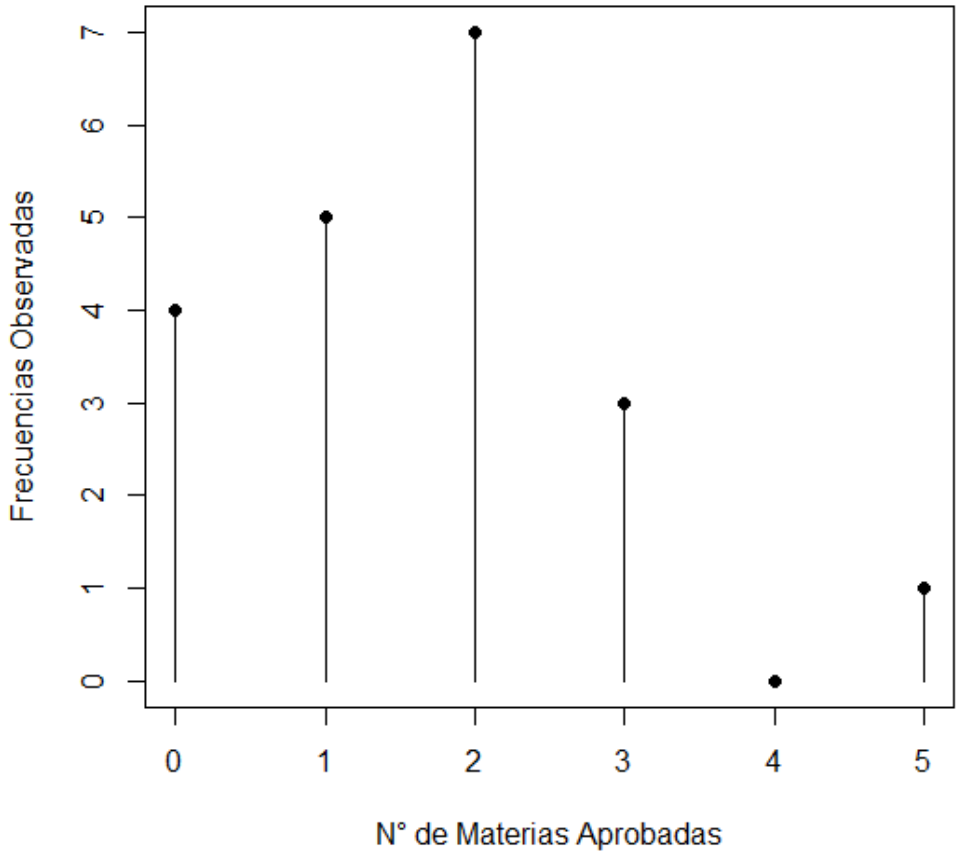
- El gráfico apropiado para resumir la información de una variable discreta es conocido como **"Gráfico de Bastones"**.
- Para la variable *"Número de Materias aprobadas"* el gráfico es:



- Algunos suelen usar "Gráfico de Barras". NO ES APROPIADO!!!

Análisis Exploratorio de Datos Unidimensionales

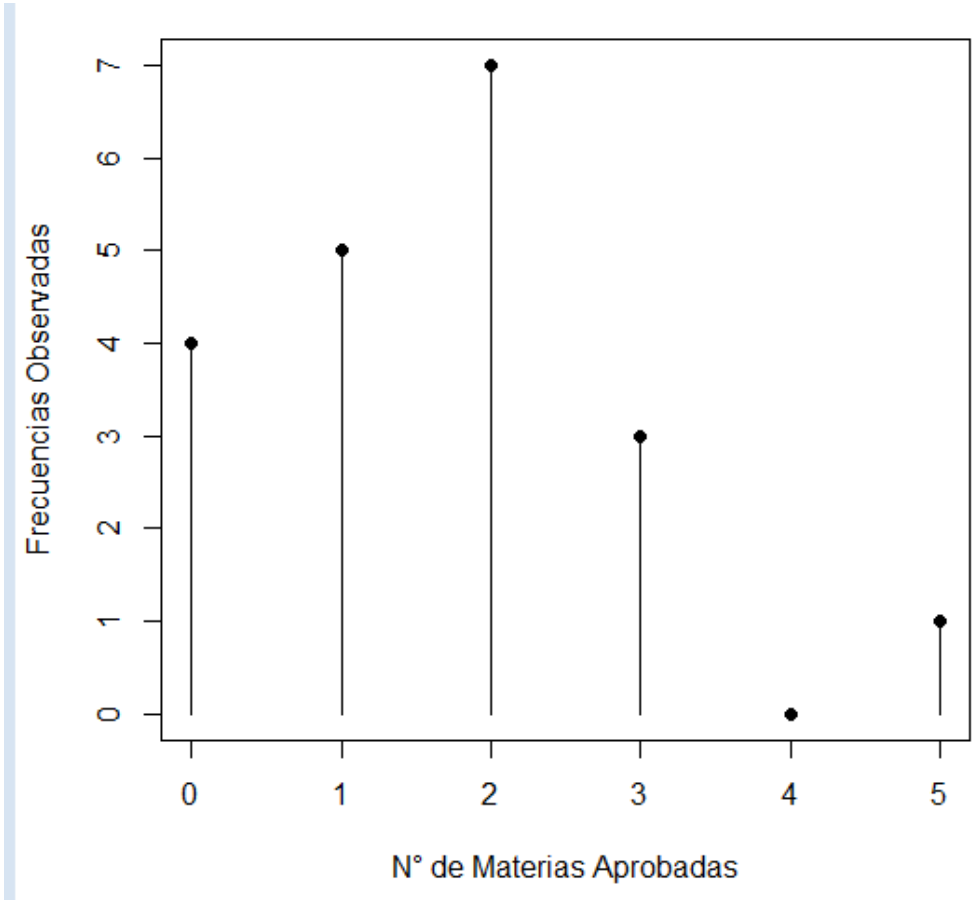
- El gráfico apropiado para resumir la información de una variable discreta es conocido como **"Gráfico de Bastones"**.
- Para la variable *"Número de Materias aprobadas"* el gráfico es:



- Algunos suelen usar "Gráfico de Barras". NO ES APROPIADO!!!

Análisis Exploratorio de Datos Unidimensionales

- El gráfico apropiado para resumir la información de una variable discreta es conocido como **"Gráfico de Bastones"**.
- Para la variable *"Número de Materias aprobadas"* el gráfico es:



- Algunos suelen usar "Gráfico de Barras". NO ES APROPIADO!!!

- **Otros gráficos:**

- **Ejemplo:** Los datos a seguir corresponden a **"Salarios anuales" (en miles de dólares)** de 40-Directores de Laboratorios Farmacéuticos, para el año 2008:

25	50	42	65	76	43	55	53	44	90
53	33	54	66	46	62	44	54	76	47
63	39	35	55	75	28	60	40	68	66
62	62	57	41	91	29	63	58	61	43

- **Otros gráficos:**

- **Ejemplo:** Los datos a seguir corresponden a **"Salarios anuales" (en miles de dólares)** de 40-Directores de Laboratorios Farmacéuticos, para el año 2008:

25	50	42	65	76	43	55	53	44	90
53	33	54	66	46	62	44	54	76	47
63	39	35	55	75	28	60	40	68	66
62	62	57	41	91	29	63	58	61	43

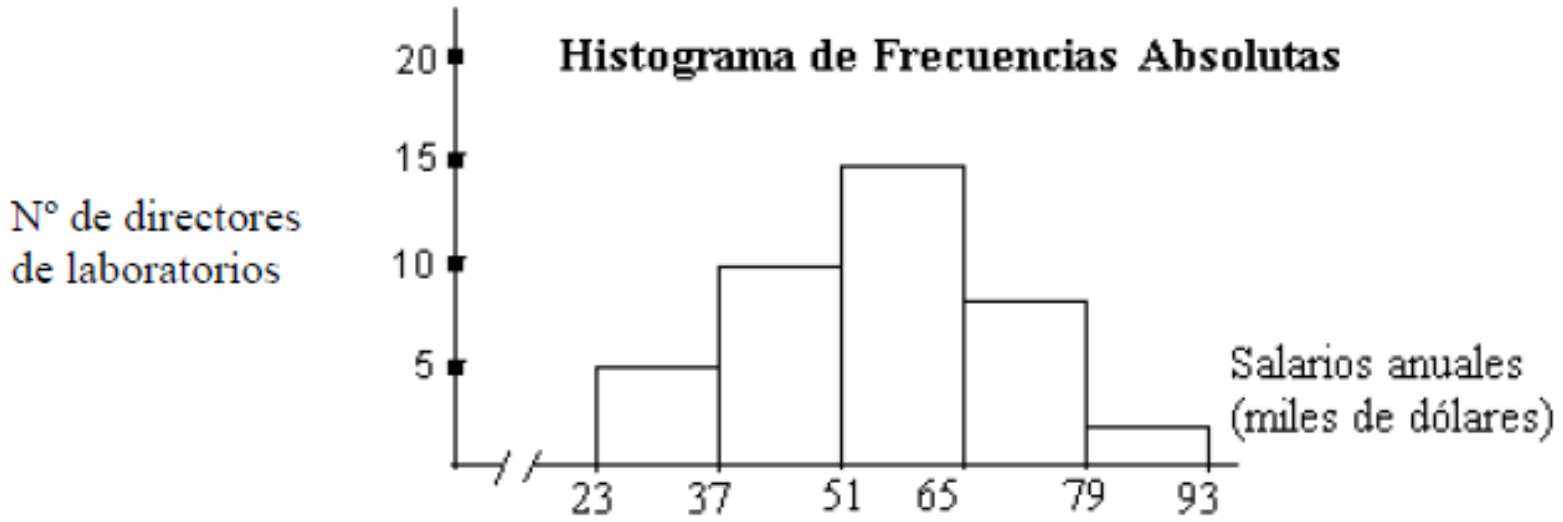
- Los investigadores ofrecieron la siguiente Tabla de Distribución de Frecuencias para estos datos de salarios, agrupados en 5-clases de igual amplitud:

Salarios anuales (miles de dólares) Clase	N° de directores de laboratorios f_i	Marca de Clase	F_i	f_{ri}	F_{ri}
[23 , 37)	5	30	5	5/40	5/40
[37 , 51)	11	44	16	11/40	16/40
[51 , 65)	15	58	31	15/40	31/40
[65 , 79)	7	72	38	7/40	38/40
[79 , 93]	2	86	40	2/40	40/40

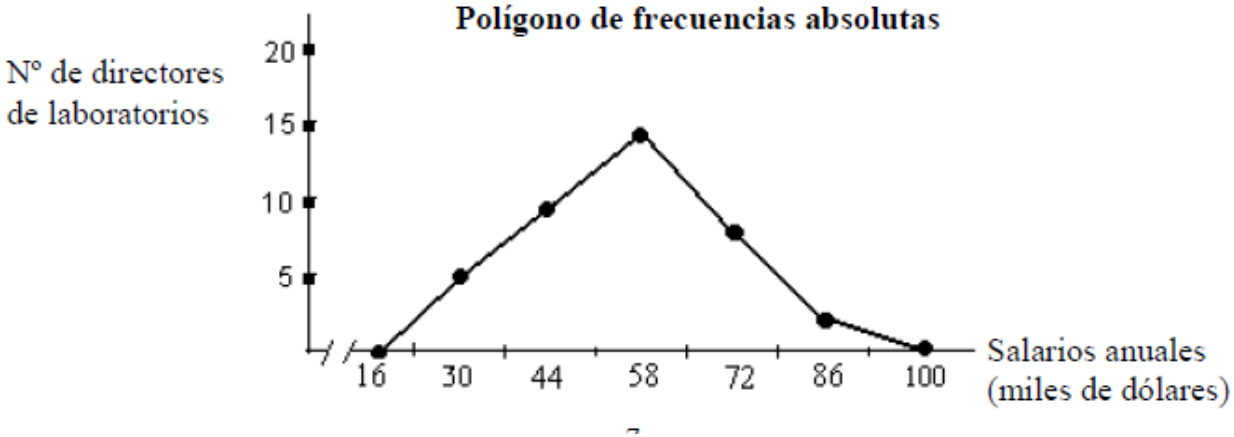
denotando por:

- f_i la frecuencia observada en la i-ésima clase;
- F_i la frecuencia observada acumulada en la clase i;
- f_{ri} la frecuencia relativa en la i-ésima clase; y
- F_{ri} la frecuencia relativa acumulada en la i-ésima clase.

- También acompañan su análisis con el siguiente **Histograma** de los Salarios anuales de 40-directores de Laboratorios Farmacéuticos:

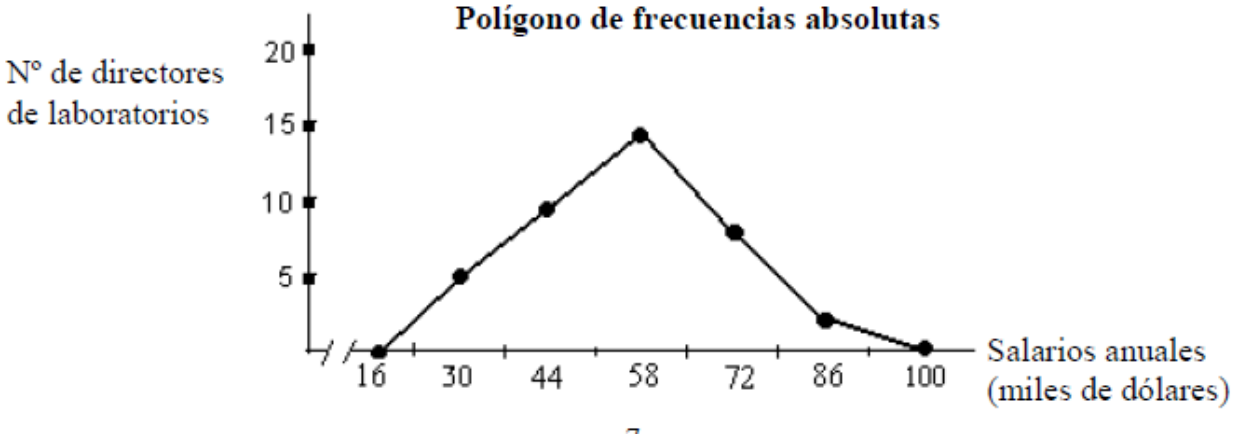


- E incluyen el siguiente gráfico, denominado **"Polígono de Frecuencias Absolutas"**, útil para visualizar, rápidamente, la forma de la distribución de frecuencias de la variable en estudio.



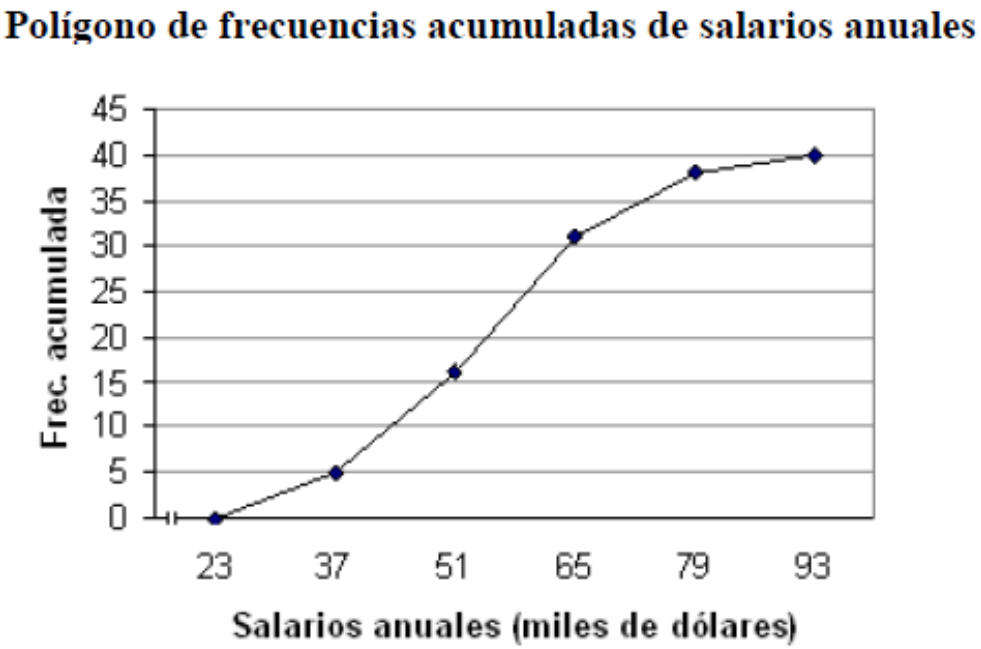
- El **Polígono de Frecuencias Absolutas** se construye uniendo mediante segmentos los puntos medios de cada intervalo (marca de clase). Se debe continuar la línea construida hasta la marca de clase de intervalos hipotéticos anterior al primero y posterior al último, de modo que el área encerrada bajo la poligonal sea igual a la suma de las áreas de los rectángulos que componen el correspondiente histograma.

- E incluyen el siguiente gráfico, denominado **"Polígono de Frecuencias Absolutas"**, útil para visualizar, rápidamente, la forma de la distribución de frecuencias de la variable en estudio.



- El **Polígono de Frecuencias Absolutas** se construye uniendo mediante segmentos los puntos medios de cada intervalo (marca de clase). Se debe continuar la línea construida hasta la marca de clase de intervalos hipotéticos anterior al primero y posterior al último, de modo que el área encerrada bajo la poligonal sea igual a la suma de las áreas de los rectángulos que componen el correspondiente histograma.

- Finalmente, los investigadores ofrecen otro gráfico para el análisis de estos datos, denominado **"Polígono de frecuencias acumuladas u OJIVA"**:



Análisis Exploratorio de Datos Unidimensionales

- El **objetivo** del **Polígono de frecuencias acumuladas u OJIVA**, al igual que el Histograma y el Polígono de Frecuencias Absolutas, es representar distribuciones de frecuencias de variables medidas en escala numérica continua, pero sólo para frecuencias acumuladas.
- Se utilizan en su confección segmentos de recta, siendo por ello útil, tanto para representar una distribución de frecuencias como para comparar dos o más distribuciones.
- La **diferencia** con el polígono de frecuencias es que la frecuencia acumulada no se grafica sobre el punto medio de la clase, sino al final de la misma, ya que representa el número de individuos acumulados hasta esa clase. Como el valor de la frecuencia acumulada es mayor a medida que se avanza en la distribución, la poligonal que se obtiene siempre va a ser creciente y esa forma particular de la misma es la que ha hecho que se le dé también el nombre de ojiva.

Análisis Exploratorio de Datos Unidimensionales

- El **objetivo** del **Polígono de frecuencias acumuladas u OJIVA**, al igual que el Histograma y el Polígono de Frecuencias Absolutas, es representar distribuciones de frecuencias de variables medidas en escala numérica continua, pero sólo para frecuencias acumuladas.
- Se utilizan en su confección segmentos de recta, siendo por ello útil, tanto para representar una distribución de frecuencias como para comparar dos o más distribuciones.
- La **diferencia** con el polígono de frecuencias es que la frecuencia acumulada no se grafica sobre el punto medio de la clase, sino al final de la misma, ya que representa el número de individuos acumulados hasta esa clase. Como el valor de la frecuencia acumulada es mayor a medida que se avanza en la distribución, la poligonal que se obtiene siempre va a ser creciente y esa forma particular de la misma es la que ha hecho que se le dé también el nombre de ojiva.

Análisis Exploratorio de Datos Unidimensionales

- El **objetivo** del **Polígono de frecuencias acumuladas u OJIVA**, al igual que el Histograma y el Polígono de Frecuencias Absolutas, es representar distribuciones de frecuencias de variables medidas en escala numérica continua, pero sólo para frecuencias acumuladas.
- Se utilizan en su confección segmentos de recta, siendo por ello útil, tanto para representar una distribución de frecuencias como para comparar dos o más distribuciones.
- La **diferencia** con el polígono de frecuencias es que la frecuencia acumulada no se grafica sobre el punto medio de la clase, sino al final de la misma, ya que representa el número de individuos acumulados hasta esa clase. Como el valor de la frecuencia acumulada es mayor a medida que se avanza en la distribución, la poligonal que se obtiene siempre va a ser creciente y esa forma particular de la misma es la que ha hecho que se le dé también el nombre de ojiva.

- **Gráfico de Líneas**
- Es uno de los más sencillos de confeccionar. Su uso estadístico fundamental es en la representación de **series cronológicas**, y en casos particulares, como el del crecimiento y desarrollo humanos, para representar los valores promedio u otras medidas descriptivas numéricas de muchas dimensiones: peso según la edad, talla según la edad, entre otras.
- Uno de los ejes (habitualmente el horizontal) se usa para la unidad de tiempo estudiada: años, días, etc. En el otro eje se representa la frecuencia o el indicador calculado a partir de esos datos.

- **Gráfico de Líneas**
- Es uno de los más sencillos de confeccionar. Su uso estadístico fundamental es en la representación de **series cronológicas**, y en casos particulares, como el del crecimiento y desarrollo humanos, para representar los valores promedio u otras medidas descriptivas numéricas de muchas dimensiones: peso según la edad, talla según la edad, entre otras.
- Uno de los ejes (habitualmente el horizontal) se usa para la unidad de tiempo estudiada: años, días, etc. En el otro eje se representa la frecuencia o el indicador calculado a partir de esos datos.

- **Gráfico de Líneas**
- Es uno de los más sencillos de confeccionar. Su uso estadístico fundamental es en la representación de **series cronológicas**, y en casos particulares, como el del crecimiento y desarrollo humanos, para representar los valores promedio u otras medidas descriptivas numéricas de muchas dimensiones: peso según la edad, talla según la edad, entre otras.
- Uno de los ejes (habitualmente el horizontal) se usa para la unidad de tiempo estudiada: años, días, etc. En el otro eje se representa la frecuencia o el indicador calculado a partir de esos datos.

Análisis Exploratorio de Datos Unidimensionales

- En estas series se investiga ***"la tendencia´´***.
- La tendencia es un movimiento de larga duración que muestra la evolución general de la serie en el tiempo. Ese movimiento puede ser estacionario, ascendente o descendente y su recorrido, una línea recta o una curva.
- Una de la posibles formas es la que se muestra a seguir:

- En el mismo gráfico se puede representar más de una serie de datos si la escala usada se adecua para todas, y cuando los valores de esas series no son extremadamente diferentes.

Análisis Exploratorio de Datos Unidimensionales

- En estas series se investiga ***"la tendencia´´***.
- La tendencia es un movimiento de larga duración que muestra la evolución general de la serie en el tiempo. Ese movimiento puede ser estacionario, ascendente o descendente y su recorrido, una línea recta o una curva.
- Una de la posibles formas es la que se muestra a seguir:

- En el mismo gráfico se puede representar más de una serie de datos si la escala usada se adecua para todas, y cuando los valores de esas series no son extremadamente diferentes.

Análisis Exploratorio de Datos Unidimensionales

- En estas series se investiga ***"la tendencia´´***.
- La tendencia es un movimiento de larga duración que muestra la evolución general de la serie en el tiempo. Ese movimiento puede ser estacionario, ascendente o descendente y su recorrido, una línea recta o una curva.
- Una de la posibles formas es la que se muestra a seguir:

- En el mismo gráfico se puede representar más de una serie de datos si la escala usada se adecua para todas, y cuando los valores de esas series no son extremadamente diferentes.

Análisis Exploratorio de Datos Unidimensionales

- En estas series se investiga ***"la tendencia´´***.
- La tendencia es un movimiento de larga duración que muestra la evolución general de la serie en el tiempo. Ese movimiento puede ser estacionario, ascendente o descendente y su recorrido, una línea recta o una curva.
- Una de la posibles formas es la que se muestra a seguir:

- En el mismo gráfico se puede representar más de una serie de datos si la escala usada se adecua para todas, y cuando los valores de esas series no son extremadamente diferentes.

Análisis Exploratorio de Datos Unidimensionales

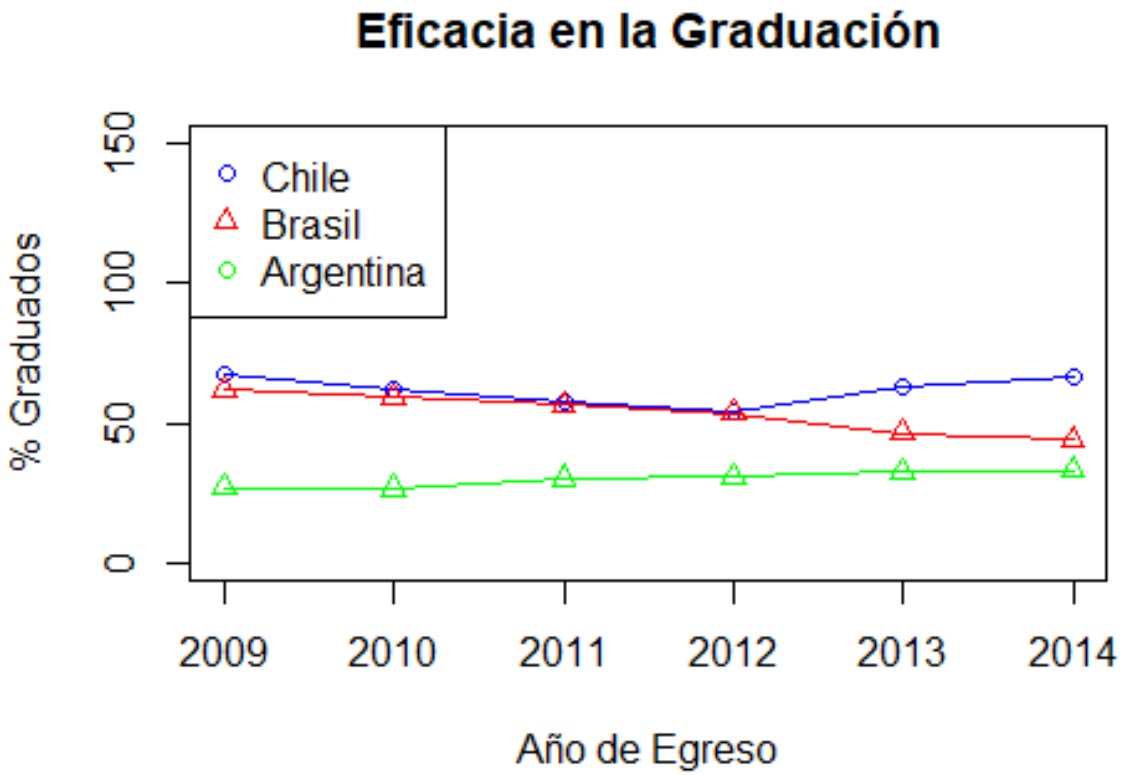
- **Ejemplo:** En el Boletin N° 5, julio-agosto 2016, del Centro de Estudios de Educación Argentina de la Universidad de Belgrano, puede encontrarse el siguiente cuadro, que compara la realidad universitaria de Argentina, Brasil y Chile, mediante la denominada *"Eficacia en la Graduación´´*, es decir la relación existente entre la cantidad graduados comparada con la cantidad de egresados:

Cuadro N° 3: Eficacia en la graduación.

Graduados/Ingresantes	Chile	Brasil	Argentina
2014/2008	66,31%	43,95%	33,03%
2013/2007	62,60%	46,35%	32,46%
2012/2006	54,37%	53,45%	30,76%
2011/2005	57,37%	56,32%	30,24%
2010/2004	61,73%	59,15%	26,69%
2009/2003	67,60%	61,70%	26,82%

Fuente: ***Argentina***: **Anuarios Estadísticos Universitarios.** ***Brasil***: **Ministerio de Educación, Censo de Educación Superior.** ***Chile***: **Servicio de Información de Educación Superior, Consejo Nacional de Educación.**

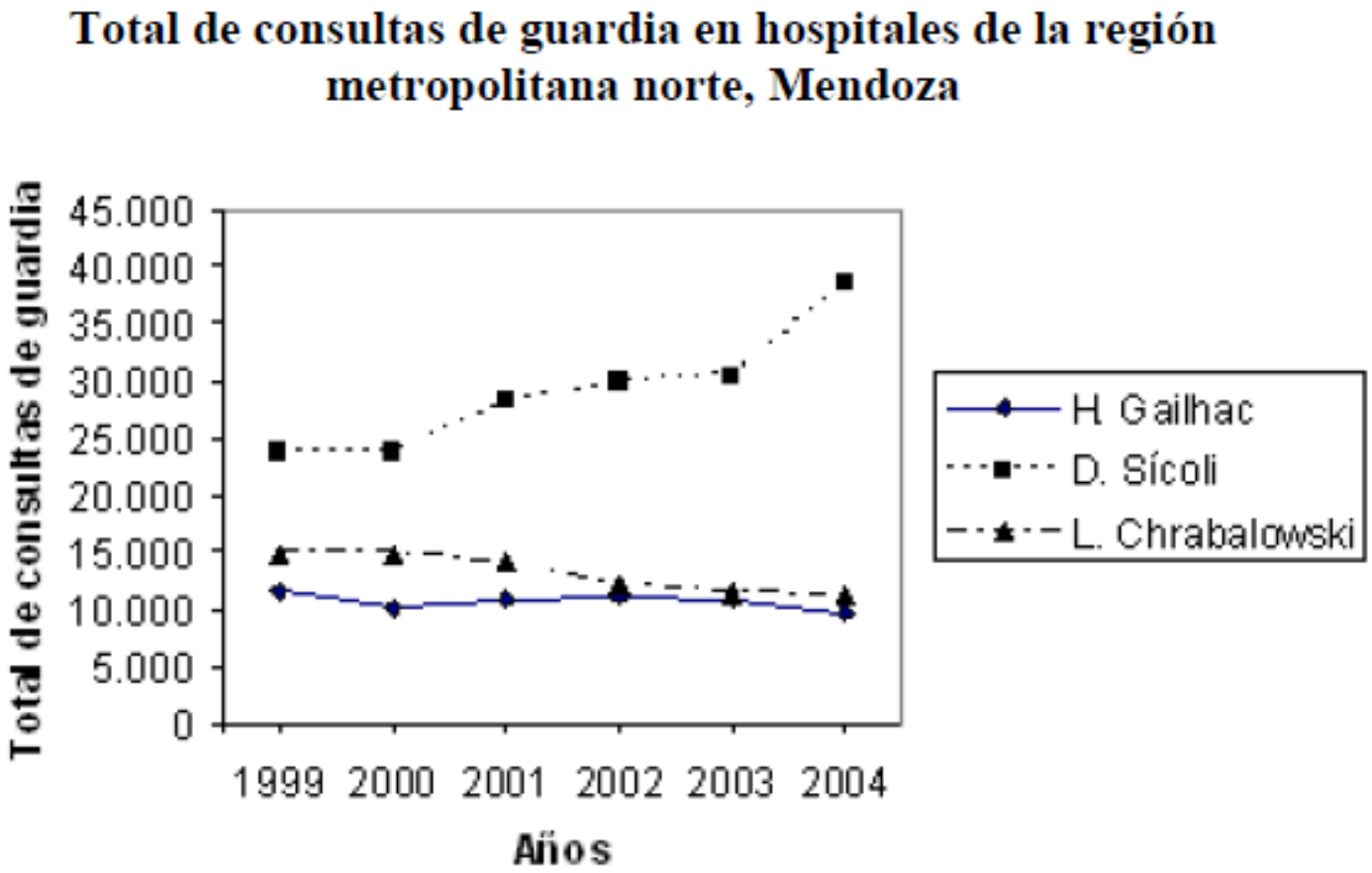
- El gráfico de líneas resultante es:



- **Ejemplo:** Los siguientes datos corresponden al total de consultas de guardia realizadas en tres hospitales de la región metropolitana norte, Mendoza, correspondiente a los años 1999 al 2004:

Año	1999	2000	2001	2002	2003	2004
H. Gailhac	11.735	10.230	10.960	11.039	10.868	9.603
D. Sícoli	23.703	23.813	28.280	29.941	30.301	38.534
L. Chrabalowski	14.902	14.918	14.000	12.034	11.269	11.059

- El gráfico de líneas resultante es:



Análisis Exploratorio de Datos Unidimensionales

- **Pictograma**: Los pictogramas son gráficos similares a los gráficos de barras, pero empleando un dibujo en una determinada escala para expresar la unidad de medida de los datos.
- Se usan para lograr el interés masivo del público.

Análisis Exploratorio de Datos Unidimensionales

- **Pictograma**: Los pictogramas son gráficos similares a los gráficos de barras, pero empleando un dibujo en una determinada escala para expresar la unidad de medida de los datos.
- Se usan para lograr el interés masivo del público.

- **Observación:** Volveremos con OTROS **Tipos de Gráficos**, útiles para estudiar la distribución de una variable, una vez que hayamos introducido una serie de medidas resumenes.

Análisis Exploratorio de Datos Unidimensionales

- Vimos que la reducción de datos a través de un gráfico o de una tabla de frecuencias nos da mucha más información sobre el comportamiento de una variable que el propio conjunto original de datos.
- Sin embargo, algunas veces queremos reducir aún más esos datos, presentando **uno o dos valores que sean "representativos" de toda la serie**.
- Al usar un único valor se tiene una reducción drástica de los datos.
- Ya hemos usado dos medidas, mínimo y máximo, como medida de posición en un conjunto de datos.
- Usualmente, se emplea una de las siguientes medidas como **Medidas de Posición Central**:

MODA, MEDIANA y MEDIA (aritmética).

Análisis Exploratorio de Datos Unidimensionales

- Vimos que la reducción de datos a través de un gráfico o de una tabla de frecuencias nos da mucha más información sobre el comportamiento de una variable que el propio conjunto original de datos.
- Sin embargo, algunas veces queremos reducir aún más esos datos, presentando **uno o dos valores que sean "representativos" de toda la serie**.
- Al usar un único valor se tiene una reducción drástica de los datos.
- Ya hemos usado dos medidas, mínimo y máximo, como medida de posición en un conjunto de datos.
- Usualmente, se emplea una de las siguientes medidas como **Medidas de Posición Central**:

MODA, MEDIANA y MEDIA (aritmética).

Análisis Exploratorio de Datos Unidimensionales

- Vimos que la reducción de datos a través de un gráfico o de una tabla de frecuencias nos da mucha más información sobre el comportamiento de una variable que el propio conjunto original de datos.
- Sin embargo, algunas veces queremos reducir aún más esos datos, presentando **uno o dos valores que sean "representativos" de toda la serie**.
- Al usar un único valor se tiene una reducción drástica de los datos.
- Ya hemos usado dos medidas, mínimo y máximo, como medida de posición en un conjunto de datos.
- Usualmente, se emplea una de las siguientes medidas como **Medidas de Posición Central**:

MODA, MEDIANA y MEDIA (aritmética).

Análisis Exploratorio de Datos Unidimensionales

- Vimos que la reducción de datos a través de un gráfico o de una tabla de frecuencias nos da mucha más información sobre el comportamiento de una variable que el propio conjunto original de datos.
- Sin embargo, algunas veces queremos reducir aún más esos datos, presentando **uno o dos valores que sean "representativos" de toda la serie**.
- Al usar un único valor se tiene una reducción drástica de los datos.
- Ya hemos usado dos medidas, mínimo y máximo, como medida de posición en un conjunto de datos.
- Usualmente, se emplea una de las siguientes medidas como **Medidas de Posición Central**:

MODA, MEDIANA y MEDIA (aritmética).

Análisis Exploratorio de Datos Unidimensionales

- Vimos que la reducción de datos a través de un gráfico o de una tabla de frecuencias nos da mucha más información sobre el comportamiento de una variable que el propio conjunto original de datos.
- Sin embargo, algunas veces queremos reducir aún más esos datos, presentando **uno o dos valores que sean "representativos" de toda la serie**.
- Al usar un único valor se tiene una reducción drástica de los datos.
- Ya hemos usado dos medidas, mínimo y máximo, como medida de posición en un conjunto de datos.
- Usualmente, se emplea una de las siguientes medidas como **Medidas de Posición Central**:

MODA, MEDIANA y MEDIA (aritmética).

Moda (Mo)

La Moda o Modo es la realización más frecuente del conjunto de valores observados.

- Por **ejemplo**, para la variable *"Número de Materias aprobadas"*, $\text{Mo} = 2$.
- **Observaciones:**
  1. La Moda es la única medida de tendencia central que podrán brindar si trabajan con varibles cualitativas.
  2. Una distribución de frecuencias unimodal es aquella que presenta un solo pico.
  3. En ocasiones una distribución de frecuencias tiene más de un pico; es bimodal si tiene dos picos y multimodal cuando presenta más de dos.
  4. Una distribución de frecuencias unimodal es **simétrica** si existe un eje que permita dividir al polígono de frecuencias absolutas en dos partes iguales que al superponerlas coincidan.

Moda (Mo)

La Moda o Modo es la realización más frecuente del conjunto de valores observados.

- Por **ejemplo**, para la variable *"Número de Materias aprobadas"*, Mo $= 2$.
- **Observaciones:**
  1. La Moda es la única medida de tendencia central que podrán brindar si trabajan con varibles cualitativas.
  2. Una distribución de frecuencias unimodal es aquella que presenta un solo pico.
  3. En ocasiones una distribución de frecuencias tiene más de un pico; es bimodal si tiene dos picos y multimodal cuando presenta más de dos.
  4. Una distribución de frecuencias unimodal es **simétrica** si existe un eje que permita dividir al polígono de frecuencias absolutas en dos partes iguales que al superponerlas coincidan.

Moda (Mo)

La Moda o Modo es la realización más frecuente del conjunto de valores observados.

- Por **ejemplo**, para la variable *"Número de Materias aprobadas"*, $\text{Mo} = 2$.
- **Observaciones:**
  1. La Moda es la única medida de tendencia central que podrán brindar si trabajan con varibles cualitativas.
  2. Una distribución de frecuencias unimodal es aquella que presenta un solo pico.
  3. En ocasiones una distribución de frecuencias tiene más de un pico; es bimodal si tiene dos picos y multimodal cuando presenta más de dos.
  4. Una distribución de frecuencias unimodal es **simétrica** si existe un eje que permita dividir al polígono de frecuencias absolutas en dos partes iguales que al superponerlas coincidan.

Moda (Mo)

La Moda o Modo es la realización más frecuente del conjunto de valores observados.

- Por **ejemplo**, para la variable *"Número de Materias aprobadas"*, Mo $= 2$.
- **Observaciones:**
  1. La Moda es la única medida de tendencia central que podrán brindar si trabajan con varibles cualitativas.
  2. Una distribución de frecuencias unimodal es aquella que presenta un solo pico.
  3. En ocasiones una distribución de frecuencias tiene más de un pico; es bimodal si tiene dos picos y multimodal cuando presenta más de dos.
  4. Una distribución de frecuencias unimodal es **simétrica** si existe un eje que permita dividir al polígono de frecuencias absolutas en dos partes iguales que al superponerlas coincidan.

Moda (Mo)

La Moda o Modo es la realización más frecuente del conjunto de valores observados.

- Por **ejemplo**, para la variable *"Número de Materias aprobadas"*, Mo $= 2$.
- **Observaciones:**
  1. La Moda es la única medida de tendencia central que podrán brindar si trabajan con varibles cualitativas.
  2. Una distribución de frecuencias unimodal es aquella que presenta un solo pico.
  3. En ocasiones una distribución de frecuencias tiene más de un pico; es bimodal si tiene dos picos y multimodal cuando presenta más de dos.
  4. Una distribución de frecuencias unimodal es **simétrica** si existe un eje que permita dividir al polígono de frecuencias absolutas en dos partes iguales que al superponerlas coincidan.

Moda (Mo)

La Moda o Modo es la realización más frecuente del conjunto de valores observados.

- Por **ejemplo**, para la variable *"Número de Materias aprobadas"*, Mo $= 2$.
- **Observaciones:**
  1. La Moda es la única medida de tendencia central que podrán brindar si trabajan con varibles cualitativas.
  2. Una distribución de frecuencias unimodal es aquella que presenta un solo pico.
  3. En ocasiones una distribución de frecuencias tiene más de un pico; es bimodal si tiene dos picos y multimodal cuando presenta más de dos.
  4. Una distribución de frecuencias unimodal es **simétrica** si existe un eje que permita dividir al polígono de frecuencias absolutas en dos partes iguales que al superponerlas coincidan.

Moda (Mo)

La Moda o Modo es la realización más frecuente del conjunto de valores observados.

- Por **ejemplo**, para la variable *"Número de Materias aprobadas"*, Mo $= 2$.
- **Observaciones:**
  1. La Moda es la única medida de tendencia central que podrán brindar si trabajan con varibles cualitativas.
  2. Una distribución de frecuencias unimodal es aquella que presenta un solo pico.
  3. En ocasiones una distribución de frecuencias tiene más de un pico; es bimodal si tiene dos picos y multimodal cuando presenta más de dos.
  4. Una distribución de frecuencias unimodal es **simétrica** si existe un eje que permita dividir al polígono de frecuencias absolutas en dos partes iguales que al superponerlas coincidan.

Análisis Exploratorio de Datos Unidimensionales

5. Al trabajar con datos reales, es poco frecuente que las distribuciones sean perfectamente simétricas, sin embargo, cuando se trata de caracterizar la forma de una distribución, si existen diferencias mínimas, éstas son ignoradas.
6. Una distribución de frecuencias unimodal es **asimétrica**, también denominada sesgada, si tiene un pico descentrado y una cola más larga que la otra.
7. Cuando la cola más larga apunta hacia la derecha se dice **asimétrica a derecha** o sesgo positivo y, en caso contrario, **asimétrica a izquierda** o sesgo negativo.
8. Para los datos de variables hidroquímicas de la Laguna Colorada Grande, la distribución de todas estas variables (recuerde, *"las concentraciones de calcio´´*) es asimétrica a derecha, mientras que la distribución de los *"Salarios anuales de directores de laboratorios (en miles de dólares)"* puede ser considerada simétrica.

Análisis Exploratorio de Datos Unidimensionales

5. Al trabajar con datos reales, es poco frecuente que las distribuciones sean perfectamente simétricas, sin embargo, cuando se trata de caracterizar la forma de una distribución, si existen diferencias mínimas, éstas son ignoradas.
6. Una distribución de frecuencias unimodal es **asimétrica**, también denominada sesgada, si tiene un pico descentrado y una cola más larga que la otra.
7. Cuando la cola más larga apunta hacia la derecha se dice **asimétrica a derecha** o sesgo positivo y, en caso contrario, **asimétrica a izquierda** o sesgo negativo.
8. Para los datos de variables hidroquímicas de la Laguna Colorada Grande, la distribución de todas estas variables (recuerde, *"las concentraciones de calcio´´*) es asimétrica a derecha, mientras que la distribución de los *"Salarios anuales de directores de laboratorios (en miles de dólares)"* puede ser considerada simétrica.

Análisis Exploratorio de Datos Unidimensionales

5. Al trabajar con datos reales, es poco frecuente que las distribuciones sean perfectamente simétricas, sin embargo, cuando se trata de caracterizar la forma de una distribución, si existen diferencias mínimas, éstas son ignoradas.
6. Una distribución de frecuencias unimodal es **asimétrica**, también denominada sesgada, si tiene un pico descentrado y una cola más larga que la otra.
7. Cuando la cola más larga apunta hacia la derecha se dice **asimétrica a derecha** o sesgo positivo y, en caso contrario, **asimétrica a izquierda** o sesgo negativo.
8. Para los datos de variables hidroquímicas de la Laguna Colorada Grande, la distribución de todas estas variables (recuerde, *"las concentraciones de calcio´´*) es asimétrica a derecha, mientras que la distribución de los *"Salarios anuales de directores de laboratorios (en miles de dólares)"* puede ser considerada simétrica.

Análisis Exploratorio de Datos Unidimensionales

5. Al trabajar con datos reales, es poco frecuente que las distribuciones sean perfectamente simétricas, sin embargo, cuando se trata de caracterizar la forma de una distribución, si existen diferencias mínimas, éstas son ignoradas.
6. Una distribución de frecuencias unimodal es **asimétrica**, también denominada sesgada, si tiene un pico descentrado y una cola más larga que la otra.
7. Cuando la cola más larga apunta hacia la derecha se dice **asimétrica a derecha** o sesgo positivo y, en caso contrario, **asimétrica a izquierda** o sesgo negativo.
8. Para los datos de variables hidroquímicas de la Laguna Colorada Grande, la distribución de todas estas variables (recuerde, *"las concentraciones de calcio´´*) es asimétrica a derecha, mientras que la distribución de los *"Salarios anuales de directores de laboratorios (en miles de dólares)"* puede ser considerada simétrica.

Mediana (Md)

La Mediana es la realización que ocupa la posición central de la serie de observaciones, cuando éstas están ordenadas de acuerdo a sus valores (creciente o decreciente).

- Por **ejemplo**, para la variable "Concentraciones de Calcio" de las aguas de la Laguna Colorada Grande, ordenando los datos a partir del Tallo y Hoja, tenemos:

[100 – 200)|52
[200 – 300)|28 36 60 76 80
[300 – 400)|04 08 20 24 24 32 44 44 44 48 48 48 52 56 56 72 76 80 88
[400 – 500)|48 60
[500 – 600)|60
[600 – 700)|
[700 – 800)|88
[800 – 900)|80

Mediana (Md)

La Mediana es la realización que ocupa la posición central de la serie de observaciones, cuando éstas están ordenadas de acuerdo a sus valores (creciente o decreciente).

- Por **ejemplo**, para la variable "Concentraciones de Calcio" de las aguas de la Laguna Colorada Grande, ordenando los datos a partir del Tallo y Hoja, tenemos:

$[100 - 200)|52$

$[200 - 300)|28\ 36\ 60\ 76\ 80$

$[300 - 400)|04\ 08\ 20\ 24\ 24\ 32\ 44\ 44\ 44\ 48\ 48\ 48\ 52\ 56\ 56\ 72\ 76\ 80\ 88$

$[400 - 500)|48\ 60$

$[500 - 600)|60$

$[600 - 700)|$

$[700 - 800)|88$

$[800 - 900)|80$

- La Mediana de este conjunto de datos, no es precisamente, un valor observado, sino que la se encontraria entre los valores 344 y 348 y, por ende, escogemos la $\text{Md}(X) = 346$.

Se define la **Mediana** de un conjunto de datos por:

$$\text{Md}(X) = \begin{cases} x_{(\frac{n+1}{2})} & si \quad n \text{ es impar} \\ \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & si \quad n \text{ es par} \end{cases}$$

- Para el **Ejemplo**, X :*"Número de materias aprobadas por un estudiante del curso de Estadística"*, comenzamos ordenando los datos:

$$0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 3\ 3\ 5$$

y determinamos que $Md(X) = 2$.

- La Mediana de este conjunto de datos, no es precisamente, un valor observado, sino que la se encontraria entre los valores 344 y 348 y, por ende, escogemos la $\text{Md}(X) = 346$.

Se define la **Mediana** de un conjunto de datos por:

$$\text{Md}(X) = \begin{cases} x_{(\frac{n+1}{2})} & si \quad n \text{ es impar} \\ \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & si \quad n \text{ es par} \end{cases}$$

- Para el **Ejemplo**, X :*"Número de materias aprobadas por un estudiante del curso de Estadística"*, comenzamos ordenando los datos:

$$0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 3\ 3\ 5$$

y determinamos que $Md(X) = 2$.

- La Mediana de este conjunto de datos, no es precisamente, un valor observado, sino que la se encontraria entre los valores 344 y 348 y, por ende, escogemos la $\text{Md}(X) = 346$.

Se define la **Mediana** de un conjunto de datos por:

$$\text{Md}(X) = \begin{cases} x_{(\frac{n+1}{2})} & si \quad n \text{ es impar} \\ \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & si \quad n \text{ es par} \end{cases}$$

- Para el **Ejemplo**, X :*"Número de materias aprobadas por un estudiante del curso de Estadística"*, comenzamos ordenando los datos:

$$0\ 0\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ 2\ 2\ 2\ 2\ 2\ 2\ 3\ 3\ 5$$

y determinamos que $Md(X) = 2$.

Media Aritmética: $\overline{x}$

La Media aritmética o, simplemente, media o promedio, es la suma de todos los valores observados divididos por el número total de ellos (o sea, por el número de elementos en la muestra).

La media se define por:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

- La **Media** de los valores de las *"Concentraciones de Calcio"* es:

$$\overline{x} = \frac{1}{30} (152 + 228 + 236 + \cdots + 788 + 880) = 371,2.$$

Media Aritmética: $\overline{x}$

La Media aritmética o, simplemente, media o promedio, es la suma de todos los valores observados divididos por el número total de ellos (o sea, por el número de elementos en la muestra).

La media se define por:

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

- La **Media** de los valores de las *"Concentraciones de Calcio"* es:

$$\overline{x} = \frac{1}{30} (152 + 228 + 236 + \cdots + 788 + 880) = 371,2.$$

Análisis Exploratorio de Datos Unidimensionales

- El resumen de un conjunto de datos, por medio de una **única** medida representativa de la posición central esconde toda la información sobre la variabilidad del conjunto de valores.
- Por **ejemplo**, supongamos que 5-grupos de alumnos se someten a un test, obteniendo las siguientes notas:

Grupo	Notas					Media
A	3	4	5	6	7	5
B	1	3	5	7	9	5
C	5	5	5	5	5	5
D	3	5	5	7		5
E	3,5	5	6,5			5

- Se observa que el promedio de las calificaciones es el mismo en cada grupo, por lo tanto, la identificación de cada uno de estos grupos de estudiantes a partir de su media, nada nos dice sobre la variabilidad de los mismos.

Análisis Exploratorio de Datos Unidimensionales

- El resumen de un conjunto de datos, por medio de una **única** medida representativa de la posición central esconde toda la información sobre la variabilidad del conjunto de valores.
- Por **ejemplo**, supongamos que 5-grupos de alumnos se someten a un test, obteniendo las siguientes notas:

Grupo	Notas					Media
A	3	4	5	6	7	5
B	1	3	5	7	9	5
C	5	5	5	5	5	5
D	3	5	5	7		5
E	3,5	5	6,5			5

- Se observa que el promedio de las calificaciones es el mismo en cada grupo, por lo tanto, la identificación de cada uno de estos grupos de estudiantes a partir de su media, nada nos dice sobre la variabilidad de los mismos.

Análisis Exploratorio de Datos Unidimensionales

- El resumen de un conjunto de datos, por medio de una **única** medida representativa de la posición central esconde toda la información sobre la variabilidad del conjunto de valores.
- Por **ejemplo**, supongamos que 5-grupos de alumnos se someten a un test, obteniendo las siguientes notas:

Grupo	Notas					Media
A	3	4	5	6	7	5
B	1	3	5	7	9	5
C	5	5	5	5	5	5
D	3	5	5	7		5
E	3,5	5	6,5			5

- Se observa que el promedio de las calificaciones es el mismo en cada grupo, por lo tanto, la identificación de cada uno de estos grupos de estudiantes a partir de su media, nada nos dice sobre la variabilidad de los mismos.

Análisis Exploratorio de Datos Unidimensionales

- Notamos entonces, la conveniencia de **crear una medida que resuma la variabilidad de una serie de datos** y que nos permita, por ejemplo, **comparar conjuntos diferentes de valores** como los del ejemplo anterior, según algún criterio establecido.
- El **criterio** frecuentemente usado es aquel que **mide la concentración de los datos alrededor de su media**.
- El principio básico es analizar los desvíos de las observaciones en relación a la media de las observación.
- Por **ejemplo:**
  - Para el Grupo A, los desvíos son: $-2; -1; 0; 1; 2$;
  - Para el Grupo B, los desvíos son: $-4; -2; 0; 2; 4$;
  - Para el Grupo C, los desvíos son: $0; 0; 0; 0; 0$;
  - Para el Grupo D, los desvíos son: $-2; 0; 0; 2$;
  - Para el Grupo E, los desvíos son: $-1,5; 0; 1,5$.

Análisis Exploratorio de Datos Unidimensionales

- Notamos entonces, la conveniencia de **crear una medida que resuma la variabilidad de una serie de datos** y que nos permita, por ejemplo, **comparar conjuntos diferentes de valores** como los del ejemplo anterior, según algún criterio establecido.
- El **criterio** frecuentemente usado es aquel que **mide la concentración de los datos alrededor de su media**.
- El principio básico es analizar los desvíos de las observaciones en relación a la media de las observación.
- Por **ejemplo:**
  - Para el Grupo A, los desvíos son: $-2; -1; 0; 1; 2$;
  - Para el Grupo B, los desvíos son: $-4; -2; 0; 2; 4$;
  - Para el Grupo C, los desvíos son: $0; 0; 0; 0; 0$;
  - Para el Grupo D, los desvíos son: $-2; 0; 0; 2$;
  - Para el Grupo E, los desvíos son: $-1,5; 0; 1,5$.

Análisis Exploratorio de Datos Unidimensionales

- Notamos entonces, la conveniencia de **crear una medida que resuma la variabilidad de una serie de datos** y que nos permita, por ejemplo, **comparar conjuntos diferentes de valores** como los del ejemplo anterior, según algún criterio establecido.
- El **criterio** frecuentemente usado es aquel que **mide la concentración de los datos alrededor de su media**.
- El principio básico es analizar los desvíos de las observaciones en relación a la media de las observación.
- Por **ejemplo:**
  - Para el Grupo A, los desvíos son: $-2; -1; 0; 1; 2$;
  - Para el Grupo B, los desvíos son: $-4; -2; 0; 2; 4$;
  - Para el Grupo C, los desvíos son: $0; 0; 0; 0; 0$;
  - Para el Grupo D, los desvíos son: $-2; 0; 0; 2$;
  - Para el Grupo E, los desvíos son: $-1,5; 0; 1,5$.

Análisis Exploratorio de Datos Unidimensionales

- Notamos entonces, la conveniencia de **crear una medida que resuma la variabilidad de una serie de datos** y que nos permita, por ejemplo, **comparar conjuntos diferentes de valores** como los del ejemplo anterior, según algún criterio establecido.
- El **criterio** frecuentemente usado es aquel que **mide la concentración de los datos alrededor de su media**.
- El principio básico es analizar los desvíos de las observaciones en relación a la media de las observación.
- Por **ejemplo:**
  - Para el Grupo A, los desvíos son: $-2; -1; 0; 1; 2;$
  - Para el Grupo B, los desvíos son: $-4; -2; 0; 2; 4;$
  - Para el Grupo C, los desvíos son: $0; 0; 0; 0; 0;$
  - Para el Grupo D, los desvíos son: $-2; 0; 0; 2;$
  - Para el Grupo E, los desvíos son: $-1,5; 0; 1,5.$

- Entonces, cualquiera sea el conjunto de datos, **la suma de los desvíos, SIEMPRE, es CERO**, i.e. $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$, y por lo tanto, $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$ **NO ES UNA BUENA MEDIDA DE DISPERSIÓN**.

- Surgen, entonces, **dos** opciones a considerar:
  - El Total de los Desvíos en Valor Alboluto, i.e. $\sum_{i=1}^{n}|x_i - \overline{x}|$;
  - El total de los Cuadrados de los Desvíos, i.e. $\sum_{i=1}^{n}(x_i - \overline{x})^2$.

- Entonces, cualquiera sea el conjunto de datos, **la suma de los desvíos, SIEMPRE, es CERO**, i.e. $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$, y por lo tanto, $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$ **NO ES UNA BUENA MEDIDA DE DISPERSIÓN**.

- Surgen, entonces, **dos** opciones a considerar:
  - El Total de los Desvíos en Valor Alboluto, i.e. $\sum_{i=1}^{n}|x_i - \overline{x}|$;
  - El total de los Cuadrados de los Desvíos, i.e. $\sum_{i=1}^{n}(x_i - \overline{x})^2$.

- Estos totales aplicados al grupo de estudiantes, son:

Grupo	$\sum_{i=1}^{n} \lvert x_i - \overline{x} \rvert$	$\sum_{i=1}^{n} (x_i - \overline{x})^2$
A	2+1+0+1+2= $\boxed{6}$	4+1+0+1+4= $\boxed{10}$
B	4+2+0+2+4= $\boxed{12}$	16+4+0+4+16= $\boxed{40}$
C	0+0+0+0+0 = $\boxed{0}$	0+0+0+0+0 = $\boxed{0}$
D	2+0+0+2 = $\boxed{4}$	4+0+0+4= $\boxed{8}$
E	1,5+0+1,5 = $\boxed{3}$	2,25+0+0+2,25= $\boxed{4,5}$

- El uso de estos totales causa dificultades cuando comparamos conjuntos de datos con un número diferente de observaciones.

- Estos totales aplicados al grupo de estudiantes, son:

Grupo	$\sum_{i=1}^{n} \lvert x_i - \overline{x} \rvert$	$\sum_{i=1}^{n} (x_i - \overline{x})^2$
A	2+1+0+1+2= $\boxed{6}$	4+1+0+1+4= $\boxed{10}$
B	4+2+0+2+4= $\boxed{12}$	16+4+0+4+16= $\boxed{40}$
C	0+0+0+0+0 = $\boxed{0}$	0+0+0+0+0 = $\boxed{0}$
D	2+0+0+2 = $\boxed{4}$	4+0+0+4= $\boxed{8}$
E	1,5+0+1,5 = $\boxed{3}$	2,25+0+0+2,25= $\boxed{4,5}$

- El uso de estos totales causa dificultades cuando comparamos conjuntos de datos con un número diferente de observaciones.

- Por lo tanto, usamos las medias o promedios de estas medidas, a saber:

$$\text{Desvío Medio} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|.$$

$$\text{Varianza} = \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

- Para el ejemplo, tenemos:

Medidas	A	B	C	D	E
Desvío Medio	1,2	2,4	0	1,0	1,0
Varianza	2,0	8,0	0	2,0	1,5

- **Observe que:**
  - Según ambas medidas, C es el **más homogéneo** y B el **más heterogéneo**.
  - Según el Desvío Medio, D es **más homogéneo** que A, mientras que, según la Varianza, son **igualmente homogéneos**.

- Por lo tanto, usamos las medias o promedios de estas medidas, a saber:

$$\text{Desvío Medio} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|.$$

$$\text{Varianza} = \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

- Para el ejemplo, tenemos:

Medidas	A	B	C	D	E
Desvío Medio	1,2	2,4	0	1,0	1,0
Varianza	2,0	8,0	0	2,0	1,5

- **Observe que:**
  - Según ambas medidas, C es el **más homogéneo** y B el **más heterogéneo**.
  - Según el Desvío Medio, D es **más homogéneo** que A, mientras que, según la Varianza, son **igualmente homogéneos**.

- Por lo tanto, usamos las medias o promedios de estas medidas, a saber:

$$\text{Desvío Medio} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}|.$$

$$\text{Varianza} = \widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

- Para el ejemplo, tenemos:

Medidas	A	B	C	D	E
Desvío Medio	1,2	2,4	0	1,0	1,0
Varianza	2,0	8,0	0	2,0	1,5

- **Observe que:**

  • Según ambas medidas, C es el **más homogéneo** y B el **más heterogéneo**.

  • Según el Desvío Medio, D es **más homogéneo** que A, mientras que, según la Varianza, son **igualmente homogéneos**.

Análisis Exploratorio de Datos Unidimensionales

- Siendo la **Varianza** una medida que expresa un desvío cuadrático medio, puede causar algunos problemas de interpretación.
- Para evitar ésto, se acostumbra usar el **DESVÍO ESTÁNDAR**, definido por:

$$D.E.(X) = \widehat{\sigma_X} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

- Así, para cada uno de los grupos de alumnos, tenemos:

$$\widehat{\sigma_A} = 1,41 \;\; ; \;\; \widehat{\sigma_B} = 2,83 \;\; ; \;\; \widehat{\sigma_C} = 0 \;\; ; \;\; \widehat{\sigma_D} = 1,41 \;\; ; \;\; \widehat{\sigma_E} = 1,22.$$

- **Observación 1:** Con el "Desvío Estándar" tenemos una medida expresada en la misma unidad de medida de los valores del conjunto de datos original.

Análisis Exploratorio de Datos Unidimensionales

- Siendo la **Varianza** una medida que expresa un desvío cuadrático medio, puede causar algunos problemas de interpretación.
- Para evitar ésto, se acostumbra usar el **DESVÍO ESTÁNDAR**, definido por:

$$D.E.(X) = \widehat{\sigma_X} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

- Así, para cada uno de los grupos de alumnos, tenemos:

$$\widehat{\sigma_A} = 1,41 \;\; ; \;\; \widehat{\sigma_B} = 2,83 \;\; ; \;\; \widehat{\sigma_C} = 0 \;\; ; \;\; \widehat{\sigma_D} = 1,41 \;\; ; \;\; \widehat{\sigma_E} = 1,22.$$

- **Observación 1:** Con el "Desvío Estándar" tenemos una medida expresada en la misma unidad de medida de los valores del conjunto de datos original.

- Siendo la **Varianza** una medida que expresa un desvío cuadrático medio, puede causar algunos problemas de interpretación.
- Para evitar ésto, se acostumbra usar el **DESVÍO ESTÁNDAR**, definido por:

$$D.E.(X) = \widehat{\sigma_X} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

- Así, para cada uno de los grupos de alumnos, tenemos:

$$\widehat{\sigma_A} = 1,41 \;\; ; \;\; \widehat{\sigma_B} = 2,83 \;\; ; \;\; \widehat{\sigma_C} = 0 \;\; ; \;\; \widehat{\sigma_D} = 1,41 \;\; ; \;\; \widehat{\sigma_E} = 1,22.$$

- **Observación 1:** Con el "Desvío Estándar" tenemos una medida expresada en la misma unidad de medida de los valores del conjunto de datos original.

- Siendo la **Varianza** una medida que expresa un desvío cuadrático medio, puede causar algunos problemas de interpretación.
- Para evitar ésto, se acostumbra usar el **DESVÍO ESTÁNDAR**, definido por:

$$D.E.(X) = \widehat{\sigma_X} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

- Así, para cada uno de los grupos de alumnos, tenemos:

$$\widehat{\sigma_A} = 1,41 \;\; ; \;\; \widehat{\sigma_B} = 2,83 \;\; ; \;\; \widehat{\sigma_C} = 0 \;\; ; \;\; \widehat{\sigma_D} = 1,41 \;\; ; \;\; \widehat{\sigma_E} = 1,22.$$

- **Observación 1:** Con el "Desvío Estándar" tenemos una medida expresada en la misma unidad de medida de los valores del conjunto de datos original.

- **Observación 2:** Si los datos originales han sido agrupados, se usa el punto medio de cada clase como representativo de esa clase y se trabaja con las expresiones anteriores como si fuera una variable discreta, obteniéndose expresiones aproximadas de media y varianza, i.e.:

$$\overline{x} \approx \frac{1}{n}\sum_{i=1}^{k} x_i' n_i = \sum_{i=1}^{k} x_i' f_i$$

$$\widehat{\sigma_i^2} \approx \frac{1}{n}\sum_{i=1}^{k} n_i (x_i' - \overline{x})^2$$

donde:

- k: es el N° de Intervalos o clases en que se agrupa la variable;
- x_i': es el punto medio de la clase i;
- n_i: es la frecuencias observada en la clase i; y
- f_i: es la frecuencia relativa de la clase i.

- **Observación 3:** Puede usarse la expresión alternativa, conocida como "**varianza muestral**´´:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Existen fundamentos teoricos para el uso de la misma, que serán desarrollados en la segunda parte de este curso.

Observe además que, si n es grande, no debieran existir diferencias numéricas importantes entre $\widehat{\sigma_X^2}$ y s_X^2.

Análisis Exploratorio de Datos Unidimensionales

- **Observación 4:** Cuando la media es una buena medida de posición, y por ende, el desvío una buena medida de la variabilidad, suele utilizarse, también, una medida de la variabilidad relativa, conocida como **"Coeficiente de Variación"**: $CV(X) = \frac{s}{\bar{x}}$
  1. El Coeficiente de Variación no posee unidades aunque se lo suele expresar en forma porcentual.
  2. Este cociente da cuenta de la desviación estándar como una proporción de la media, y es a veces un indicador bastante útil.
  3. Por ejemplo, un valor $s = 10$ no es significativo, a menos que se lo compare con algo diferente.
     - Si $s = 10$ y $\bar{x} = 1000$, entonces la variación es muy pequeña con respecto a la media.
     - Sin embargo, si $s = 10$ y $\bar{x} = 5$, la variación con respecto a la media es grande.
     - Por ejemplo, si se estudia la precisión (variación en mediciones repetidas) de un instrumento de medición, el primer caso $CV = \frac{10}{1000} = 0,01$ presentaría una precisión aceptable, pero en el segundo caso $CV = \frac{10}{5} = 2$ sería totalmente inaceptable.

Análisis Exploratorio de Datos Unidimensionales

- **Observación 4:** Cuando la media es una buena medida de posición, y por ende, el desvío una buena medida de la variabilidad, suele utilizarse, también, una medida de la variabilidad relativa, conocida como **"Coeficiente de Variación"**: $CV(X) = \frac{s}{\bar{x}}$
  1. El Coeficiente de Variación no posee unidades aunque se lo suele expresar en forma porcentual.
  2. Este cociente da cuenta de la desviación estándar como una proporción de la media, y es a veces un indicador bastante útil.
  3. Por ejemplo, un valor $s = 10$ no es significativo, a menos que se lo compare con algo diferente.
     - Si $s = 10$ y $\bar{x} = 1000$, entonces la variación es muy pequeña con respecto a la media.
     - Sin embargo, si $s = 10$ y $\bar{x} = 5$, la variación con respecto a la media es grande.
     - Por ejemplo, si se estudia la precisión (variación en mediciones repetidas) de un instrumento de medición, el primer caso $CV = \frac{10}{1000} = 0,01$ presentaría una precisión aceptable, pero en el segundo caso $CV = \frac{10}{5} = 2$ sería totalmente inaceptable.

Análisis Exploratorio de Datos Unidimensionales

- **Observación 4:** Cuando la media es una buena medida de posición, y por ende, el desvío una buena medida de la variabilidad, suele utilizarse, también, una medida de la variabilidad relativa, conocida como **"Coeficiente de Variación"**: $CV(X) = \frac{s}{\bar{x}}$
  1. El Coeficiente de Variación no posee unidades aunque se lo suele expresar en forma porcentual.
  2. Este cociente da cuenta de la desviación estándar como una proporción de la media, y es a veces un indicador bastante útil.
  3. Por ejemplo, un valor $s = 10$ no es significativo, a menos que se lo compare con algo diferente.
     - Si $s = 10$ y $\bar{x} = 1000$, entonces la variación es muy pequeña con respecto a la media.
     - Sin embargo, si $s = 10$ y $\bar{x} = 5$, la variación con respecto a la media es grande.
     - Por ejemplo, si se estudia la precisión (variación en mediciones repetidas) de un instrumento de medición, el primer caso $CV = \frac{10}{1000} = 0,01$ presentaría una precisión aceptable, pero en el segundo caso $CV = \frac{10}{5} = 2$ sería totalmente inaceptable.

Análisis Exploratorio de Datos Unidimensionales

- **Observación 4:** Cuando la media es una buena medida de posición, y por ende, el desvío una buena medida de la variabilidad, suele utilizarse, también, una medida de la variabilidad relativa, conocida como **"Coeficiente de Variación"**: $CV(X) = \frac{s}{\bar{x}}$
  1. El Coeficiente de Variación no posee unidades aunque se lo suele expresar en forma porcentual.
  2. Este cociente da cuenta de la desviación estándar como una proporción de la media, y es a veces un indicador bastante útil.
  3. Por ejemplo, un valor $s = 10$ no es significativo, a menos que se lo compare con algo diferente.
     • Si $s = 10$ y $\bar{x} = 1000$, entonces la variación es muy pequeña con respecto a la media.
     • Sin embargo, si $s = 10$ y $\bar{x} = 5$, la variación con respecto a la media es grande.
     • Por ejemplo, si se estudia la precisión (variación en mediciones repetidas) de un instrumento de medición, el primer caso $CV = \frac{10}{1000} = 0,01$ presentaría una precisión aceptable, pero en el segundo caso $CV = \frac{10}{5} = 2$ sería totalmente inaceptable.

Análisis Exploratorio de Datos Unidimensionales

4. En el caso de datos de campo o de laboratorio, el coeficiente de variación refleja una mezcla desconocida de la variabilidad natural, **la variabilidad introducida durante el proceso de muestreo y de causas aleatorias**.
5. El coeficiente de variación de una **población homogénea** es típicamente menor que la unidad.
6. Si es mayor que $1,5$ conviene investigar posibles fuentes de **heterogeneidad en los datos**, también puede indicar la **existencia de valores extremos**.
7. A pesar de esto, en numerosas variables geológicas de las ciencias experimentales (geología, por ejemplo) el coeficiente de variación toma valores entre $0,2$ y $2,5$.
8. El coeficiente de variación también resulta útil para **comparar la variabilidad entre varias muestras**, incluso **la variabilidad entre mediciones realizadas en diferentes unidades**.
9. Pero cuando la media es cercana a cero **no es útil calcular el coeficiente de variación**.

Análisis Exploratorio de Datos Unidimensionales

4. En el caso de datos de campo o de laboratorio, el coeficiente de variación refleja una mezcla desconocida de la variabilidad natural, **la variabilidad introducida durante el proceso de muestreo y de causas aleatorias**.
5. El coeficiente de variación de una **población homogénea** es típicamente menor que la unidad.
6. Si es mayor que $1,5$ conviene investigar posibles fuentes de **heterogeneidad en los datos**, también puede indicar la **existencia de valores extremos**.
7. A pesar de esto, en numerosas variables geológicas de las ciencias experimentales (geología, por ejemplo) el coeficiente de variación toma valores entre $0,2$ y $2,5$.
8. El coeficiente de variación también resulta útil para **comparar la variabilidad entre varias muestras**, incluso **la variabilidad entre mediciones realizadas en diferentes unidades**.
9. Pero cuando la media es cercana a cero **no es útil calcular el coeficiente de variación**.

Análisis Exploratorio de Datos Unidimensionales

4. En el caso de datos de campo o de laboratorio, el coeficiente de variación refleja una mezcla desconocida de la variabilidad natural, **la variabilidad introducida durante el proceso de muestreo y de causas aleatorias**.
5. El coeficiente de variación de una **población homogénea** es típicamente menor que la unidad.
6. Si es mayor que $1,5$ conviene investigar posibles fuentes de **heterogeneidad en los datos**, también puede indicar la **existencia de valores extremos**.
7. A pesar de esto, en numerosas variables geológicas de las ciencias experimentales (geología, por ejemplo) el coeficiente de variación toma valores entre $0,2$ y $2,5$.
8. El coeficiente de variación también resulta útil para **comparar la variabilidad entre varias muestras**, incluso **la variabilidad entre mediciones realizadas en diferentes unidades**.
9. Pero cuando la media es cercana a cero **no es útil calcular el coeficiente de variación**.

Análisis Exploratorio de Datos Unidimensionales

4. En el caso de datos de campo o de laboratorio, el coeficiente de variación refleja una mezcla desconocida de la variabilidad natural, **la variabilidad introducida durante el proceso de muestreo y de causas aleatorias**.
5. El coeficiente de variación de una **población homogénea** es típicamente menor que la unidad.
6. Si es mayor que $1,5$ conviene investigar posibles fuentes de **heterogeneidad en los datos**, también puede indicar la **existencia de valores extremos**.
7. A pesar de esto, en numerosas variables geológicas de las ciencias experimentales (geología, por ejemplo) el coeficiente de variación toma valores entre $0,2$ y $2,5$.
8. El coeficiente de variación también resulta útil para **comparar la variabilidad entre varias muestras**, incluso **la variabilidad entre mediciones realizadas en diferentes unidades**.
9. Pero cuando la media es cercana a cero **no es útil calcular el coeficiente de variación**.

Análisis Exploratorio de Datos Unidimensionales

4. En el caso de datos de campo o de laboratorio, el coeficiente de variación refleja una mezcla desconocida de la variabilidad natural, **la variabilidad introducida durante el proceso de muestreo y de causas aleatorias**.
5. El coeficiente de variación de una **población homogénea** es típicamente menor que la unidad.
6. Si es mayor que $1,5$ conviene investigar posibles fuentes de **heterogeneidad en los datos**, también puede indicar la **existencia de valores extremos**.
7. A pesar de esto, en numerosas variables geológicas de las ciencias experimentales (geología, por ejemplo) el coeficiente de variación toma valores entre $0,2$ y $2,5$.
8. El coeficiente de variación también resulta útil para **comparar la variabilidad entre varias muestras**, incluso **la variabilidad entre mediciones realizadas en diferentes unidades**.
9. Pero cuando la media es cercana a cero **no es útil calcular el coeficiente de variación**.

Análisis Exploratorio de Datos Unidimensionales

4. En el caso de datos de campo o de laboratorio, el coeficiente de variación refleja una mezcla desconocida de la variabilidad natural, **la variabilidad introducida durante el proceso de muestreo y de causas aleatorias**.
5. El coeficiente de variación de una **población homogénea** es típicamente menor que la unidad.
6. Si es mayor que $1,5$ conviene investigar posibles fuentes de **heterogeneidad en los datos**, también puede indicar la **existencia de valores extremos**.
7. A pesar de esto, en numerosas variables geológicas de las ciencias experimentales (geología, por ejemplo) el coeficiente de variación toma valores entre $0,2$ y $2,5$.
8. El coeficiente de variación también resulta útil para **comparar la variabilidad entre varias muestras**, incluso **la variabilidad entre mediciones realizadas en diferentes unidades**.
9. Pero cuando la media es cercana a cero **no es útil calcular el coeficiente de variación**.

- **Observación 5:** Si los datos muestran una distribución simétrica, la *"media muestral"* $\overline{x}$ es una buena medida de posición central, y por lo tanto, el *"desvío estándar"* $\widehat{\sigma_X}$ es una buena medida de dispersión.

  Sin embargo, en distribuciones asimétricas, la media muestral se ve afectada por valores extremos, por lo que es necesario buscar otras medidas (tanto de posición central, como de dispersión) representativas del conjunto de datos.

Análisis Exploratorio de Datos Unidimensionales

- **Ejemplo:** Para comprobar si la tolerancia a la glucosa en sujetos sanos tiende a decrecer con la edad, se realizó un test oral de glucosa a dos muestras de pacientes sanos, una de ellas correspondiente a sujetos jóvenes y la otra correspondiente a adultos. Dicho test consistió en medir el nivel de glucosa en sangre en el momento de la ingestión (nivel basal) de 100 gramos de glucosa y a los 60 minutos de la toma.
- Los resultados correspondientes a esta última medición fueron los siguientes:

Nivel de glucosa a los 60 minutos de la ingestión

Jóvenes	135	136	138	141	141	142	144	145	145	145	147	147	149	150	154	156
Adultos	170	176	182	182	185	187	189	189	190	190	191	192	192	193	196	197

Análisis Exploratorio de Datos Unidimensionales

- **Ejemplo:** Para comprobar si la tolerancia a la glucosa en sujetos sanos tiende a decrecer con la edad, se realizó un test oral de glucosa a dos muestras de pacientes sanos, una de ellas correspondiente a sujetos jóvenes y la otra correspondiente a adultos. Dicho test consistió en medir el nivel de glucosa en sangre en el momento de la ingestión (nivel basal) de 100 gramos de glucosa y a los 60 minutos de la toma.
- Los resultados correspondientes a esta última medición fueron los siguientes:

Nivel de glucosa a los 60 minutos de la ingestión

Jóvenes	135	136	138	141	141	142	144	145	145	145	147	147	149	150	154	156
Adultos	170	176	182	182	185	187	189	189	190	190	191	192	192	193	196	197

Análisis Exploratorio de Datos Unidimensionales

- Para la variable, cuantitativa **continua**, *X* : *"Nivel de glucosa en sangre a los 60 minutos de la ingestión de 100 gramos de glucosa´´*, medida en las poblaciones tanto de jóvenes como de adultos, se construyeron las siguientes Tablas de Distribución de Frecuencias, considerando clases de amplitud 5 a partir del mínimo:

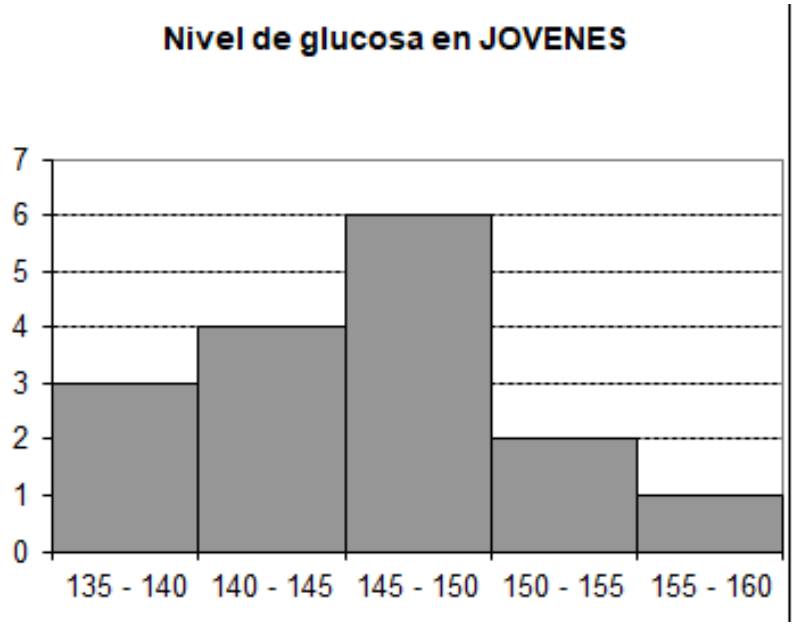
Tabla de distribución de frecuencias del nivel de glucosa a los 60 minutos de la ingestión en sujetos jóvenes

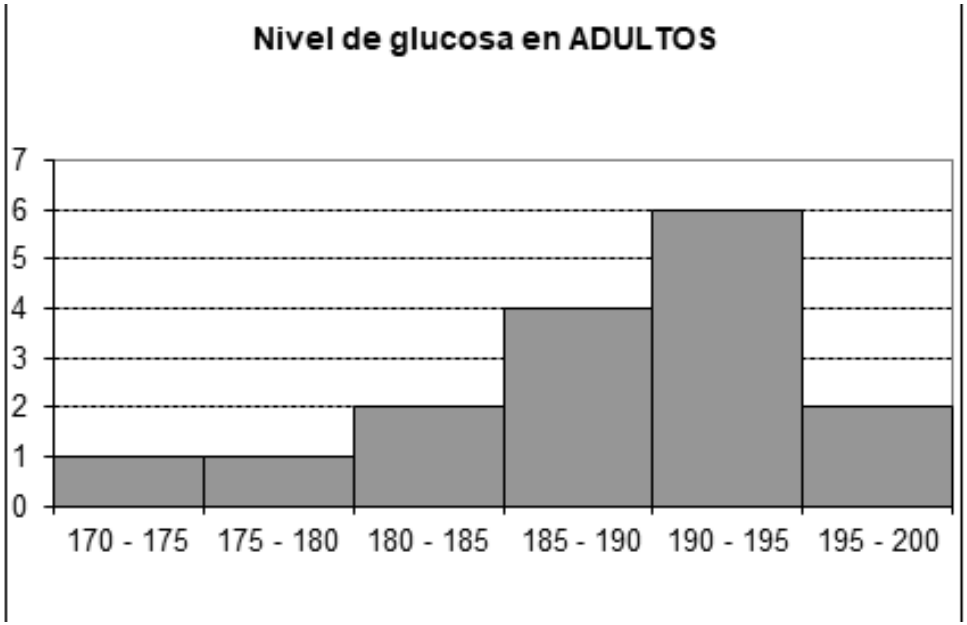
Clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia relativa acumulada
135 - 140	3	0,1875	0,1875
140 - 145	4	0,2500	0,4375
145 - 150	6	0,3750	0,8125
150 - 155	2	0,1250	0,9375
155 - 160	1	0,0625	1,0000
TOTAL	16	1	

Tabla de distribución de frecuencias del nivel de glucosa a los 60 minutos de la ingestión en sujetos adultos

Clase	Frecuencia absoluta	Frecuencia relativa	Frecuencia relativa acumulada
170 - 175	1	0,0625	0,0625
175 - 180	1	0,0625	0,1250
180 - 185	2	0,1250	0,2500
185 - 190	4	0,2500	0,5000
190 - 195	6	0,3750	0,8750
195 - 200	2	0,1250	1,0000
TOTAL	16	1,0000	

- Los respectivos Histogramas son:

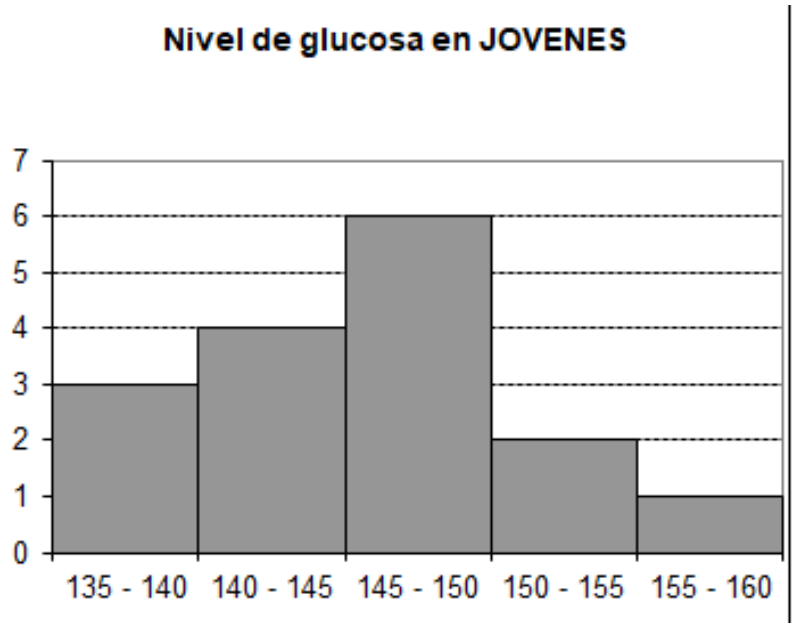


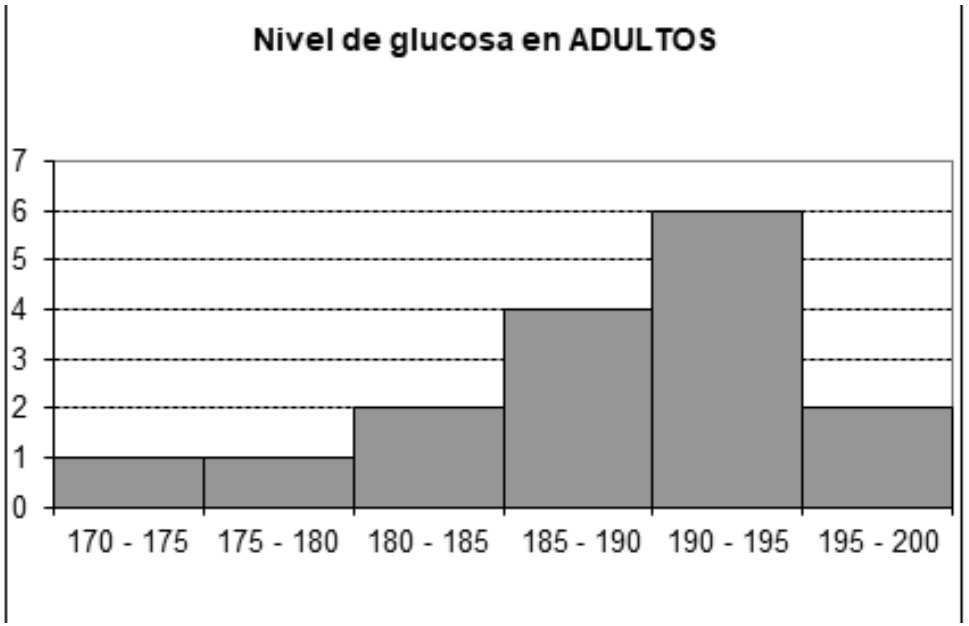


- Puede observarse una leve asimetría hacia valores grandes (sesgada a derecha) en los jóvenes, mientras que la asimetría parece más pronunciada, pero hacia la izquierda (sesgada a izquierda), en los adultos.

Análisis Exploratorio de Datos Unidimensionales

- Los respectivos Histogramas son:





- Puede observarse una leve asimetría hacia valores grandes (sesgada a derecha) en los jóvenes, mientras que la asimetría parece más pronunciada, pero hacia la izquierda (sesgada a izquierda), en los adultos.

Análisis Exploratorio de Datos Unidimensionales

- Algunas Medidas Resumen para esta variable, discriminadas para las poblaciones de jóvenes y adultos, son:

Medidas	Jóvenes	Adultos
Mínimo	135	170
Media	144,6875	187,5625
Desvío	5,5275	6,7079
Mediana	145	189,5
Máximo	156	197

- De acuerdo a lo observado previamente, debiera existir otra medida de posición central (y de dispersión), mejor que la media (y el desvío) para este conjunto de datos.
- Cuando los datos presentan una distribución asimétrica, la mejor medida de posición central es la **MEDIANA**, y entonces, **MAD** (**M**edian **A**bsolute **D**eviation, por sus siglas en inglés) en la mejor medida de dispersión.

Análisis Exploratorio de Datos Unidimensionales

- Algunas Medidas Resumen para esta variable, discriminadas para las poblaciones de jóvenes y adultos, son:

Medidas	Jóvenes	Adultos
Mínimo	135	170
Media	144,6875	187,5625
Desvío	5,5275	6,7079
Mediana	145	189,5
Máximo	156	197

- De acuerdo a lo observado previamente, debiera existir otra medida de posición central (y de dispersión), mejor que la media (y el desvío) para este conjunto de datos.
- Cuando los datos presentan una distribución asimétrica, la mejor medida de posición central es la **MEDIANA**, y entonces, **MAD** (**M**edian **A**bsolute **D**eviation, por sus siglas en inglés) en la mejor medida de dispersión.

Análisis Exploratorio de Datos Unidimensionales

- Algunas Medidas Resumen para esta variable, discriminadas para las poblaciones de jóvenes y adultos, son:

Medidas	Jóvenes	Adultos
Mínimo	135	170
Media	144,6875	187,5625
Desvío	5,5275	6,7079
Mediana	145	189,5
Máximo	156	197

- De acuerdo a lo observado previamente, debiera existir otra medida de posición central (y de dispersión), mejor que la media (y el desvío) para este conjunto de datos.
- Cuando los datos presentan una distribución asimétrica, la mejor medida de posición central es la **MEDIANA**, y entonces, **MAD** (**M**edian **A**bsolute **D**eviation, por sus siglas en inglés) en la mejor medida de dispersión.

Análisis Exploratorio de Datos Unidimensionales

- MAD se define por:

$$\text{MAD} = \text{Md}\left(|x_i - \text{Md}(X)|\right).$$

- Para calcular MAD, se procede de la siguiente manera:
  1. Ordenar el conjunto de datos originales y determinar la mediana, i.e. $\text{Md}(X)$;
  2. Calcular las n-desviaciones de cada dato a la $\text{Md}(X)$, i.e. $x_i - \text{Md}(X)$;
  3. Tomar el valor absoluto de las n-desviaciones, i.e. $|x_i - \text{Md}(X)|$;
  4. Re-ordenar los valores absolutos de las desviaciones (de menor a mayor);
  5. Encontrar la mediana de los valores absolutos de las desviaciones ordenadas. Esta Mediana es MAD.
- Para el ejemplo, para los jóvenes $\text{MAD} = 5,9391$ y para los adultos $\text{MAD} = 4,4593$.

Análisis Exploratorio de Datos Unidimensionales

- MAD se define por:

$$\text{MAD} = \text{Md}\left(|x_i - \text{Md}(X)|\right).$$

- Para calcular MAD, se procede de la siguiente manera:
  1. Ordenar el conjunto de datos originales y determinar la mediana, i.e. $\text{Md}(X)$;
  2. Calcular las n-desviaciones de cada dato a la $\text{Md}(X)$, i.e. $x_i - \text{Md}(X)$;
  3. Tomar el valor absoluto de las n-desviaciones, i.e. $|x_i - \text{Md}(X)|$;
  4. Re-ordenar los valores absolutos de las desviaciones (de menor a mayor);
  5. Encontrar la mediana de los valores absolutos de las desviaciones ordenadas. Esta Mediana es MAD.
- Para el ejemplo, para los jóvenes $\text{MAD} = 5,9391$ y para los adultos $\text{MAD} = 4,4593$.

Análisis Exploratorio de Datos Unidimensionales

- MAD se define por:

$$\text{MAD} = \text{Md}\left(|x_i - \text{Md}(X)|\right).$$

- Para calcular MAD, se procede de la siguiente manera:
  1. Ordenar el conjunto de datos originales y determinar la mediana, i.e. $\text{Md}(X)$;
  2. Calcular las n-desviaciones de cada dato a la $\text{Md}(X)$, i.e. $x_i - \text{Md}(X)$;
  3. Tomar el valor absoluto de las n-desviaciones, i.e. $|x_i - \text{Md}(X)|$;
  4. Re-ordenar los valores absolutos de las desviaciones (de menor a mayor);
  5. Encontrar la mediana de los valores absolutos de las desviaciones ordenadas. Esta Mediana es MAD.
- Para el ejemplo, para los jóvenes $\text{MAD} = 5,9391$ y para los adultos $\text{MAD} = 4,4593$.

Análisis Exploratorio de Datos Unidimensionales

- MAD se define por:

$$\text{MAD} = \text{Md}\left(|x_i - \text{Md}(X)|\right).$$

- Para calcular MAD, se procede de la siguiente manera:
  1. Ordenar el conjunto de datos originales y determinar la mediana, i.e. $\text{Md}(X)$;
  2. Calcular las n-desviaciones de cada dato a la $\text{Md}(X)$, i.e. $x_i - \text{Md}(X)$;
  3. Tomar el valor absoluto de las n-desviaciones, i.e. $|x_i - \text{Md}(X)|$;
  4. Re-ordenar los valores absolutos de las desviaciones (de menor a mayor);
  5. Encontrar la mediana de los valores absolutos de las desviaciones ordenadas. Esta Mediana es MAD.
- Para el ejemplo, para los jóvenes $\text{MAD} = 5,9391$ y para los adultos $\text{MAD} = 4,4593$.

- MAD se define por:

$$\text{MAD} = \text{Md}\left(|x_i - \text{Md}(X)|\right).$$

- Para calcular MAD, se procede de la siguiente manera:
  1. Ordenar el conjunto de datos originales y determinar la mediana, i.e. $\text{Md}(X)$;
  2. Calcular las n-desviaciones de cada dato a la $\text{Md}(X)$, i.e. $x_i - \text{Md}(X)$;
  3. Tomar el valor absoluto de las n-desviaciones, i.e. $|x_i - \text{Md}(X)|$;
  4. Re-ordenar los valores absolutos de las desviaciones (de menor a mayor);
  5. Encontrar la mediana de los valores absolutos de las desviaciones ordenadas. Esta Mediana es MAD.
- Para el ejemplo, para los jóvenes $\text{MAD} = 5,9391$ y para los adultos $\text{MAD} = 4,4593$.

- MAD se define por:

$$\text{MAD} = \text{Md}\left(|x_i - \text{Md}(X)|\right).$$

- Para calcular MAD, se procede de la siguiente manera:
  1. Ordenar el conjunto de datos originales y determinar la mediana, i.e. $\text{Md}(X)$;
  2. Calcular las n-desviaciones de cada dato a la $\text{Md}(X)$, i.e. $x_i - \text{Md}(X)$;
  3. Tomar el valor absoluto de las n-desviaciones, i.e. $|x_i - \text{Md}(X)|$;
  4. Re-ordenar los valores absolutos de las desviaciones (de menor a mayor);
  5. Encontrar la mediana de los valores absolutos de las desviaciones ordenadas. Esta Mediana es MAD.
- Para el ejemplo, para los jóvenes $\text{MAD} = 5,9391$ y para los adultos $\text{MAD} = 4,4593$.

- MAD se define por:

$$\text{MAD} = \text{Md}\left(|x_i - \text{Md}(X)|\right).$$

- Para calcular MAD, se procede de la siguiente manera:
  1. Ordenar el conjunto de datos originales y determinar la mediana, i.e. $\text{Md}(X)$;
  2. Calcular las n-desviaciones de cada dato a la $\text{Md}(X)$, i.e. $x_i - \text{Md}(X)$;
  3. Tomar el valor absoluto de las n-desviaciones, i.e. $|x_i - \text{Md}(X)|$;
  4. Re-ordenar los valores absolutos de las desviaciones (de menor a mayor);
  5. Encontrar la mediana de los valores absolutos de las desviaciones ordenadas. Esta Mediana es MAD.
- Para el ejemplo, para los jóvenes $\text{MAD} = 5,9391$ y para los adultos $\text{MAD} = 4,4593$.

- MAD se define por:

$$\text{MAD} = \text{Md}\left(|x_i - \text{Md}(X)|\right).$$

- Para calcular MAD, se procede de la siguiente manera:
  1. Ordenar el conjunto de datos originales y determinar la mediana, i.e. $\text{Md}(X)$;
  2. Calcular las n-desviaciones de cada dato a la $\text{Md}(X)$, i.e. $x_i - \text{Md}(X)$;
  3. Tomar el valor absoluto de las n-desviaciones, i.e. $|x_i - \text{Md}(X)|$;
  4. Re-ordenar los valores absolutos de las desviaciones (de menor a mayor);
  5. Encontrar la mediana de los valores absolutos de las desviaciones ordenadas. Esta Mediana es MAD.
- Para el ejemplo, para los jóvenes $\text{MAD} = 5,9391$ y para los adultos $\text{MAD} = 4,4593$.

Análisis Exploratorio de Datos Unidimensionales

- Otras medidas de posición importantes, llamadas **Cuartiles**, son útiles para estudiar la variabilidad de una muestra.
- Los cuartiles q_1, q_2 y q_3 son valores numéricos que dividen una muestra de observaciones en grupos, de manera tal que:
  - $\frac{1}{4}$ de los datos (25 %) son menores que q_1;
  - $\frac{1}{2}$ (la mitad) de los datos (50 %) son menores que q_2;
  - $\frac{3}{4}$ de los datos (75 %) son menores que q_3.
- Observemos que $q_2 = \text{Md}(X)$.

Análisis Exploratorio de Datos Unidimensionales

- Otras medidas de posición importantes, llamadas **Cuartiles**, son útiles para estudiar la variabilidad de una muestra.
- Los cuartiles q_1, q_2 y q_3 son valores numéricos que dividen una muestra de observaciones en grupos, de manera tal que:
  - $\frac{1}{4}$ de los datos (25 %) son menores que q_1;
  - $\frac{1}{2}$ (la mitad) de los datos (50 %) son menores que q_2;
  - $\frac{3}{4}$ de los datos (75 %) son menores que q_3.
- Observemos que $q_2 = \text{Md}(X)$.

- Otras medidas de posición importantes, llamadas **Cuartiles**, son útiles para estudiar la variabilidad de una muestra.
- Los cuartiles q_1, q_2 y q_3 son valores numéricos que dividen una muestra de observaciones en grupos, de manera tal que:
  - $\frac{1}{4}$ de los datos (25 %) son menores que q_1;
  - $\frac{1}{2}$ (la mitad) de los datos (50 %) son menores que q_2;
  - $\frac{3}{4}$ de los datos (75 %) son menores que q_3.
- Observemos que $q_2 = \text{Md}(X)$.

Análisis Exploratorio de Datos Unidimensionales

- Para determinar los cuartiles, procedemos de la siguiente manera:

1. Ordenar los datos: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.
2. Si n es impar, hacer $q = \frac{n+1}{2}$. Si n es par, hacer $q = \frac{n}{2}$. Entonces:

$$\text{Md}(X) = q_2 = \begin{cases} x_{(q)} & si \quad n \text{ es impar} \\ \\ \frac{x_{(q)} + x_{(q+1)}}{2} & si \quad n \text{ es par} \end{cases}$$

3.

$$\text{Si } q \text{ es impar, hacer } \; r = \frac{q+1}{2} \; \text{ y definir : } \begin{cases} q_1 = x_{(r)} \\ \\ q_3 = x_{(n+1-r)} \end{cases}$$

$$\text{Si } q \text{ es par, hacer } \; r = \frac{q}{2} \; \text{ y definir : } \begin{cases} q_1 = \frac{x_{(r)} + x_{(r+1)}}{2} \\ \\ q_3 = \frac{x_{(n+1-r)} + x_{(n-r)}}{2} \end{cases}$$

4. El rango intercuartílico se calcula como $RIQ = q_3 - q_1$.

Análisis Exploratorio de Datos Unidimensionales

- Para determinar los cuartiles, procedemos de la siguiente manera:

1. Ordenar los datos: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.
2. Si n es impar, hacer $q = \frac{n+1}{2}$. Si n es par, hacer $q = \frac{n}{2}$. Entonces:

$$\text{Md}(X) = q_2 = \begin{cases} x_{(q)} & si \quad n \text{ es impar} \\ \\ \frac{x_{(q)} + x_{(q+1)}}{2} & si \quad n \text{ es par} \end{cases}$$

3.

$$\text{Si } q \text{ es impar, hacer } \; r = \frac{q+1}{2} \; \text{ y definir : } \begin{cases} q_1 = x_{(r)} \\ \\ q_3 = x_{(n+1-r)} \end{cases}$$

$$\text{Si } q \text{ es par, hacer } \; r = \frac{q}{2} \; \text{ y definir : } \begin{cases} q_1 = \frac{x_{(r)} + x_{(r+1)}}{2} \\ \\ q_3 = \frac{x_{(n+1-r)} + x_{(n-r)}}{2} \end{cases}$$

4. El rango intercuartílico se calcula como $RIQ = q_3 - q_1$.

Análisis Exploratorio de Datos Unidimensionales

- Para determinar los cuartiles, procedemos de la siguiente manera:

1. Ordenar los datos: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.
2. Si n es impar, hacer $q = \frac{n+1}{2}$. Si n es par, hacer $q = \frac{n}{2}$. Entonces:

$$\text{Md}(X) = q_2 = \begin{cases} x_{(q)} & si \quad n \text{ es impar} \\ \frac{x_{(q)} + x_{(q+1)}}{2} & si \quad n \text{ es par} \end{cases}$$

3. Si q es impar, hacer $r = \frac{q+1}{2}$ y definir : $\begin{cases} q_1 = x_{(r)} \\ q_3 = x_{(n+1-r)} \end{cases}$

   Si q es par, hacer $r = \frac{q}{2}$ y definir : $\begin{cases} q_1 = \frac{x_{(r)} + x_{(r+1)}}{2} \\ q_3 = \frac{x_{(n+1-r)} + x_{(n-r)}}{2} \end{cases}$

4. El rango intercuartílico se calcula como $RIQ = q_3 - q_1$.

Análisis Exploratorio de Datos Unidimensionales

- Para determinar los cuartiles, procedemos de la siguiente manera:

1. Ordenar los datos: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.
2. Si n es impar, hacer $q = \frac{n+1}{2}$. Si n es par, hacer $q = \frac{n}{2}$. Entonces:

$$\text{Md}(X) = q_2 = \begin{cases} x_{(q)} & si \quad n \text{ es impar} \\ \\ \frac{x_{(q)} + x_{(q+1)}}{2} & si \quad n \text{ es par} \end{cases}$$

3.

$$\text{Si } q \text{ es impar, hacer } \; r = \frac{q+1}{2} \; \text{ y definir : } \begin{cases} q_1 = x_{(r)} \\ \\ q_3 = x_{(n+1-r)} \end{cases}$$

$$\text{Si } q \text{ es par, hacer } \; r = \frac{q}{2} \; \text{ y definir : } \begin{cases} q_1 = \frac{x_{(r)} + x_{(r+1)}}{2} \\ \\ q_3 = \frac{x_{(n+1-r)} + x_{(n-r)}}{2} \end{cases}$$

4. El rango intercuartílico se calcula como $RIQ = q_3 - q_1$.

Análisis Exploratorio de Datos Unidimensionales

- Para determinar los cuartiles, procedemos de la siguiente manera:

1. Ordenar los datos: $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.
2. Si n es impar, hacer $q = \frac{n+1}{2}$. Si n es par, hacer $q = \frac{n}{2}$. Entonces:

$$\text{Md}(X) = q_2 = \begin{cases} x_{(q)} & si \quad n \text{ es impar} \\ \\ \frac{x_{(q)} + x_{(q+1)}}{2} & si \quad n \text{ es par} \end{cases}$$

3.

$$\text{Si } q \text{ es impar, hacer } \; r = \frac{q+1}{2} \; \text{ y definir : } \begin{cases} q_1 = x_{(r)} \\ \\ q_3 = x_{(n+1-r)} \end{cases}$$

$$\text{Si } q \text{ es par, hacer } \; r = \frac{q}{2} \; \text{ y definir : } \begin{cases} q_1 = \frac{x_{(r)} + x_{(r+1)}}{2} \\ \\ q_3 = \frac{x_{(n+1-r)} + x_{(n-r)}}{2} \end{cases}$$

4. El rango intercuartílico se calcula como $RIQ = q_3 - q_1$.

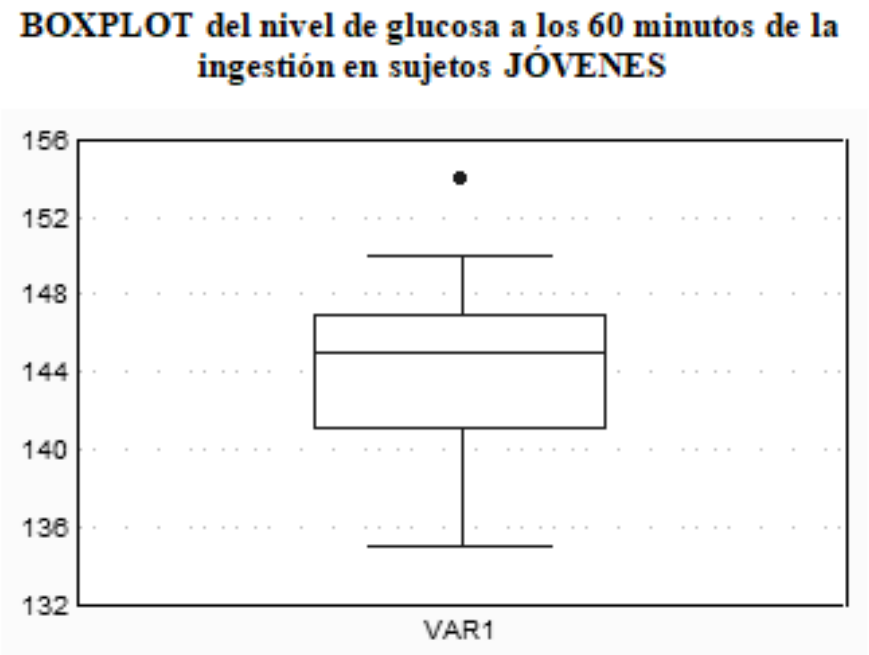
Análisis Exploratorio de Datos Unidimensionales

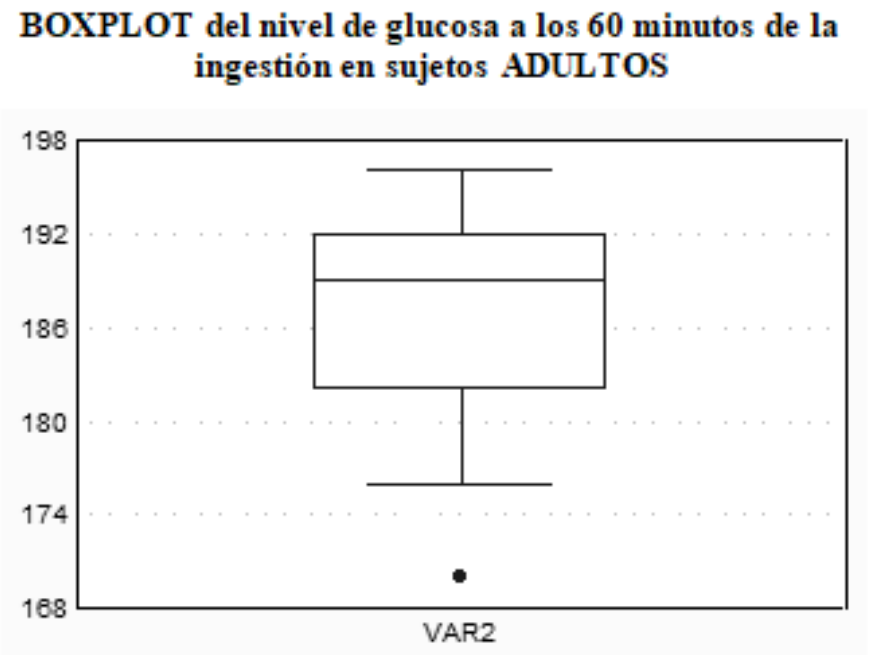
- Todas las medidas, hasta aquí definidas y calculadas para el ejemplo de la tolerancia a la glucosa, en jóvenes y adultos, se muestran en la siguiente tabla resumen:

	JOVENES	ADULTOS
Media	144,6875	187,5625
Desvío	5,5275	6,7079
Mínimo	135,0	170,0
Primer cuartil	141,0	183,5
Mediana	145,0	189,5
Tercer cuartil	148,0	192,0
Máximo	156,0	197,0
MAD	5,9391	4,4543
Amplitud intercuartílica	7,0	8,5

Análisis Exploratorio de Datos Unidimensionales

- Los 5-puntos (mínimo, q_1, q_2, q_3 y máximo) pueden ser representados en un gráfico conocido como Diseño de Caja (o *"Box Plot"* en inglés):





Análisis Exploratorio de Datos Unidimensionales

- El **Gráfico de Caja** es una forma de presentación estadística destinada, fundamentalmente, a resaltar aspectos de la distribución de las observaciones en una o más series de datos cuantitativos.
- Reemplaza, en consecuencia, al histograma y a la curva de distribución de frecuencias sobre los que tiene ventajas en cuanto a la información que brinda y a la apreciación global que surge de la lectura.
- Fue ideado por John Tukey, de la Universidad de Princeton (U.S.A.) en 1977 y los detalles que daremos a seguir corresponden a la descripción dada por este autor.
- Cabe destacar que en diferentes textos (y presentaciones del gráfico), algunos elementos se utilizan de manera diferente a las señaladas por su creador.

Análisis Exploratorio de Datos Unidimensionales

- El **Gráfico de Caja** es una forma de presentación estadística destinada, fundamentalmente, a resaltar aspectos de la distribución de las observaciones en una o más series de datos cuantitativos.
- Reemplaza, en consecuencia, al histograma y a la curva de distribución de frecuencias sobre los que tiene ventajas en cuanto a la información que brinda y a la apreciación global que surge de la lectura.
- Fue ideado por John Tukey, de la Universidad de Princeton (U.S.A.) en 1977 y los detalles que daremos a seguir corresponden a la descripción dada por este autor.
- Cabe destacar que en diferentes textos (y presentaciones del gráfico), algunos elementos se utilizan de manera diferente a las señaladas por su creador.

Análisis Exploratorio de Datos Unidimensionales

- El **Gráfico de Caja** es una forma de presentación estadística destinada, fundamentalmente, a resaltar aspectos de la distribución de las observaciones en una o más series de datos cuantitativos.
- Reemplaza, en consecuencia, al histograma y a la curva de distribución de frecuencias sobre los que tiene ventajas en cuanto a la información que brinda y a la apreciación global que surge de la lectura.
- Fue ideado por John Tukey, de la Universidad de Princeton (U.S.A.) en 1977 y los detalles que daremos a seguir corresponden a la descripción dada por este autor.
- Cabe destacar que en diferentes textos (y presentaciones del gráfico), algunos elementos se utilizan de manera diferente a las señaladas por su creador.

Análisis Exploratorio de Datos Unidimensionales

- El **Gráfico de Caja** es una forma de presentación estadística destinada, fundamentalmente, a resaltar aspectos de la distribución de las observaciones en una o más series de datos cuantitativos.
- Reemplaza, en consecuencia, al histograma y a la curva de distribución de frecuencias sobre los que tiene ventajas en cuanto a la información que brinda y a la apreciación global que surge de la lectura.
- Fue ideado por John Tukey, de la Universidad de Princeton (U.S.A.) en 1977 y los detalles que daremos a seguir corresponden a la descripción dada por este autor.
- Cabe destacar que en diferentes textos (y presentaciones del gráfico), algunos elementos se utilizan de manera diferente a las señaladas por su creador.

- **Elementos:**
- Este gráfico utiliza una sola escala: la correspondiente a la variable de los datos que se presentan. Es decir, no utiliza escala de frecuencias. Por lo tanto, no corresponde asociarlo a los que utilizan el sistema de coordenadas cartesianas.
- Los elementos que lo constituyen son:
  1. **La caja:** Es un rectángulo que abarca el recorrido o Rango InterCuartílico (RIC) de la distribución; o sea, el tramo de la escala que va desde el primer cuartil (q_1) al tercer cuartil (q_3). Esto incluye el 50 % de las observaciones centrales.
  2. **Mediana:** Se dibuja mediante una línea (algunos lo marcan con un asterisco, otros con una cruz o rectangulo chiquito sobre la línea) dentro de la caja y a la altura de la escala que corresponde al valor de esa medida. También (a pedido) puede destacarse un punto que corresponde al valor de la media

Análisis Exploratorio de Datos Unidimensionales

- **Elementos:**
- Este gráfico utiliza una sola escala: la correspondiente a la variable de los datos que se presentan. Es decir, no utiliza escala de frecuencias. Por lo tanto, no corresponde asociarlo a los que utilizan el sistema de coordenadas cartesianas.
- Los elementos que lo constituyen son:
  1. **La caja:** Es un rectángulo que abarca el recorrido o Rango InterCuartílico (RIC) de la distribución; o sea, el tramo de la escala que va desde el primer cuartil (q_1) al tercer cuartil (q_3). Esto incluye el 50 % de las observaciones centrales.
  2. **Mediana:** Se dibuja mediante una línea (algunos lo marcan con un asterisco, otros con una cruz o rectangulo chiquito sobre la línea) dentro de la caja y a la altura de la escala que corresponde al valor de esa medida. También (a pedido) puede destacarse un punto que corresponde al valor de la media

- **Elementos:**
- Este gráfico utiliza una sola escala: la correspondiente a la variable de los datos que se presentan. Es decir, no utiliza escala de frecuencias. Por lo tanto, no corresponde asociarlo a los que utilizan el sistema de coordenadas cartesianas.
- Los elementos que lo constituyen son:
  1. **La caja:** Es un rectángulo que abarca el recorrido o Rango InterCuartílico (RIC) de la distribución; o sea, el tramo de la escala que va desde el primer cuartil (q_1) al tercer cuartil (q_3). Esto incluye el 50 % de las observaciones centrales.
  2. **Mediana:** Se dibuja mediante una línea (algunos lo marcan con un asterisco, otros con una cruz o rectangulo chiquito sobre la línea) dentro de la caja y a la altura de la escala que corresponde al valor de esa medida. También (a pedido) puede destacarse un punto que corresponde al valor de la media

- **Elementos:**
- Este gráfico utiliza una sola escala: la correspondiente a la variable de los datos que se presentan. Es decir, no utiliza escala de frecuencias. Por lo tanto, no corresponde asociarlo a los que utilizan el sistema de coordenadas cartesianas.
- Los elementos que lo constituyen son:
  1. **La caja:** Es un rectángulo que abarca el recorrido o Rango InterCuartílico (RIC) de la distribución; o sea, el tramo de la escala que va desde el primer cuartil (q_1) al tercer cuartil (q_3). Esto incluye el 50 % de las observaciones centrales.
  2. **Mediana:** Se dibuja mediante una línea (algunos lo marcan con un asterisco, otros con una cruz o rectangulo chiquito sobre la línea) dentro de la caja y a la altura de la escala que corresponde al valor de esa medida. También (a pedido) puede destacarse un punto que corresponde al valor de la media

- **Elementos:**
- Este gráfico utiliza una sola escala: la correspondiente a la variable de los datos que se presentan. Es decir, no utiliza escala de frecuencias. Por lo tanto, no corresponde asociarlo a los que utilizan el sistema de coordenadas cartesianas.
- Los elementos que lo constituyen son:
  1. **La caja:** Es un rectángulo que abarca el recorrido o Rango InterCuartílico (RIC) de la distribución; o sea, el tramo de la escala que va desde el primer cuartil (q_1) al tercer cuartil (q_3). Esto incluye el 50 % de las observaciones centrales.
  2. **Mediana:** Se dibuja mediante una línea (algunos lo marcan con un asterisco, otros con una cruz o rectangulo chiquito sobre la línea) dentro de la caja y a la altura de la escala que corresponde al valor de esa medida. También (a pedido) puede destacarse un punto que corresponde al valor de la media

Análisis Exploratorio de Datos Unidimensionales

3. **Bigotes:** Son líneas que salen a los costados de la caja y que sirven como referencia para ubicar las observaciones que están por fuera del 50 % central de la distribución. (Para determinar su longitud: ver explicación sobre construcción del diseño a seguir).
4. **Puntos Periféricos:** Es el señalamiento de las observaciones que se encuentran entre la finalización del bigote y una medida más del tamaño del bigote. Se marcan, usualmente con un asterístico .
(Algunos paquetes informáticos utilizan una “O”).
5. **Puntos extremos o periféricos lejanos o *“outliers”*:** Es el señalamiento de las observaciones que se encuentran fuera de, por lo menos, dos rangos intercuartílicos y medios. Se marcan con un punto grande •.
(Algunos paquetes informáticos utilizan una “E” y otros una “X”).

Análisis Exploratorio de Datos Unidimensionales

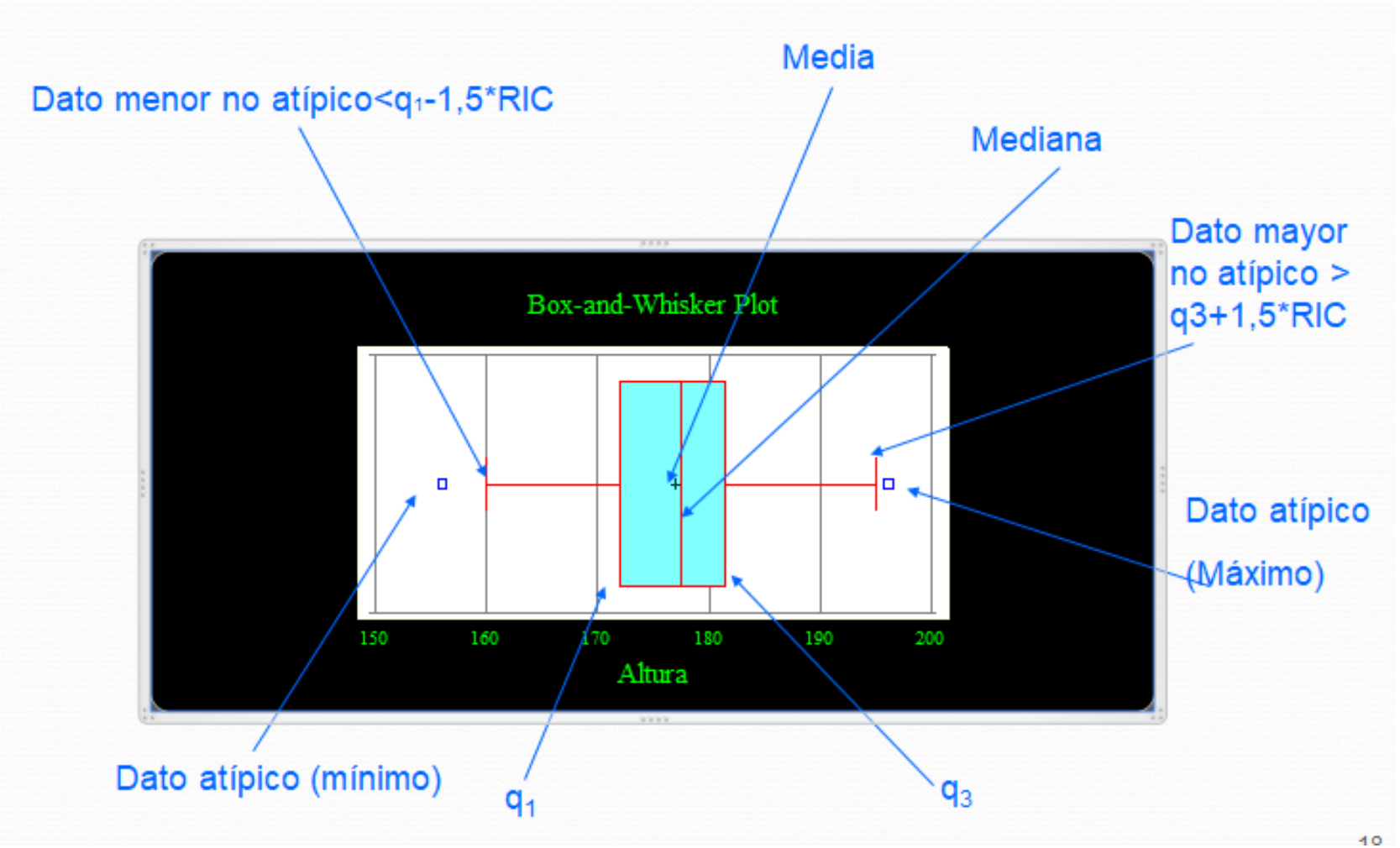
3. **Bigotes:** Son líneas que salen a los costados de la caja y que sirven como referencia para ubicar las observaciones que están por fuera del 50 % central de la distribución. (Para determinar su longitud: ver explicación sobre construcción del diseño a seguir).
4. **Puntos Periféricos:** Es el señalamiento de las observaciones que se encuentran entre la finalización del bigote y una medida más del tamaño del bigote. Se marcan, usualmente con un asterístico . (Algunos paquetes informáticos utilizan una “O”).
5. **Puntos extremos o periféricos lejanos o *“outliers”*:** Es el señalamiento de las observaciones que se encuentran fuera de, por lo menos, dos rangos intercuartílicos y medios. Se marcan con un punto grande •.
(Algunos paquetes informáticos utilizan una “E” y otros una “X”).

Análisis Exploratorio de Datos Unidimensionales

3. **Bigotes:** Son líneas que salen a los costados de la caja y que sirven como referencia para ubicar las observaciones que están por fuera del 50 % central de la distribución. (Para determinar su longitud: ver explicación sobre construcción del diseño a seguir).
4. **Puntos Periféricos:** Es el señalamiento de las observaciones que se encuentran entre la finalización del bigote y una medida más del tamaño del bigote. Se marcan, usualmente con un asterístico .
(Algunos paquetes informáticos utilizan una “O”).
5. **Puntos extremos o periféricos lejanos o *“outliers”*:** Es el señalamiento de las observaciones que se encuentran fuera de, por lo menos, dos rangos intercuartílicos y medios. Se marcan con un punto grande •.
(Algunos paquetes informáticos utilizan una “E” y otros una “X”).

Análisis Exploratorio de Datos Unidimensionales

- Los elementos citados se destacan en el siguiente **Diseño de Caja**



- **Construcción del Box Plot:**
  1. La creación de la caja no ofrece dificultades ya que se extiende entre q_1 y q_3; quedando, así, determinados los lados izquierdo y derecho por los puntos de la escala a la que corresponden esas medidas.
     Los lados superior e inferior de la caja no están determinados más que por la conveniencia estética de la presentación; es decir, lo define quien lo dibuja (o el programa informático).
  2. Tampoco representa dificultad trazar la mediana: se lo hace a la altura de la escala donde se encuentre el valor correspondiente a ese estadístico.
  3. Los bigotes merecen alguna descripción más detallada. En principio, cada uno debe tener un largo "máximo" equivalente a $1,5$ veces el largo de la caja. Si los valores máximo y/o mínimo se encuentra dentro de estos rangos de valores (es decir, a lo largo de los bigotes o de alguno de ellos) el bigote se corta en ese valor (sea el máximo o sea el mínimo).
  4. Valores observados más grandes que los puntos donde se cortan los bigotes son considerados en la categoría periféricos (extremos).

- **Construcción del Box Plot:**
  1. La creación de la caja no ofrece dificultades ya que se extiende entre q_1 y q_3; quedando, así, determinados los lados izquierdo y derecho por los puntos de la escala a la que corresponden esas medidas.
     Los lados superior e inferior de la caja no están determinados más que por la conveniencia estética de la presentación; es decir, lo define quien lo dibuja (o el programa informático).
  2. Tampoco representa dificultad trazar la mediana: se lo hace a la altura de la escala donde se encuentre el valor correspondiente a ese estadístico.
  3. Los bigotes merecen alguna descripción más detallada. En principio, cada uno debe tener un largo "máximo" equivalente a $1,5$ veces el largo de la caja. Si los valores máximo y/o mínimo se encuentra dentro de estos rangos de valores (es decir, a lo largo de los bigotes o de alguno de ellos) el bigote se corta en ese valor (sea el máximo o sea el mínimo).
  4. Valores observados más grandes que los puntos donde se cortan los bigotes son considerados en la categoría periféricos (extremos).

- **Construcción del Box Plot:**
  1. La creación de la caja no ofrece dificultades ya que se extiende entre q_1 y q_3; quedando, así, determinados los lados izquierdo y derecho por los puntos de la escala a la que corresponden esas medidas.
     Los lados superior e inferior de la caja no están determinados más que por la conveniencia estética de la presentación; es decir, lo define quien lo dibuja (o el programa informático).
  2. Tampoco representa dificultad trazar la mediana: se lo hace a la altura de la escala donde se encuentre el valor correspondiente a ese estadístico.
  3. Los bigotes merecen alguna descripción más detallada. En principio, cada uno debe tener un largo "máximo" equivalente a $1,5$ veces el largo de la caja. Si los valores máximo y/o mínimo se encuentra dentro de estos rangos de valores (es decir, a lo largo de los bigotes o de alguno de ellos) el bigote se corta en ese valor (sea el máximo o sea el mínimo).
  4. Valores observados más grandes que los puntos donde se cortan los bigotes son considerados en la categoría periféricos (extremos).

- **Construcción del Box Plot:**
  1. La creación de la caja no ofrece dificultades ya que se extiende entre q_1 y q_3; quedando, así, determinados los lados izquierdo y derecho por los puntos de la escala a la que corresponden esas medidas.
     Los lados superior e inferior de la caja no están determinados más que por la conveniencia estética de la presentación; es decir, lo define quien lo dibuja (o el programa informático).
  2. Tampoco representa dificultad trazar la mediana: se lo hace a la altura de la escala donde se encuentre el valor correspondiente a ese estadístico.
  3. Los bigotes merecen alguna descripción más detallada. En principio, cada uno debe tener un largo "máximo" equivalente a $1,5$ veces el largo de la caja. Si los valores máximo y/o mínimo se encuentra dentro de estos rangos de valores (es decir, a lo largo de los bigotes o de alguno de ellos) el bigote se corta en ese valor (sea el máximo o sea el mínimo).
  4. Valores observados más grandes que los puntos donde se cortan los bigotes son considerados en la categoría periféricos (extremos).

- **Construcción del Box Plot:**
  1. La creación de la caja no ofrece dificultades ya que se extiende entre q_1 y q_3; quedando, así, determinados los lados izquierdo y derecho por los puntos de la escala a la que corresponden esas medidas.
     Los lados superior e inferior de la caja no están determinados más que por la conveniencia estética de la presentación; es decir, lo define quien lo dibuja (o el programa informático).
  2. Tampoco representa dificultad trazar la mediana: se lo hace a la altura de la escala donde se encuentre el valor correspondiente a ese estadístico.
  3. Los bigotes merecen alguna descripción más detallada. En principio, cada uno debe tener un largo "máximo" equivalente a $1,5$ veces el largo de la caja. Si los valores máximo y/o mínimo se encuentra dentro de estos rangos de valores (es decir, a lo largo de los bigotes o de alguno de ellos) el bigote se corta en ese valor (sea el máximo o sea el mínimo).
  4. Valores observados más grandes que los puntos donde se cortan los bigotes son considerados en la categoría periféricos (extremos).

Análisis Exploratorio de Datos Unidimensionales

- En el caso del Diseño de Caja de la *"tolerancia a la glucosa"* para los jóvenes por ejemplo, si se hubiera usado la propuesta original de Tukey, el RIC sería igual a 10,5 $((q_3 - q_1) * 1,5)$, equivalente a la extensión de los bigotes, llegando éstos hasta 130,5 por izquierda y 158,5 por derecha; pero se cortarian en en el mínimo=135 y en el máximo=156. Sin embargo, con el paquete usado (*Statistica*) se definió un largo más pequeño para cada bigote.
- Aunque la descripción de la construcción se ha realizado de manera horizontal, generalmente el gráfico se presenta verticalmente: la escala de la variable trazada sobre una línea vertical y a la derecha el resto de los elementos correspondientes a una (o más) distribuciones. En su presentación vertical (la más habitual) los valores de la escala se incrementan de abajo hacia arriba.
- Así, el gráfico es útil para representar más de una distribución de frecuencias, siempre que las series utilicen la misma escala y en un tramo de ella que permita la comparación en un mismo gráfico.

Análisis Exploratorio de Datos Unidimensionales

- En el caso del Diseño de Caja de la *"tolerancia a la glucosa"* para los jóvenes por ejemplo, si se hubiera usado la propuesta original de Tukey, el RIC sería igual a 10,5 $((q_3 - q_1) * 1,5)$, equivalente a la extensión de los bigotes, llegando éstos hasta 130,5 por izquierda y 158,5 por derecha; pero se cortarian en en el mínimo=135 y en el máximo=156. Sin embargo, con el paquete usado (*Statistica*) se definió un largo más pequeño para cada bigote.
- Aunque la descripción de la construcción se ha realizado de manera horizontal, generalmente el gráfico se presenta verticalmente: la escala de la variable trazada sobre una línea vertical y a la derecha el resto de los elementos correspondientes a una (o más) distribuciones. En su presentación vertical (la más habitual) los valores de la escala se incrementan de abajo hacia arriba.
- Así, el gráfico es útil para representar más de una distribución de frecuencias, siempre que las series utilicen la misma escala y en un tramo de ella que permita la comparación en un mismo gráfico.

Análisis Exploratorio de Datos Unidimensionales

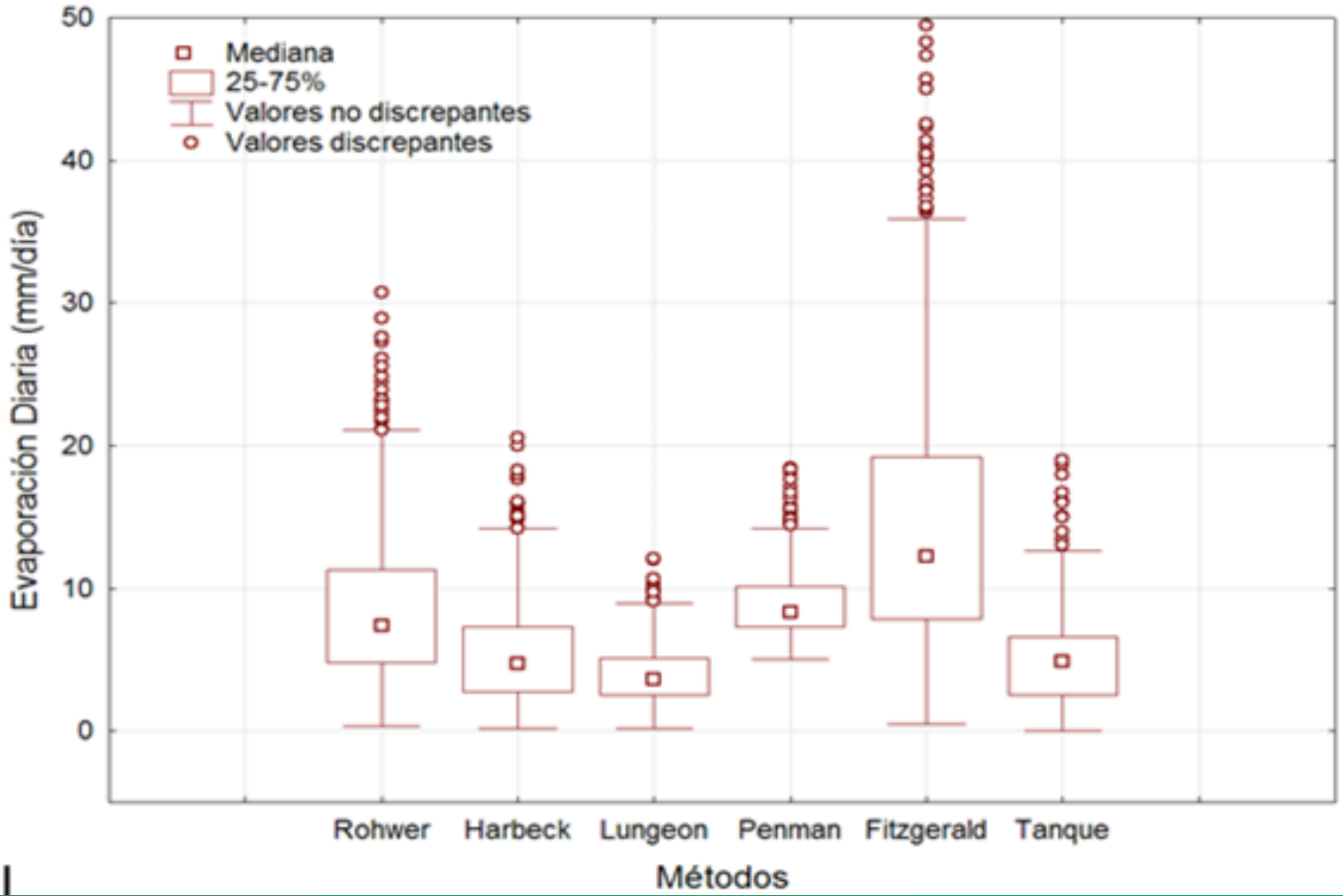
- En el caso del Diseño de Caja de la *"tolerancia a la glucosa"* para los jóvenes por ejemplo, si se hubiera usado la propuesta original de Tukey, el RIC sería igual a 10,5 $((q_3 - q_1) * 1,5)$, equivalente a la extensión de los bigotes, llegando éstos hasta 130,5 por izquierda y 158,5 por derecha; pero se cortarian en en el mínimo=135 y en el máximo=156. Sin embargo, con el paquete usado (*Statistica*) se definió un largo más pequeño para cada bigote.
- Aunque la descripción de la construcción se ha realizado de manera horizontal, generalmente el gráfico se presenta verticalmente: la escala de la variable trazada sobre una línea vertical y a la derecha el resto de los elementos correspondientes a una (o más) distribuciones. En su presentación vertical (la más habitual) los valores de la escala se incrementan de abajo hacia arriba.
- Así, el gráfico es útil para representar más de una distribución de frecuencias, siempre que las series utilicen la misma escala y en un tramo de ella que permita la comparación en un mismo gráfico.

Análisis Exploratorio de Datos Unidimensionales

- **Ejemplo:** La Tesis de Maestría en Ciencias Hídricas-UNLPam, de la Dra. María Laura LÓPEZ (marzo de 2012), dirigida por la Dra. María Cristina MARTÍN, titulada *"Análisis de los Métodos de Estimación de la Evaporación y Evapotranspiración a las condiciones locales de la ciudad de Córdoba"*, tiene como uno de sus objetivos comparar distintos procedimientos que aparecen en la bibliografía, que son y han sido ampliamente usados, para calcular la evaporación y la evapotranspiración potencial, tanto a escala diaria como mensual, partiendo de datos meteorológicos y experimentales medidos en la estación del Servicio Meteorológico Nacional y en el predio de Ciudad Universitaria de la ciudad de Córdoba a fin de determinar el(los) mejor(es) procedimiento(s), y contrastar los resultados de evaporación con las mediciones en *"tanque de evaporación clase A"* y los resultados de evapotranspiración potencial con una estimación a partir de dichas mediciones.

Análisis Exploratorio de Datos Unidimensionales

- Cuando la investigadora compara los métodos que le permiten estimar la *"Evaporación Diaria"*, resume sus resultados en el siguiente gráfico de Diseños de Caja:



Bibliografía

ALPERÍN, M. (2013): ***"Introducción al Análisis Estadístico de Datos Geológicos"***. Editorial Universidad Nacional de La Plata, 1° edición. 281 páginas.

ARRIAGA GOMEZ, A. J., FERNANDEZ PALACÍN, F., LÓPEZ SÁNCHEZ M.A., MUÑOZ MARQUEZ, M., PÉREZ PLAZA, S. y SÁNCHEZ NAVAS, A.(2008): ***"Estadística Básica con R y R-Commander"***. Recuperado de: http://knuth.uca.es/repos/ebrcmdr/pdf/actual/ebrcmdr.pdf.

DANIEL, W. W. (1999). ***"Bioestadística. Base para el Análisis de las Ciencias de la Salud"***. UTEHA: Noriega Editores. México, 878 p.

DEVORE, J.L. (2001): ***"Probabilidad y Estadística para ingeniería y ciencias"***. International Thomson Editores, S.A., México. Quinta edición. 762 páginas.

ELSTON, R.C. y JOHNSON, W.D. (1990): ***"Principios de Bioestadística"***. Ed. Manual Moderno. 298 páginas

GARCÍA, R.M. (2004). ***"Inferencia Estadística y Diseño Experimental"***. Editorial Universitaria de Buenos Aires. EUDEBA-, Buenos Aires, 734 p.

GOMEZ VILLEGAS, M.A. (2005). ***"Inferencia Estadística"***. Ediciones Díaz de Santos, 518 p.

JOHNSON, R. (1990). ***"Estadística Elemental"***. Grupo Editorial Iberoamericano. México, D.F. 592 p.

Bibliografía

MENDENHALL, W.; WACKERLY, D.D. y SCHEAFFER, R.L. (1994). ***"Estadística Matemática con Aplicaciones"***. 2da. Edición. Grupo Editorial. Iberoamericano, México, D.F. 751 p.

MOORE, D. (2004): ***"Estadística Aplicada Básica"***. Editorial ANTONI BOCH. 880 páginas.

MOSCHETTI, E.E., FERRERO, S., PALACIO, G. y RUIZ, M. (2000): ***"Introducción a la Estadística para las Ciencias de la Vida"***. Editorial de la Fundación Universidad Nacional de Río Cuarto. Córdoba, 171 p.

SONVICO, V. (1984): ***"Modelos Estadisticos y Experimentación Biológica"***. Cuaderno N° 2 - SAE, 1-14.

ZAR, J.H. (1996): ***"Bioestatistical Analysis"***. 3th. Edition. Prentice Hall. New Jersey. 663 p. + 212 App.

R Development Core Team (2010). ***"R: A language and environment for statistical computing"***. (Versión 2.12.0) [Software] Disponible en http://www.R-project.org/ .