Documento Docente

Estadística Descriptiva: Análisis de Datos en Sección Transversal

Claudio Poloni Izeta
2012

El autor es Licenciado en Ciencias en la Administración de Empresas e Ingeniero Comercial de la Pontificia Universidad Católica de Valparaíso, profesor jornada completa de la misma Universidad.

Documento que recopila y complementa textos elaborados anteriormente por docentes de la Escuela de Ingeniería Comercial con el fin de ser un aporte y guiar al estudiante en el desarrollo de la asignatura "Análisis Cuantitativo 1". Se agradece en especial a los profesores Rodrigo Vergara, Renzo Devoto y Angélica Casaletti, por todo el trabajo y aporte realizado para la elaboración de las versiones anteriores de este documento.

Serie de Documentos Docentes, Escuela de Ingeniería Comercial, Pontificia Universidad Católica de Valparaíso, Documento Docente Nº, Claudio Poloni Izeta.

Contenido

DOCUMENTO DOCENTE
Estadística Descriptiva: Análisis de Datos en Sección Transversal

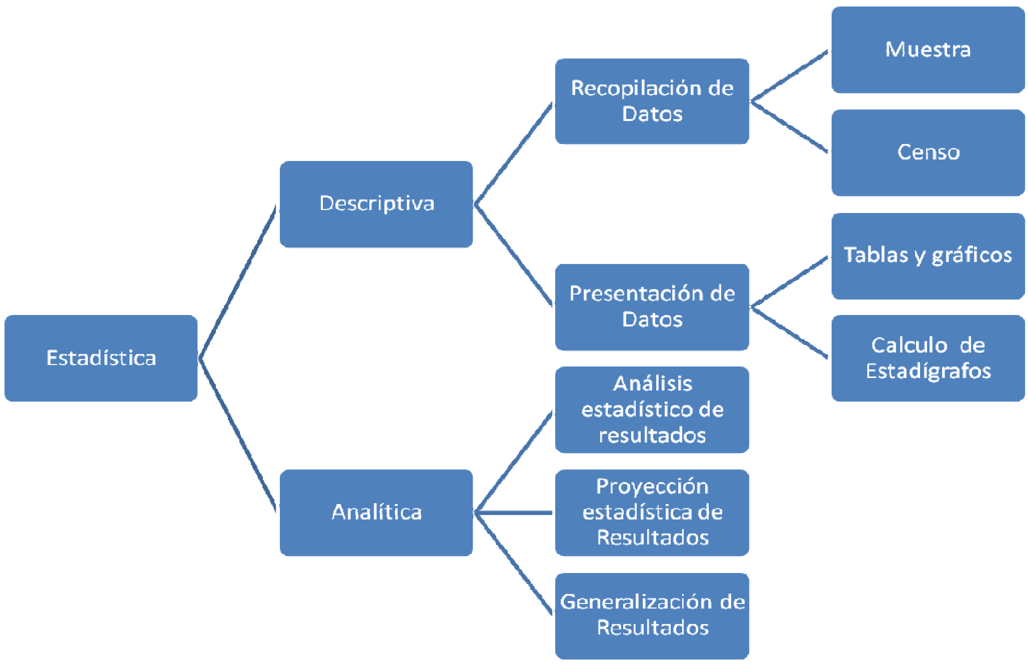
INTRODUCCION AL ANALISIS CUANTITATIVO

1. *Introducción*

La estadística es una disciplina que reúne un conjunto de procedimientos y métodos útiles para conocer el comportamiento de algún objeto de estudio a través de un conjunto de observaciones adquiridas a partir de este. Lo anterior lo realiza a través de la:

a) Obtención y descripción de un conjunto de datos, en forma cuantitativa.
b) Interpretación estadística de esa información.
c) Predicción de fenómenos en base a esa información.
d) Demostración de fenómenos físicos y sociales.

De manera formal se puede definir estadística como la tecnología (forma de hacer las cosas) capaz de entregarle significado a los datos, entregando un resumen conveniente de ellos, para generar información y toma de decisiones.
Es posible clasificar la estadística en **Descriptiva** o **Analítica**, dependiendo del objetivo que se persiga. La estadística descriptiva busca describir un conjunto de datos en forma cuantitativa, mientras que la estadística analítica busca demostrar y predecir (inferencia) fenómenos físicos y sociales.



Así, cada vez que se realiza una investigación estadística, se debe seguir el siguiente procedimiento marcado por tres procesos secuenciales sobre las observaciones que se presentan a continuación:

a) Recopilación de datos:
- definición de objetivos del estudio. ¿Qué es lo que se quiere medir?
- definición de variables a medir. ¿Qué aristas del objeto se van a medir?
- definición de la población a medir. ¿Qué objetos concretos se deben medir?
- definición de forma de hacer la medición. ¿Cómo se llevará a cabo el proceso de medición?

b) Organización y presentación de datos (Primer estado de la Información)
- construcción de tablas. Resumen amplio del conjunto de datos.
- construcción de gráficas. Visión general del comportamiento de los datos.
- calculo de estadígrafos. Resumen representativo del conjunto de datos.

c) Análisis de la información:
- descripción de los sucesos. Interpretar el comportamiento del conjunto.
- búsqueda de relaciones significativas. Variables que pueden afectar el comportamiento de otras.
- explicación de fenómenos. Causas o consecuencias del comportamiento descrito.
- proyección de los resultados. Construcción de la(s) situaciones futuras que afectará el comportamiento del conjunto de datos.
- obtención de una medida de confiabilidad de este análisis. Certificación probabilística del comportamiento de los análisis realizados y supuestos a cumplir.

La estadística es una disciplina que puede facilitar el complejo proceso de toma de decisiones, el cual es esencial en una profesión dedicada a la planificación, análisis económico y comportamiento del consumidor, en que es frecuente tomar decisiones bajo condiciones de riesgo o incertidumbre, en situaciones relacionadas tanto con la administración de personas, de clientes, proveedores y como del recurso financiero.

1.1 <u>Conjunto de datos</u>

En cualquier estudio de fenómenos sociales o físicos es necesario observar y recoger algunas características de los objetos en estudio. La intuición nos guiará a entender que con el fin de obtener información, completa, precisa y acuciosa del objeto, se debe recoge características de la totalidad del conjunto de datos. Si este es el caso, hablamos de un "estudio poblacional" donde **Población**, es el conjunto de los individuos de interés y **Censo** corresponde la forma de obtener las observaciones. (En Chile cada 10 años se realiza un levantamiento de datos de todos los habitantes del país, lo que se denomina Censo).

Sin embargo, en la mayoría de los casos es impracticable (cantidad de recursos) o imposible (espacio temporal) observar a todos los elementos de la población. Para poder observar su comportamiento, se selecciona un subconjunto de la población, esperando que este sea un buen representante de la totalidad[1]. Se habla entonces, de un "estudio muestral", donde **Muestra** es una colección de mediciones seleccionadas de la población de interés o en otros términos un subconjunto de datos representativo de la población.

Existe una exhaustiva metodología que pretende diseñar y llevar a cabo censos y muestreos, con el fin que estos reflejen fielmente el comportamiento de la población del objeto de estudio, cometiendo la menor cantidad de errores, metodología que escapa a los objetivos de este curso. Por el momento, se supondrá que la etapa de recolección de datos ya ha sido realizada de modo adecuado, y que el objetivo es organizar y presentar esos datos en una forma adecuada.

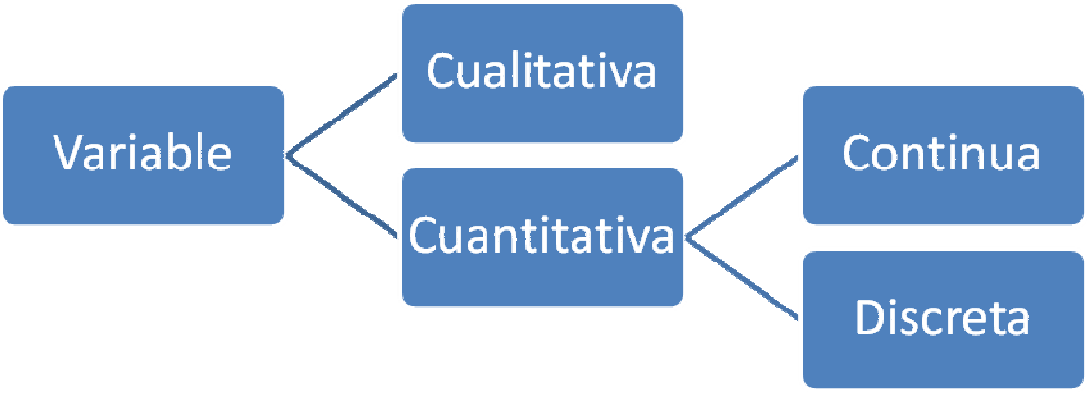
2. ORGANIZACIÓN Y PRESENTACION DE DATOS

2.1 Antecedentes Generales

La estadística es aplicable a toda clase de observaciones o cada uno de los datos, sean estas de carácter cuantitativo o cualitativo, aunque el principal interés, en este curso en particular, es considerar ciertos procedimientos adecuados para el estudio de datos cuantitativos.

Una cantidad medible, que puede variar entre un individuo a otro, se denomina **Variable**. Recibe el nombre de **variable continua** la que puede tomar cualquier valor, mientras que se llama **variable discreta** (o discontinua) aquella que sólo puede tomar valores enteros (por ejemplo: número de hijos por familia, cantidad de contenedores por bodega, número de habitaciones en una casa, artículos vendidos mensualmente, etc.). Esquematizando:

[1] Considere por ejemplo que al momento de realizarle un hemograma (estudio de la sangre) sólo se seleccionan una proporción relativamente pequeña de su sangre para entender el comportamiento de toda ella cometiendo relativamente pocos errores de estimación.



Existen diversas formas de trasladar variables del tipo cualitativa, como gustos, opiniones, percepciones, entre otras, a variables del tipo cuantitativas que permiten su análisis y la utilización de los métodos estadísticos a ser utilizados durante este curso. Una de ellas es transformación a escalas jerárquicas las cuales pretenden ordenar y distanciar las variables cualitativas unas de otras (escalas tipo Likert).

Por otro lado, en estadística puede interesar una sola característica o atributo de un objeto (caso unidimensional), por ejemplo las rentas obtenidas por los ejecutivos de una empresa. También, es posible considerar, en un mismo estudio, **dos o más** atributos en forma simultánea (caso multidimensional) como por ejemplo las rentas de los ejecutivos de una empresa, relacionadas con la antigüedad en la misma, y con la preparación académica de cada uno de ellos[2].

El interés en este curso, es referirse a la organización y presentación de **datos unidimensionales** de una variable cuantitativa, así como la descripción resumida pero sumamente representativa del conjunto de observaciones, que llamaremos **Estadígrafos** (como la media aritmética, la mediana y la desviación estándar, entre otros).

2.2 __Distribuciones de frecuencia unidimensionales__

Una forma frecuentemente usada con el fin de conocer el comportamiento de alguna variable, es conociendo su historia o lo sucedido con ella en algún instante del tiempo. Conocer el comportamiento de un conjunto de datos nos permite a través de alguna metodología poder proyectar o inferir algo de su comportamiento futuro. Una forma de presentar ese comportamiento pasado es a través de su **distribución de frecuencia** (la manera en que se repiten las observaciones de una variable de interés en un conjunto de datos).

2.2.1 __Tablas de frecuencia__

Con el fin de lograr conocer la distribución de frecuencia, la cual permitirá conocer el comportamiento de una variable de un conjunto de datos, es conveniente dar un cierto grado de agregación a los datos, con el fin de visualizar de mejor manera esa distribución

[2] No importando cual es la cantidad de variables que describan el atributo al cual se refiere.

de frecuencia y apreciar de manera mas general, esto permite tener una visión de conjunto y total del comportamiento de cualquier conjunto de datos. Para lograr lo anterior, es recomendable trabajar en base a tablas[3].

Ejemplo

Suponga que las colocaciones, en miles de unidades monetarias (M$), de las 12 sucursales de un banco en un mismo periodo de tiempo, son las dadas en la tabla 1, en orden aleatorio.

Tabla nº 1 Colocaciones (M$) mes Enero (datos no agrupados)
205.300
110.433
95.250
76.982
262.070
123.980
240.710
53.005
207.025
100.001
248.005
76.843

Tabla nº 2 Colocaciones (M$) mes Enero (tabla de clasificación con datos no agrupados)
53.005
76.843
76.982
95.250
100.001
110.433
123.980
170.340
207.025
240.710
248.005
262.070

Evidentemente, la forma de presentar las remuneraciones de la tabla N^{ro} 1 es pobre, y no permite formar una idea de la distribución de las colocaciones. No conocer de forma sencilla y directa el máximo y el mínimo de una distribución dificulta el entendimiento de cual es el comportamiento de un conjunto de datos.

Una primera mejora que se sugiere es la de ordenar de forma ascendente o descendente los datos de cualquier tabla, consiguiéndose una primera **Tabla de Clasificación**, tal como se muestra inmediatamente en la tabla N^{ro} 2.

No obstante, la tabla N^{ro} 2 tampoco resulta cómoda para presentar los datos y mucho menos para el análisis de los mismos, ya que continúa mostrando las observaciones de forma extensa y no general, lo cual es primordial con el fin de conocer el comportamiento de un conjunto de datos.

Para solucionar la poca generalidad que presenta esta tabla, lo cual permite tener una percepción adecuada del comportamiento del conjunto de datos, es intentar agrupar estos

[3] La manera en la cual se trabaja con los datos, es siempre a partir de resúmenes cómodos como son las tablas, las cuales permiten presentar información, realizar gráficos y algún nivel de análisis general.

en categorías, que denominaran **clases**. Esto permitirá la construcción de un resumen de mayor utilidad con los fines de identificar u observar el comportamiento de un conjunto de datos que se llama **Tabla de clasificación**.

A partir del ejemplo podemos notar que los datos son continuos lo que genera una mayor dificultad en la posibilidad de poder agruparlos en algún subconjunto ya que en general los datos de variables continuas son distintos entre sí. En cambio en datos continuos la probabilidad de encontrar datos iguales es relativamente grande. Con este tipo de datos, es posible construir tablas de frecuencias resumen, las cuales agrupan datos iguales en valor reconociendo la frecuencia de ocurrencia.

*2.2.2 **Tablas de Clasificación***

Una forma sencilla de proceder sería trabajar con dos clases, dividiendo dicotómicamente los datos clasificándolos (los mayores a un valor en un subconjunto y los menores a dicho valor en otro subconjunto).

Si bien esta opción permite reducir (objetivo de la segmentación) la cantidad y variedad de los datos en análisis, la pérdida de información inherente a la agrupación aludida es sumamente grande ya que no logra este resumen describir adecuadamente el comportamiento del conjunto de datos.

Clases			(veces que ocurre el evento)
\$53.005	-	\$209.065	8
\$209.066	-	\$262.070	4

Esto sucede principalmente ya que en esta agrupación efectuada no se sabe si los datos se reparten de forma uniforme o bien todos los datos están junto a cualquiera de los extremos de cada uno de los subconjuntos o clases construidos.

Con el fin de solucionar esta pérdida de información, una de las soluciones comúnmente aceptadas, se puede encontrar en una agrupación de los datos en varias clases, determinadas, cada una de ellas, por un límite inferior (LI_i) y por un límite superior (LS_i), los cuales agrupan una proporción de datos que se asemejan entre sí o no difieren mucho del comportamiento central del espacio en termino de valores que existe en la clase.

Las condiciones necesaria para formular esta segmentación es que no deben existir clases sin datos y que las clases sean un continuo en termino de valores y claramente identificables con respecto a la pertenencia de los datos.

Se conoce como **Amplitud de clase** (C_i), al intervalo numérico comprendido entre esos dos límites descritos anteriormente, es decir:

$$C_i = LS_i - LI_i$$

Su principal utilidad es conocer cuál es la magnitud en término de valores en la variable que identifica este subconjunto, lo cual nada tiene que ver con la cantidad de eventos contenido en cada una de las clases descritas o frecuencia de ocurrencia.

A su vez, cada clase tiene un valor central que llamaremos **Marca de clase** (M_i), que puede calcularse como:

$$M_i = \frac{(LI_i + LS_i)}{2}$$

Este dato será, a falta de otros estimadores el comportamiento habitual de los datos que existen dentro de una clase, la cual asume que los datos se distribuyen dentro de él de manera **uniforme,** lo cual no es necesariamente cierto y se da en una proporción menor de los conjuntos de datos. Recuerde que posterior a la agrupación de los datos en cada una de las clases usted nunca podrá conocer cuál es el real valor que tomó la variable para cada una de las observaciones contenidas en las clases.
Una vez identificadas todas las clases (vital para conocer su comportamiento de manera general y lo primero que se debe construir con el fin de realizar análisis de un conjunto de datos) nos interesará entonces, determinar cuántos o que proporción de los datos observados quedan contenidos en cada clase definida previamente.

Llamaremos **Frecuencia absoluta** (f_i) de la clase "i" a la cantidad de datos contenidos en dicha clase.
Llamaremos **Frecuencia relativa** (fi/n) a la proporción del total del conjunto de datos que quedan contenidos en la clase "i".

Definamos la siguiente tabla de clasificación para nuestro ejemplo, en base a cuatro clases de amplitud 80.000 u.m cada una:

Tabla N^{ro} 3	Tabla de Clasificación Colocaciones mes de Enero (M$) (Datos Agrupados)		
CLASE	Marca (M_i)	Frecuencia Absoluta (f_i)	Frecuencia Relativa (f_i/n)
50.000 – 100.000	75.000	4	0.333
100.001 – 150.000	125.500	3	0.25
150.001 – 240.000	195.000	2	0.166
240.001 – 263.000	251.500	3	0.25

Con este resumen de la información presentada a partir de la tabla de clasificación podemos apreciar cuatro subconjunto de datos, donde estos se asemejan entre si y nos permiten ver de forma resumida pero con algún detalle el comportamiento que tienen todas los subconjunto de datos que pertenecen a él.
Como se puede apreciar la elección del número de clases y su amplitud fue realizada arbitrariamente, sin ningún tipo de criterio. Parte ciencia y en parte arte es poder construir

una tabla de clasificación que de forma acotada, sea capaz de mostrar el conjunto de datos de forma representativa, lo cual es fundamental para interpretar de mejor manera el comportamiento de algún objeto.

A continuación se detallarán algunos criterios básicos que pueden ser utilizados con el fin de que la tabla de clasificación muestre de forma adecuada el comportamiento de un conjunto de datos que pretenden mostrar el comportamiento de algún objeto. Algunos criterios o sugerencias son:

a) Un número de clases superior a 20 dificulta la lectura y el análisis de la información. Sin embargo dependiendo de la situación y la información que se desea mostrar no quedará otra alternativa (no fuerce la tabla de clasificación a tener menos de 20 clases).
b) Un número de clases inferior a 5 implica perder mucha información en aras de la comodidad de presentación. Sin embargo, existen muchos tipos de conjunto donde la información queda adecuadamente organizada en menos de 5 clases (No subdivida clases de forma artificial).
c) Las clases deben ser mutuamente excluyentes entre sí. Debe haber continuidad en las clases adyacentes. Si desea dar énfasis, en un conjunto de datos discretos y en algunos continuos, que algún límite inferior o superior está incluido en alguna clase puede agregar un signo **(+)** adyacente a el límite en referencia. De igual modo puede utilizar un signo **(-)** si desea indicar que el limite no se encuentra incluido en dicha clase.
d) Conviene (aunque no es indispensable) trabajar con clases de igual amplitud, para permitir un análisis visual más fácil. Posteriormente es necesario acomodar clases dependiendo de la lejanía que presentan los datos extremos del resto del conjunto.
e) La **Marca de clase** es un valor representativo de cada clase, por lo que se la debe intentar asociar a aquellos datos que tengan una mayor frecuencia de ocurrencia (sobre todo cuando se trabaja con datos discretos) ya que se asume una distribución uniforme.

Podemos entonces, mejorar la presentación de los datos del ejemplo anterior, en base a estos criterios. De esa forma, la amplitud de clase determinada por el rango de los datos y por la cantidad de clases que se desea tener:

- Recorrido de los datos del ejemplo: $\begin{array}{r}\$262.070\\ \underline{-\$53.005}\\ \$209.065\end{array}$
- Número de clase sugerido : 5
- Con dichos valores se obtiene un amplitud de clase tentativa (C_i) :
  $C_i = \$209.065/5 = \$41.813$
- luego, por comodidad, la amplitud de clase puede ser : $C_i = \$42.000$

2.2.2.1 Información adicional en una tabla de clasificación

Adicionalmente podemos complementar la tabla de clasificación utilizando los datos reflejados en ella, agregando la siguiente información:

- **Frecuencia Acumulada Menos de** (F^-): Cantidad de observaciones contenidas bajo el límite superior de la clase "i".
- **Frecuencia Acumulada Más de** (F^+): Cantidad de observaciones contenidas sobre el límite inferior de la clase "i".
- **Frecuencia Acumulada Relativa Menos de** (F^-/n): proporción de observaciones contenidas bajo el límite superior de la clase "i".
- **Frecuencia Acumulada Relativa Más de** (F^+/n): proporción de observaciones contenidas sobre el límite inferior de la clase "i".

La siguiente tabla[4] corresponde a aquella ajustada a los criterios enumerados anteriormente:

Tabla N^{ro} 4	Tabla de clasificación de Colocaciones del mes de enero (Datos Agrupados)						
Clases (M$)	**M_i**	**F_i**	**F_i/n**	**F_i^-**	**F_i^-/n**	**F_i^+**	**F_i^+/n**
53 a 95 (-)	74	3	0,25	3	0,25	12	1
95 - 137 (-)	116	4	0,33	7	0,583	9	0,75
137-179 (-)	158	1	0,083	8	0,666	5	0,416
179-221 (-)	200	1	0,083	9	0,75	4	0,333
221-263 (-)	242	3	0,25	12	1	3	0,25

2.2.3 Representación gráfica.

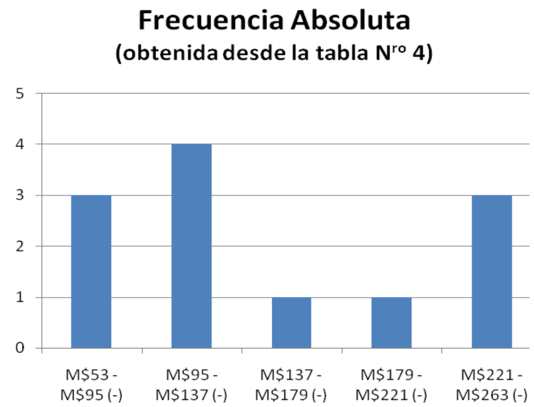
No obstante que las tablas de frecuencia y de clasificación son muy útiles para presentar los datos, pero por sobretodo para su análisis, para una mejor comprensión y visión de ellos, estos se pueden presentar a través de gráficas. Para este efecto, habitualmente son utilizados los siguientes tipos:

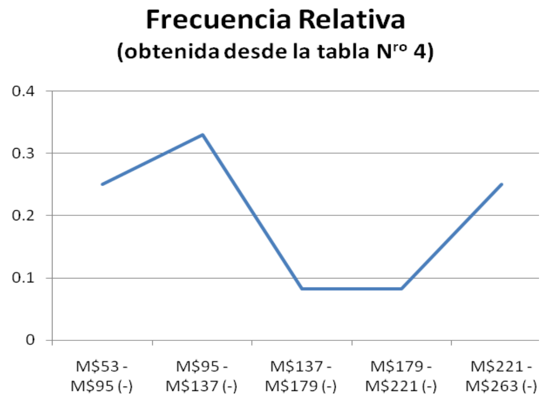
a) **Histogramas:** diagrama que muestra la distribución de frecuencia (absoluta o relativa) a través de un gráfico de barras. Útil para conocer la distribución de los datos desde las categorías construidas.
b) **Polígono de frecuencia:** diagrama que muestra la distribución de frecuencia (absoluta o relativa) a través de un gráfico de líneas rectas. Útil para analizar la proporcion de cada una de las clases entre si.

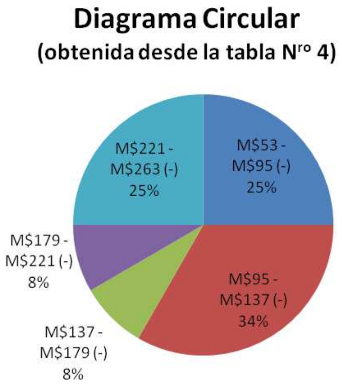
[4] La notación utilizada en los límites superiores de las clases, es útil para indicar la exclusión entre ellas. Así M$ 95 (-) quiere decir que el valor mencionado es en realidad un poco inferior al indicado, como 94.999,99999. Esto implica que un valor de M$ 95, caerá en la segunda clase (95.000 a 137.000).

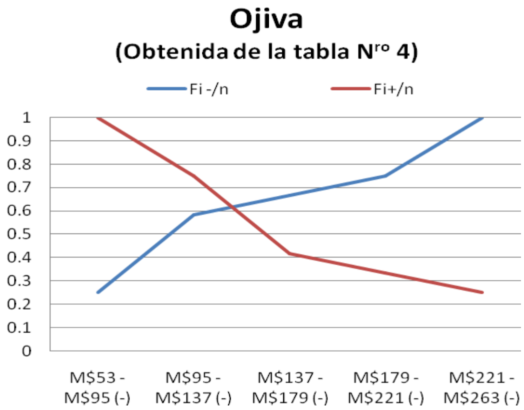
c) **Circular:** diagrama que muestra el porcentaje de datos contenidos en cada clase (participación de cada clase) a través de un gráfico circular o "de torta". Útil para conocer la concentración relativa de cada una de las clases construidas.

d) **Ojiva:** diagrama que muestra la distribución de frecuencia acumulada más de y menos de, a través de un grafico de líneas rectas. Útil para conocer y aproximar el centro de un conjunto de datos.







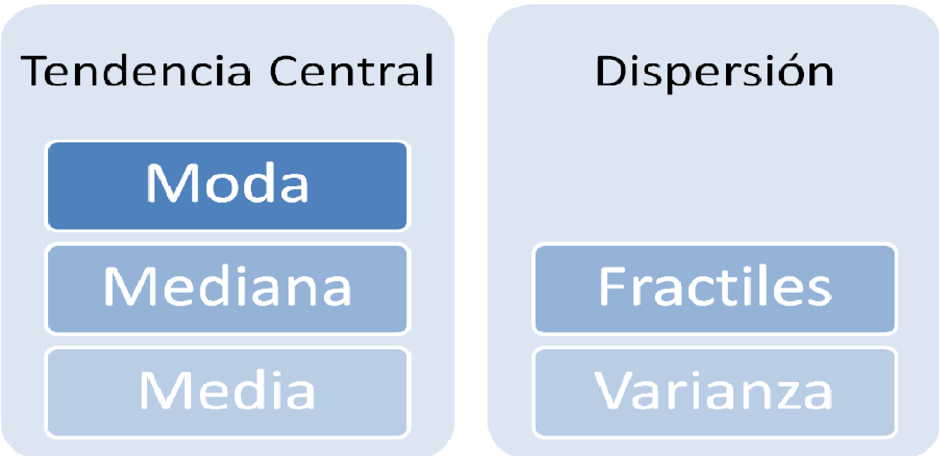


3. Estadígrafos Descriptivos

Los métodos gráficos son útiles para lograr una descripción cómoda de un conjunto de datos. Sin embargo, tanto las gráficas como las tablas tienen limitaciones para la descripción y el análisis de conjuntos de datos, especialmente, en lo que respecta a la inferencia estadística. Esta limitación de los métodos gráficos y tablas puede superarse con el uso de medidas descriptivas numéricas, que pueden utilizar de forma independiente a las gráficas y tablas. A su vez, estos nuevos procedimientos son de mayor facilidad en obtención, elaboración y comprensión, ya que poseen un menor grado de criterio entregando información intersubjetivamente contrastable de forma sencilla y clara.

Un **Estadígrafo** es una medida, calculada en función de los datos observados de un conjunto en particular, que describe o caracteriza algo esencial de ese conjunto de datos, poseyendo de forma intrínseca una serie de propiedades deseables que permiten la descripción de cualquier conjunto entregando información representativa.

En general, podemos clasificar los estadígrafos a través de lo que intentan mostrar o encontrar:



Cabe recordad que existen otros dos estadígrafos de dispersión que son calculados a través de la varianza (Desviación Estándar y Coeficiente de Variación) Sin embargo, los mostrados en el cuadro anterior corresponden a estadígrafos que no necesitan de otros para ser calculados.

Con el fin de desarrollar cada una de estos estadígrafos de forma adecuada, es necesario definirlos de manera conceptual, sus propiedades, relaciones, su procedimiento de cálculo y su aplicación a una situación concreta.

Con el fin de enfatizar las diferencias que existen en los procedimientos de calculo de los estadígrafos dependiendo de la agrupación en la cual se presentan los datos, se presentará la metodología especifica cuando se debe trabajar con datos no agrupados (en bruto) y aquella que se utiliza con datos agrupados (de una tabla de clasificación).

3.1 Estadígrafos de posición o tendencia Central

Con el fin de conocer y analizar un conjunto de datos, es necesario conocer de alguna forma **un dato** que sea capaz de representar del mejor modo posible a todo el conjunto de observaciones.

Cuando es preciso dar alguna información sobre un solo dato, es sumamente sencillo, vasta medirlo y con ello entendemos exactamente cuál es o ha sido su comportamiento. Sin embargo, cómo se puede lograr lo mismo que con un dato, cuando se trata de un conjunto de ellos, los cuales tienen distintos valores para la misma variable.

El fin es encontrar un dato que se encuentre al centro del conjunto de datos y que tenga la capacidad de representarlos a todos, con el menor esfuerzo posible.

Una de las primeras medidas de interés para el análisis de los datos es una medida **de tendencia central** o **de posición**, o sea una medida del centro de la distribución estudiada. Se revisarán algunas de estas medidas:

3.1.1 La moda. (Md)

Con el fin de encontrar a ese dato que se encuentra en el centro o bien que represente de buena manera a un conjunto, una primera aproximación de carácter intuitivo puede ser aquel que tiene mayor ocurrencia dentro de un conjunto de datos.

La **moda** es el valor de la o las observaciones que más se repite (que tiene mayor frecuencia) dentro del conjunto de datos. Si el conjunto de datos tiene una sola moda, se le denomina “monomodal”. De la misma forma, puede ser “bimodal” o “multimodal” según se aprecien dos o más valores modales. Existen dos métodos de cálculo, los cuales dependen de su forma de presentación:

a) Para datos no agrupados: corresponde al valor más repetido del conjunto.
b) Para datos agrupados: corresponde a la **marca de clase** (M_i) de aquella(s) con mayor frecuencia absoluta o relativa.

Ejemplo:

En este caso, no se justifica calcular la moda con datos no agrupados, dada la escala continua de las colocaciones, ya que en este tipo de variables los datos habitualmente son únicos. Con datos agrupados, a partir de la tabla N^{ro} 4, la moda es M$116.000.

De una manera muy sencilla es posible percatarse, que la moda puede no ser (habitualmente no lo es) un buen dato para representar a un conjunto de observaciones por algunas razones:

- Existen conjuntos que no tienen moda o que tienen más de una de ellas.
- No siempre la moda es un dato que se encuentra en el centro del conjunto de datos ya que muchas veces los datos extremos se repiten mas (por ser valores

máximos o mínimos), sin embargo el resto de los datos no se repite tantas veces, pero son cercanos en valor y mayores en cantidad.

- Rara vez existe moda para los datos continuos.

Por lo tanto es necesario encontrar otro representante para el conjunto de datos. Como se sospecha que un buen representante debe estar en el centro del conjunto de datos buscaremos a la observación que se encuentra en el medio del conjunto de datos.

3.1.2 La mediana (Me)

La mediana de un conjunto de datos es el valor de aquella observación que se ubica en el centro del conjunto de datos cuando este ha sido previamente ordenado en forma ascendente o descendente. Esto indica que la observación que corresponde a la mediana supera, en magnitud, a la mitad de las observaciones y es superado por la otra mitad. En palabras familiares es el valor del dato del centro del conjunto. Se calcula, según la forma de presentación de los datos de la siguiente manera:

a) Para datos no agrupados, tenemos dos casos:

- Si el número de observaciones es par:

$$\mathrm{Me} = \frac{\mathrm{X}_{(n/2)} + \mathrm{X}_{(n/2+1)}}{2}$$

Es decir, se escoge como mediana el valor medio entre las dos observaciones del centro.

- Si el número de observaciones es impar:

$$Me = X_{([n+1]/2)}$$

Es decir el dato del medio del conjunto de datos ordenados.

Ejemplo: para el ejemplo anterior, podemos calcular ya que es un conjunto par de datos.

$$Me = \frac{\$110443 - \$123980}{2} = \$117206{,}5$$

b) para datos agrupados:

$$Me = LI_i + (C_i / F_i)(n/2 - F^{-}{}_{i-1})$$

Donde ***i*** es la clase en donde se encuentra la mediana; ***i*-1** es la clase inmediatamente anterior a la clase de la mediana. Por lo que primero que usted debe hacer es encontrar el lugar (la clase) que contiene a la mediana, en términos de observación. Todo esto es metodológicamente válido suponiendo una **distribución uniforme**.

También puede utilizar, ya que supone una distribución uniforme para los datos agrupados, cual debiese ser la proporción de datos que desea encontrar junto con la participación que posee cada uno de los datos en la clase para la variable.
Ejemplo: para el ejemplo anterior, podemos calcular:

$$Me = M\$95.000 + (M\$42.000 / 4)(6-3) = M\$126.500$$

Si bien en apariencia encontramos al candidato perfecto para ser el representante del conjunto de datos, nos damos cuenta de inmediato que habitualmente no se cumple con el principal supuesto que sustenta que la mediana sea igual al centro del conjunto de datos y este es que la distribución o como se ordenan los datos sea uniforme.

Sucede habitualmente que muchos de los datos en cuestión se agrupan en alguno de los extremos (máximo o mínimo), por lo tanto el centro en términos de observaciones, no es necesariamente el centro en términos de valor de los datos de la variable. Por lo que para construir un mejor representante de las observaciones es necesario encontrar a uno que sea capaz de mezclar estos dos aspectos (centro de las observaciones en términos de cuenta y centro de las observaciones en términos de valor).

3.1.3 La media aritmética $(\overline{X})$

La media aritmética, también conocida como el promedio aritmético de un conjunto de ***n*** observaciones (sean correspondientes a la población o a la muestra), es igual a la suma de esas observaciones dividida entre ***n*** (número total de observaciones).

Planteado de otra manera, diferente a su forma de cálculo, es aquel dato que minimiza en valor de forma ponderada (llevando a cero) la distancia que existe entre cada uno de los datos y sí mismo. Lo cual convierte a la media, en el dato que debe realizar la menor cantidad de esfuerzos por convertirse en cualquiera de ellos, convirtiéndose efectivamente en el mejor representante que existe de un conjunto de datos.
En términos de cálculo, para que la distancia entre la media y cada uno de los datos sea cero, es necesario que ***n*** veces la media sea igual a la suma total del valor de las observaciones. En términos de demostración tenemos lo siguiente:

Una de las propiedades que necesitamos para que este dato sea representativo, es que la diferencia de los datos con respecto al promedio sea igual a cero:

$$\sum_{i=1}^{n}(x_i - \overline{X}) = 0$$

Dicha diferencia se puede separa en dos sumas por propiedad de la sumatoria:

$$\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{X} = 0$$

Despejando la diferencia podemos obtener:

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \overline{X}$$

Como el término de la derecha es la suma de una constante tenemos:

$$\sum_{i=1}^{n} x_i = n * \overline{X}$$

Reordenando los términos:

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

Con esta metodología nos encontramos en condiciones de calcular el dato mas representativo de un conjunto de datos.
Clasificando la metodología según la forma de presentar los datos se tiene lo siguiente:

a) para datos no agrupados:

$$\overline{X} = \sum_{i=1}^{n} X_i / n$$

Ejemplo: Para el ejemplo anterior, de las colocaciones de las 12 oficinas, podemos calcular la media aritmética: $\overline{X} = \frac{M\$1.764.644}{12} = M\$147.053,\overline{6}$

b) para datos agrupados: $\overline{X} = (\sum_{i=1}^{P} f_i M_i)/n$, donde P es, la cantidad de clases.

Suponiendo que la **marca** de cada una de las clases, es el mejor representante del conjunto de datos, asumiendo una distribución uniforme de estos.

Si dispone de mejor información o de algún estadígrafo más acucioso o insesgado, por ejemplo la media aritmética de cada clase, debe utilizar éste para el cálculo del promedio del conjunto.

Ejemplo:
Si quisiéramos obtener la media aritmética a partir de la tabla 4, la forma de cálculo debiera ser:

$$\overline{X} = \frac{M\$74*3+M\$116*4+M\$158*1+M\$200*1+M\$242*3}{12} = M\$147.500$$

Este valor de media aritmética difiere del obtenido con los datos no agrupados, debido a la pérdida de información en la que se incurre al agrupar los datos. Recuerde que el cálculo lo realiza a través de la marca de cada clase lo cual no es un dato perfectamente representativo del comportamiento de cada clase (recuerde: asume una distribución uniforme dentro de la clase si utiliza su marca).
Existen ciertas propiedades interesantes de estudiar de la media aritmética, lo que se convertirá de ahora en adelante en el mejor estadígrafo con el cual representar un conjunto de datos en lo que se refiere a la tendencia central que poseen estos:

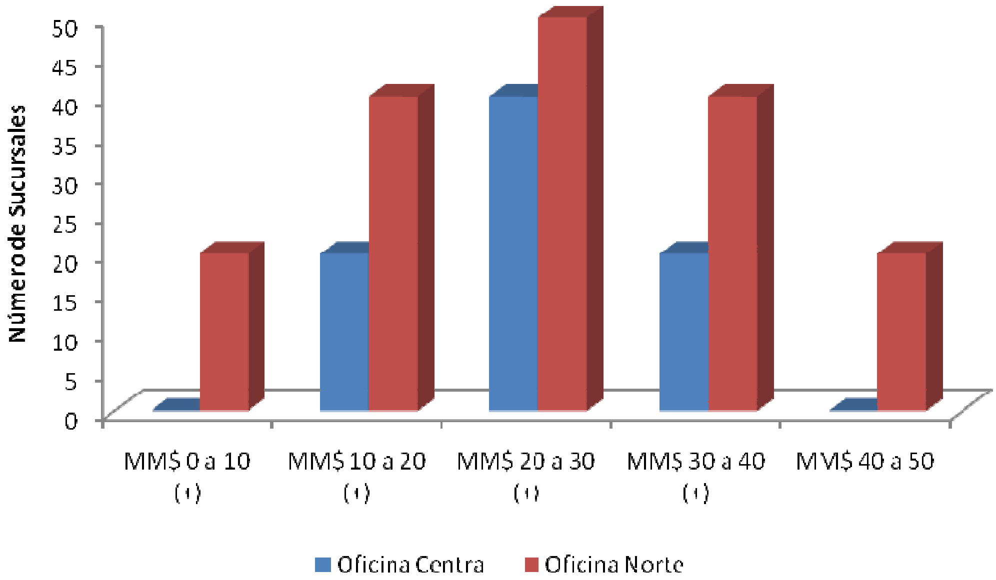
i) Si cada una de las ***n*** observaciones aumenta en "k" unidades, entonces la nueva media aritmética del conjunto de datos también aumenta en "k" unidades.
ii) Si cada una de las ***n*** observaciones es multiplicada por una constante "k", entonces la nueva media aritmética del conjunto de datos es la media aritmética anterior multiplicada por la constante "k".
iii) La suma de las desviaciones (diferencias) de cada dato respecto de la media aritmética es **siempre** cero, es decir: $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$
iv) La **media aritmética** de un conjunto de datos, es igual a la media de las medias de ***n*** subconjuntos de datos creados a partir del conjunto original.

3.2 Estadígrafos de dispersión

Claramente con el dato más representativo de un conjunto de datos, **la media aritmética**, es posible conocer el comportamiento de forma general de los datos. Sin embargo, a priori conocemos que no todas las observaciones toman en un conjunto de datos el mismo valor que la media (cada uno de los datos difiere del promedio utilizado, sin embargo la suma de las diferencias es cero). Por lo que dos grupos con el mismo promedio no necesariamente se comportan de forma similar, ni es posible identificar sus diferencias solamente a través del promedio. Por lo que si bien el promedio es el mejor representante de ellos, no permite diferenciarlo con respecto a otros conjuntos ni caracterizarlo de buena manera en comparación de otros.

Además, dependiendo de cómo se distribuyan los datos existen promedios que realizan más o menos esfuerzos por convertirse en cada uno de las observaciones. Este esfuerzo que debe realizar la media no se visibiliza directamente con la simple observación de esta.

Una vez que se tiene una medida del centro de la distribución (estadígrafo de posición o de tendencia central) representativa (la media), el siguiente paso es obtener una medida de la dispersión o variabilidad de los datos. Si tenemos dos distribuciones, como las mostradas a continuación, que están centradas en un mismo valor pero que muestran un comportamiento de los extremos diferente entre ambos, podemos decir que el comportamiento general del conjunto no es el mismo. Por lo tanto se debe construir una medida que cuantifique esa diferencia entre los conjuntos.



Algunas de estas medidas:

3.2.1 El recorrido

El recorrido de un conjunto de datos es la diferencia entre el mayor y el menor de estos. Ayuda a conocer que tan lejos están los extremos en referencia al promedio encontrado.

Ejemplo: Se tiene un recorrido de M$262.070 – M$53.005=M$209.065.-

Por lo que con esta referencia podemos obtener una primera aproximación de la diferencia que existe entre grupos de observaciones que pueden tener la misma media. Lo que no conocemos con esta medida, donde se encuentra la totalidad de los datos, ya que aun cuando los extremos sean los descritos anteriormente, la mayor parte se encontrar a un extremo o bien la gran mayoría podrían ser exactamente la media y solo 2 datos son los extremos. Por lo que el recorrido no nos entrega suficiente información sobre el comportamiento general de un conjunto de datos.

3.2.2 Los fractiles

Son subdivisiones del conjunto de datos que pretenden de forma ordenada, agrupar en una porción conocida de datos la distribución en términos de valor que poseen estos. En otras palabras pretende conocer el valor que toma ciertos datos conocidos y relevantes de la distribución de un conjunto de datos con el fin de develar la forma en la cual se distribuyen estos y poder obtener conclusiones relevantes. Estas divisiones son fundamentalmente clasificables de forma reconocidas por las siguientes:

- Cuartiles (Q_j): divide el conjunto de observaciones en cuatro sectores de 25% de los datos cada uno de ellos.
- Quintiles (K_j): divide el conjunto de observaciones en cinco sectores de 20% de los datos cada uno de ellos.
- Deciles (D_j): divide el conjunto de observaciones en diez sectores de 10% de los datos cada uno.
- Percentiles (P_j): divide el conjunto de observaciones en cien sectores de 1% de los datos cada uno.

a) Para datos no agrupados se muestran algunos ejemplos:

- Q1: valor de la observación que supera aprox. Al 25% de los datos y es superada por el 75% restante. Equivalente al P(25)
- Q2: valor de la observación que supera aprox. Al 50% de los datos y es superada por el 50% restante. Equivalente al P(50) y Mediana.
- D8: valor de la observación que supera aprox. Al 80% de los datos y es superada por el 20% restante. Equivalente al P(80) y al K(4)
- P10: valor de la observación que supera aprox. Al 10% de los datos y es superada por el 90% restante. Equivalente al D(9)
- P60: valor de la observación que supera aprox. Al 60% de los datos y es superada por el 40% restante. Equivalente al K(3) y D(6)

b) Para datos agrupados, primero debemos encontrar en que clase se debe encontrar el fractil que se busca. Dependiendo de la clasificación se tiene:

Cuartil: $$Q_j = LI_J + (C_J / F_J)(j * n / 4 - F_{J-1}^{-})$$

Quintil $$K_j = LI_J + (C_J / F_J)(j * n / 5 - F_{J-1}^{-})$$

Decil $$D_j = LI_J + (C_J / F_J)(j * n / 10 - F_{J-1}^{-})$$

Percentil $$P_j = LI_J + (C_J / F_J)(j * n / 100 - F_{J-1}^{-})$$

Donde j es el valor del fractil que se busca y J es la clase donde se ubica el fractil buscado.

Los estadígrafos que hemos revisado hasta ahora, solo han sido capaces de reflejar el comportamiento de distintos conjuntos de datos a través de rangos, los cuales no permiten obtener información concreta y comparable entre conjuntos, ni concluyentes para el conjunto en concreto. Un dato que sea capaz de resumir la información sobre la dispersión que tengan estos será vital con el fin de poder comparar, analizar y obtener información útil a la hora de analizar el comportamiento de un objeto.

Usted conoce cuál es el dato más representativo de un conjunto de datos: La Media. Lo que se requiere ahora es calcular de alguna forma la media de la dispersión que tienen los datos de un conjunto.

3.2.3 <u>La varianza (S^2)</u>

Es el promedio del cuadrado de la desviación de cada observación respecto a la media aritmética del conjunto de datos. El objeto de este dato es poder agregar la diferencia que existe por exceso o por defecto con respecto de la media de cada uno de las observaciones. Esta diferencia no se puede rescatar de forma inmediata ya que los datos por defecto y por exceso, se netean entre si ya que la suma de la diferencia entre los datos y el promedio por definición y propiedad es cero.

A partir del conjunto de datos originales podemos calcular un nuevo conjunto referido a las diferencias que existente entre cada uno de los datos del conjunto y el promedio y a este calcular su media, ya que este estadígrafo es el dato más representativo de un conjunto. La razón que justifica la creación de este conjunto es que dicha diferencia es una buena medida de la dispersión de un conjunto ya que compara cada uno de los datos con respecto al centro del conjunto.

Sin embargo, este cálculo de promedio posee un problema, la suma de las diferencias necesario para el cálculo del promedio, es igual a cero siempre, ya que por propiedad de la media del conjunto original la suma de las diferencias debe ser igual a cero. La manera de solucionar aquello, es elevando al cuadrado cada una de las diferencias y después sumarlas, ya que con ello se elimina el signo de la diferencia, lo que permite acumular en una suma el valor de la suma de la diferencia de las observaciones.

a) Para datos no agrupados:

$$S^2 = \sum_{J=1}^{n} (X_J - \overline{X})^2 / n$$

En nuestro ejemplo se tiene: S^2 = M$5.148.835,01(M^2)

b) Para datos agrupados:

Este es un caso para diferenciar algunas propiedades que posee la varianza, si se observa la tabla, podrá apreciar que el dato que se compara con la media es la marca de clase entre cada una de las clases con respecto al promedio.

Cabe recordar que en cada clase los datos no son iguales a la marca, por lo tanto lo que usted asume es que la varianza de cada clase es cero.

Así podemos decir que la varianza de los datos agrupados es:

$$S^2 = (\sum_{J=1}^{P} F_J (M_J - \overline{X})^2) / n$$

En nuestro ejemplo se tiene S^2 = M$4.152.750 (M^2)

Recuerde que la varianza no sirve en sí misma, sino solo es un medio para el cálculo del promedio de las desviaciones entre las observaciones y la media.

3.2.3.1 Propiedades de la Varianza

La varianza tiene algunas propiedades interesantes de estudiar. Supongamos que se conoce la varianza de un conjunto de ***n*** datos, entonces:

i) Si cada una de las ***n*** observaciones aumenta en "k" unidades, entonces la nueva varianza del conjunto de datos es igual a la anterior (no ha sufrido variación).
ii) Si cada una de las ***n*** observaciones es multiplicada por una constante "k", entonces la nueva varianza del conjunto de datos es la varianza anterior multiplicada por la constante "k" al cuadrado.
iii) Si el conjunto total de ***n*** de datos está subdividido en "P" submuestras mutuamente excluyentes de tamaños n_1, n_2, …,n_k respectivamente y con medias aritméticas y varianzas conocidas y diferenciables para cada una de esas submuestras, entonces:

Varianza total	=	Intravarianza	+	Intervarianza
S^2	=	$\sum_{J=1}^{P} W_j S_j^2$	+	$\sum_{J=1}^{P} W_j (\overline{X}_j - \overline{X})^2$

Donde: $W_j=n_j/n$; S^2_j=varianza grupo "j": $\overline{X}_j$ = media aritmética del Grupo "j"

3.2.4 La desviación estándar (S)

La desviación estándar o desviación típica es la raíz cuadrada positiva de la varianza. Notemos que la varianza queda medida en términos del cuadrado de las unidades originales y por lo tanto, la desviación estándar queda en las mismas medidas originales, siendo la primera medida de dispersión representativa del conjunto de datos y con ello se ha encontrado la diferencia media que existe entre las observaciones y el promedio.

Ejemplo: si calculamos la desviación estándar a partir de la varianza calculada antes, se tiene:

- Para los datos no agrupados: S= $71.553,8.-
- Para los datos agrupados: S= $64.441,8.-

3.2.5 El coeficiente de variación (c.v.)

El coeficiente de variación entrega una medida relativa de la desviación estándar de un conjunto de datos, con el fin de poder comparar entre conjuntos de observaciones. Esta es la única forma en la cual es posible comparar la dispersión que tienen los datos de un conjunto de observaciones de forma directa.

$$CV = \frac{S}{\overline{X}}$$

Si calculamos el coeficiente de variación a partir de la media aritmética y la varianza calculada antes, se tiene:

- Para los datos no agrupados: CV= 0,486
- para los datos agrupados: CV= 0,437

La desviación estándar entrega el valor promedio al cual se distribuyen los datos, por lo que esta medida entrega la información en términos relativos, y por lo tanto comparables, de la cercanía o lejanía relativa de la distancia promedio con respecto al valor central. Implica que mientras menor sea el valor que se obtiene en términos porcentuales la variación que existe es menor. En otras palabras, reconoce que la distancia entre los datos y el promedio depende de la unidad de medida y del promedio del conjunto.

EJERCICIOS INTEGRATIVOS

Ejercicio:

En el plan de estudio de la carrera de ingeniería comercial de la U. de Harvardlandia existe una asignatura llamada Decision's Tools. Dado que esa asignatura se encuentra ubicada en el primer semestre cronológico de los alumnos que se matriculan en dicha carrera, se contemplaron dos cursos paralelos 01 y 02, los que han mostrado los siguientes datos al final del semestre:

Tabla 1: Asignatura Decision´s Tools, resumen de calificaciones por paralelo y sexo

Cursos Paralelos	Damas			Varones		
	Cant.	Prom. Notas	desv. St.	Cant.	Prom. Notas	desv. St.
01	20	4,2	1,2	28	3,7	0,8
02	30	3,9	1,1	20	4,2	1,2

Tabla 2: Asignatura Decision´s Tools, resumen global de calificaciones

X (min.)=1,2	X (máx.)=5,7	P(15)=2,1	
P(45)=3,8	P(69)=4,1	P(75)=4,5	P(90)=5,2

Se pide:

a) Calcule los respectivos promedios de notas para damas y para varones matriculados en la asignatura.
b) Compare la dispersión de las notas entre damas y varones en la asignatura. ¿Quiénes tienen notas finales más homogéneas? Justifique.
c) Compare la dispersión de las notas entre el paralelo 1 y el paralelo 2 en la asignatura. ¿Quiénes tienen notas finales más homogéneas? Justifique.
d) La diferencia de notas que se da entre los alumnos es a causa de su sexo o por participar en alguno de los dos paralelos.
e) A partir de los datos de la Tabla 2, reconstruya un histograma para las notas de los alumnos (<u>no es necesario</u> que contemple bases de igual amplitud).
f) A partir de la situación anterior, calcule qué cantidad de alumnos de la mencionada asignatura aprobaron el curso (considere nota 4,0 como mínima de aprobación).

Ejercicio:

Un profesor dedicado a la docencia en la escuela de filosofía, está muy preocupado por el rendimiento que han tenido sus alumnos observando que la mitad de sus alumnos estaría reprobando la asignatura. Sin embargo no conoce ninguna metodología que le indique el comportamiento que tienen los alumnos en su curso.

Se le entrega la lista de notas de 1 a 7 donde la diferencia mínima entre las observaciones es de 0,1 unidades.

2.0; 4.7; 3.8; 4.2; 3.4; 3.8; 4.6; 2.7; 3.0; 4.8; 2.9; 4.1; 4.7; 1.1; 4.9; 4.3; 3.7; 4.2; 5.5; 5.1; 4.3; 4.5; 6.1; 3.3; 1.9; 5.7.

1. ¿De qué tipo son los datos? Cuáles son sus principales características.
2. Construya una tabla de clasificación completa de 2 maneras alternativas.

Ejercicio:

En algún semestre y en algún ramo de la escuela de ingeniería comercial, con el fin de analizar el comportamiento que tienen los estudiantes de dicho curso, se entrega la siguiente información relevante para el caso. En el curso aprobaron el 66,7% de los alumnos inscritos durante dicho semestre, además se conoce que la mediana del curso fue un 4,2.

1. ¿Cuál es el porcentaje de alumnos que aprobando el curso su nota se encuentra debajo la mediana?
2. ¿Cuál debiese ser la moda para este conjunto?
3. Para sus cálculos realizó una serie de supuestos ¿Cuáles fueron?

Ejercicio:

El problema de accidentabilidad vehicular ha ocupado las primeras páginas de los diarios de Santiago. En el año 2003, había 385.388 vehículos circulando por esa capital.

Las cifras de accidentes causados por vehículos particulares en el 2003, así como el costo hospitalario por la atención de las personas accidentadas en ellos, fueron las siguientes:

			Costo hospitalario por accidentado ($)	
Causas Accidentes	**Cdad. de Accidentes**	**Personas Accidentadas**	**Promedio**	**Desviación Estándar**
Atropello	2.693	3.015	33.890	5.800
Colisiones	4.482	10.202	15.040	6.100
Choques	4.682	16.338	21.000	9.550
Volcamiento	271	602		7.250
Otros	225	198	8.344	3.600

Además se conoce que el coeficiente de variación para el conjunto de accidentes es 0,4858088.

1. Encuentre el costo hospitalario promedio por accidentado ocurrido para los volcamientos. (no es necesario encontrar el valor sino dejar expresado la ecuación que lo resuelve en términos de la variable que deseamos encontrar).

Ejercicio:

Una sucursal de un importante banco nacional ubicado en la ciudad de Valparaíso, presenta los datos de la cantidad de dinero prestado por día por la banca minorista (banca personas). A continuación se adjunta la siguiente tabla:

Promedio	Desviación Estándar	Valor Créditos cursados por día	Cantidad de días
$ 47.786	$ 1.703	$ 44.450 - $ 50.000	
$ 54.848	$ 2.694	$ 50.000 - $ 59.500	
$ 63.928	$ 2.604	$ 59.500 - $ 69.800	
$ 75.238	$ 2.810	$ 69.800 - $ 79.200	
$ 81.648	$ 880	$ 79.200 - $ 83.100	

Además se dispone de la siguiente información:

Decil(1)	$ 48.616	Cuartil(1)	$ 52.558	Mediana	$ 60.874	P(95)	$ 81.292

Se requiere:

a) Encuentre la frecuencia absoluta de la ocurrencia que reflejan en la tabla.
b) Con la información anterior y sabiendo que el conjunto es representativo de un año de colocaciones, es posible calcular cual fue la cantidad de dinero que entrego el banco a sus clientes durante un año (la banca trabaja solo de lunes a viernes).
c) El gerente de la sucursal en cuestión afirma que los días con un nivel de ventas similares explican principalmente la desviación estándar del conjunto de datos.
d) El gerente desea conocer además, cual es el promedio y la dispersión de los días donde se vendió más de M$69.800.-

Ejercicio:

El administrado de un video club está preocupado por los ingresos que está generando su local, para lo cual ha rescatado los siguientes antecedentes para lo que lleva de este año:

Cdad. de videos arrendados diariamente	Nº de días	Dispersión de Cdad. de videos diarios	% películas para mayores de 14	Total de videos arrendados
10-50	35	15	60%	1225
51- 100	215	25	55%	17200
101- 200	70	45	15%	11200
201 y más	2	0	50%	s/ información

- El precio de arriendo es de $1.100 por video
- El máximo de videos que se ha arrendado en un día es de 255.

Determine:

1. El administrador desea saber los ingresos diarios que genera este negocio.
2. El administrador está empeñado en enfocarse al mercado de películas para mayores de 14 años, ya que está seguro que es ese tipo de películas las que genera mayor cantidad de ingresos qué opina usted al respeto.
3. Cuál es su opinión respecto de la estabilidad del número de arriendos diarios de este video club, qué implica en este caso particular, una menor o mayor dispersión.

Ejercicio:

En un estudio estadístico sobre los años de antigüedad en la empresa de un grupo de ejecutivos, se determino que la media aritmética era 8,3 años y la desviación estándar era 2,5 años. Comparando ejecutivos del área de producción con los del área de finanzas, ambos grupos presentaros la misma media: 7,5 años, pero la desviación estándar de los de producción era de 2,4 años y la de finanzas 1,8 años.

Si se sabe que esta empresa solo posee ejecutivos en el área de producción, finanzas y en el área comercial y además en cada área posee la misma cantidad de ejecutivos:

1. Cuál es su opinión respecto de la varianza total para los años de antigüedad de los ejecutivos de esta empresa, que es lo que más influye.
2. El sindicato Nº1 de la empresa, propone que el porcentaje de aumento de salarios que se quiere realizar no sea variable por tramos de salarios, sino que sea igual para todos, independiente del monto ganado. La razón que se da es que esta sería la única forma de que la dispersión relativa (CV) se mantuviese constante, característica deseable por los trabajadores de este sindicato. Comente.

Ejercicio:

De acuerdo a una encuesta realizada el año 2010 a 33 alumnos de la escuela de Ingeniería Comercial de la PUCV, se obtuvo la siguiente información respecto al gasto mensual en telefonía móvil y el tipo de contrato utilizado, información que se presenta en la siguiente tabla:

			Prepago		***Plan***
Gasto Mensual en TM($)	**Gasto Total**	**Estudiantes**	**% Estudiantes**	**Desv. Estándar**	**Desv. Estándar**
1,000 - 5,000 +	$34.000	10	100%	$1.158	
5,000 - 10,000 +	$126.400	14	50%	$1.457	$940
10,000 - 15,000 +	$75.000	6			$1.500
15,000 - 20,000	$55.300	3			$2.216

El grupo de alumnos elegido para realizar esta encuesta fue hecho de a cuerdo a todos los procedimiento de selección de objetos muestrales con el fin de mantener del conjunto elegido la representatividad de los 665 alumnos matriculados a noviembre de 2010 en la carrera. El gasto total refiere al total pagado mensualmente incluyendo todos los cargos que las distintas compañías de telefonía móvil existentes cobran por la utilización de sus servicios tales como cargos fijos, arriendo de celulares y llamadas realizadas. Además, se logró obtener que los estudiantes que utilizan prepago y gastan entre M$5 y M$10, tienen un promedio de gasto de $8.857 pesos en promedio. Por último, las únicas dos formas de contrato para telefonía móvil existentes en dicho periodo y mutuamente excluyentes son el prepago y el plan.

Se pide:

a) ¿Cuánto gastan en promedio los estudiantes que utilizan plan? Interprete claramente el resultado.
b) ¿Es cierto que la mitad de los estudiantes gasta menos de $7,500 pesos en telefonía móvil?
c) De su respuesta anterior, alguien rebate diciendo "*en realidad no es posible saber lo que usted acaba de afirmar*". ¿Por qué quien le rebate tiene toda la razón de dudar?
d) Cuál es su opinión acerca de la dispersión del gasto de los estudiantes en telefonía móvil.
e) Proviene la dispersión de la diferencia que existe entre los estudiantes dependiendo del tipo de contrato (plan o prepago) o de las diferencias entre lo que pagan efectivamente (de menor a mayor pago). Responda fundadamente.
f) La escuela, preocupada por las diferencias que existen en el gasto en telefonía móvil, dispondrá de $10,000 para aumentar el gasto de los estudiantes en telefonía móvil, con el fin de disminuir la dispersión existente. ¿Explique las

razones por las cuales debiese entregárselas equitativamente a los que gastan entre M$1 y M$5 para cumplir su objetivo? (no pueden dejar de gastar de sus recursos en telefonía y la beca es sólo para gastos en telefonía móvil).

Ejercicio:

En la exclusiva universidad de Harvardlandia, los alumnos tienen solo dos posibilidades en lo que respecta al pago de la matricula semestral: pago en un sola cuota la comienzo del semestre, o bien, obtener un crédito estatal por un determinado porcentaje del valor de la matricula y pagar solo la diferencia en una sola cuota al comienzo del semestre. La universidad no otorga becas de matrícula, aun cuando mantiene un gran número de servicios de bienestar para alumnos.

Al comienzo de un determinado semestre ya se cuenta con los resultados de las solicitudes de crédito estatal para el semestre en cuestión y se tienen los siguientes datos para ese semestre en la carrera X:

Carrera X	
Valor matrícula semestral	120 (miles $)
Nº alumnos regulares	150
Nº alumnos solicitantes crédito	128
Monto total crédito otorgado	9.060 (miles $)
Coef. de Variación de los montos de créditos individuales asignados a alumnos solicitantes	62%

Nota: el cálculo de CV incluye a los 30 alumnos solicitantes que no obtuvieron crédito estatal (0% del crédito). Ningún alumno obtuvo 100% del crédito.

Se pide (asumiendo que no habrá deserción de alumnos):

1. Calcular la desviación estándar de los montos individuales de crédito estatal asignados a los alumnos que obtuvieron crédito estatal en la carrera X.
2. Calcular la cuantía media por alumno del pago de matricula a efectuar a comienzos de semestre por los alumnos de la carrera X.
3. Calcular una intervarianza de los pagos de matricula a efectuar a comienzos de semestre por los alumnos de la carrera X.
4. Comente: "En una facultad de esta universidad la variante total de los montos individuales de crédito estatal asignados a los alumnos solicitantes es $S^{2,}$ con intravarianza igual a la intervarianza" (considerando a las 3 carreras que la conforman como grupos individuales).

Ejercicio:

En un país lejano existen dos ciudades balnearios, La Vina y La Sere, que compiten todos los años por atraer turistas nacionales y extranjeros.

Existen dos grandes problemas en estos balnearios, por un lado el grado de contaminación de las platas y por otro, la gran cantidad de asaltos que se han presentado en ambas ciudades. El ministerio de salud y medio ambiente realizo un estudio del contenido de coniformes en el agua de las distintas playas de ambas ciudades, lo que se presenta a continuación:

Coliformes por MT.	Nº de playas LA VIÑA	Nº playas LA SERE
0 – 100 (+)	2	4
100 – 300 (+)	11	6
300 – 1200 (+)	9	13
1200 – 2500 (+)	3	9
2500 – 7000 (+)	7	8

Según las normas dictadas por el ministerio, se clasifica a una playa como contaminada si siente 900 o más coniformes por metro de agua, y como libre de contaminación si contiene 450 o menos coniformes por metro de agua.

En cuanto al número de asaltos durante la temporada de verano en las playas de la viña se sabe lo siguiente:

P (10) = 30 D (4)=45 Me=65 P (80)=80

Además, se sabe que en la playa más asaltada de la viña se produjeron 96 asaltos durante la temporada. Por otro lado, en La Sere, el promedio de asaltos por playa durante la temporada es de 48 y la dispersión es de 13.

Se solicita a usted, verifique una serie de declaraciones que se han estado formulando y las comente fundamentando con cálculos.

1. Las autoridades de La Sere señalan “tenemos menos playas contaminadas que en La Vina”
2. Las autoridades de La Vina afirman lo siguiente (en promedio las aguas de nuestras playas son más limpias que en La Sere, e incluso las playas contaminadas en promedio contienen menos coniformes”
3. El ministerio ha diseñado un ingenioso sistema para limpiar el agua del mar. en términos monetarios implica que se puede reducir 100 coliformes por metro con un costo de 320 unidades monetarias en cada playa. Si se desea dejar a las playas contaminadas libres de contaminación (450 coliformes por metro de agua) y solo se puede aplicar a una ciudad, en cuál significa un menos costo para el ministerio
4. El ministerio afirma que no existen diferencias significativas entre las distintas platas del país en cuanto a la cantidad de coniformes por MT. En las agua. Además cree que la dispersión solo se explica por las diferencias en los niveles de contaminación de las playas pertenecientes a cada ciudad.
5. determine en cuál de las ciudades se produce una mayor cantidad de asaltos durante a temporada.

Ejercicio:

1. Dos baterías de auto de distinta marca tienen el mismo precio. Se sabe que para la variable "duración" $\bar{x}_a = 700 días$ y $Me_a = 800 días$; $\bar{x}_b = 800 días$ y $Me_b = 700 días$, siendo las varianzas de la duración aproximadamente iguales para ambas baterías, tras efectuar un gran numero de observaciones. Entonces, dados tales datos, un consumidor debiera preferir la marca B.

2. El 80% de las empresas de un sector industrial mantiene deudas individuales iguales o inferiores a 19.000UF y el 35% de las empresas de ese mismo sector industrial mantiene deudas individuales iguales o inferiores a 5.000Uf. Además, la deuda individual mediana de las empresas de ese sector asciende a 12.000UF, la empresa mas endeudada debe 22.000UF y un 3% de las empresas de ese sector no tiene deudas.
   Calcular la deuda media por empresa y la deuda media por empresa deudora que prevalecería si, inmediatamente, aquellas empresas que deben 8.000UDF o menos liquidasen, cada una de ellas, el 40% de su respectiva deuda.