Estadística y Probabilidad I
Estadística Interactiva en la Red.
Laboratorio Virtual de Estadística.

Contenidos Teóricos Unidad Temática 1.
Estadística Descriptiva Unidimensional.

A. Gámez, L.M. Marín, R. Rodríguez y S. Fandiño

Noviembre - 2005

Índice general

Capítulo 1

Estadística Descriptiva Unidimensional

"El pensamiento estadístico será un día tan necesario para el ciudadano eficiente como la capacidad de leer y escribir".
(H.G. Wells)
"Llegará un día en que la Estadística ocupe en la enseñanza un puesto ligeramente posterior al de la Aritmética"
(L. H. C Tippett, 1947). (Discípulo de Fisher y de Pearson)

1.1. Introducción a la Estadística

1.1.1. Concepto

Desde un punto de vista muy primitivo usamos la estadística continuamente en nuestra vida. A veces oímos frases como las siguientes: "No voy a comprar todavía un ordenador porque espero que baje de precio", "No voy a salir a las diez de la noche porque es casi seguro que aún no habrá salido ninguno de mis amigos", "El precio de las casas subirá para el año que viene más del 10 %"... Expresamos opiniones sobre muchos temas: Los trenes llegan a menudo con retraso, las mujeres conducen peor que los hombres, ciertos profesores suspenden mucho. También hay opiniones sobre algunos temas que se discuten en las conversaciones y a veces en los medios de comunicación: ¿Los catalanes son peseteros?, ¿Los andaluces somos vagos? Las respuestas a estas preguntas dependen de la experiencia personal de cada individuo, que la interpreta subjetivamente según su estado de animo, su situación social, su educación o su ideología. Como es natural, es difícil ponerse de acuerdo. Algo similar sucede en el campo de las ciencias experimentales. Si una cierta propiedad se presenta en un número finito de experimentos el científico pretenderá declarar esta propiedad experimental en forma de ley general. Pero los experimentadores pueden cometer errores en la realización de experimentos, el material puede sufrir variaciones. Si el experimentador desea comprobar una hipótesis en la que confía, inconscientemente tenderá a dar más importancia a los datos que la corroboran que a los que la rebaten. Por lo tanto deberá seleccionar los datos e interpretarlos de una forma que no sea subjetiva. En todos los casos comentados hay una idea general: se dispone de una información particular que deseamos generalizar.

En situación similar se encuentra el economista, que disponiendo de datos anteriores, desea hacer previsiones sobre las subidas de interés o la variación del índice de precios de consumo, las compañías de seguros que necesitan actualizar los precios de las pólizas, los empresarios que desean organizar la producción de sus fábricas, etc...

La Estadística nos va a ayudar a seleccionar las conclusiones generales más adecuadas a partir de datos parciales y representativos. Distinguiremos los tres campos básicos de la Estadística: La estadística descriptiva, el cálculo de probabilidades y la estadística inferencial.

La **estadística descriptiva** trata del estudio de los datos particulares (la **muestra**). La **estadística inferencial** se ocupa de lo referente a la selección de las conclusiones generales. Pero como estas conclusiones dependen de la muestra considerada, tendremos que considerar la probabilidad de error que se origina por la selección de una muestra inadecuada por no ser suficientemente representativa. Cuestiones de este tipo son las que se resuelven por medio del **cálculo de probabilidades**.

En los razonamientos estadísticos se emplea con frecuencia el **método inductivo**. Existen dos formas principales de pensamiento lógico: deductivo e inductivo. El pensamiento deductivo se debe principalmente a los griegos. Consiste en proponer axiomas, hechos admitidos, y deducir de ellos otras propiedades. El razonamiento inductivo, que es el más usado en las aplicaciones estadísticas, nos conduce a inferir conclusiones generales a partir de hechos experimentales.

En el razonamiento deductivo, usado frecuentemente en Matemáticas, los teoremas se deducen de los axiomas siguiendo las leyes de la Lógica. En este sentido son absolutamente ciertos. En cambio, en el razonamiento inductivo las conclusiones tienen un cierto grado de incertidumbre.

La base del razonamiento inductivo es admitir que los fenómenos de la naturaleza son demasiado complejos para permitir una información completa, así que no podemos recolectar toda la información y debemos contentarnos con la información parcial suministrada por una **muestra**. Las cuestiones principales que uno puede hacerse sobre las muestras son las siguientes:

¿Cómo se describe una muestra de forma útil y clara?

¿Cómo sacar conclusiones a partir de una muestra que sea generalizable al colectivo total?

¿Hasta qué punto son de fiar estas conclusiones?

¿Cómo se deben tomar las muestras para que realicen las funciones anteriores de la forma más eficaz posible?

La materia que responde a la primera pregunta es la Estadística Descriptiva. Este es el tipo de Estadística más divulgado por los medios de comunicación: tablas donde se resumen los datos, gráficas más o menos sugerentes y quizá, algunos valores promedios. También es interesante en esta fase dar algún parámetro que nos indique si los datos son entre sí más o menos parecidos. A estos parámetros se les suele llamar medidas de dispersión. La gente piensa frecuentemente que este mero reflejo de una realidad observada es el único papel de la Estadística. Sin embargo éste es sólo el primer escalón. Estaríamos en la fase que hemos llamado experimental en el método inductivo y aún quedaría la fase consistente en establecer las conclusiones (Estadística Inferencial) y determinar el grado de fiabilidad de estas conclusiones (Cálculo de Probabilidades).

La respuesta a la última pregunta que hemos formulado sobre las muestras se estudia en una rama de la Estadística que se llama Teoría de Muestras. Es fácil darse cuenta que algunas muestras no serán reflejo de la realidad global que pretendemos investigar. Si

deseamos estimar el sueldo medio de los trabajadores gaditanos no sería adecuado tener en cuenta solamente el sueldo de las personas que viven en los barrios residenciales, o si queremos tener una idea de la salud de los españoles, no sería lógico investigar exclusivamente en un único hospital. La muestra debe ser representativa de la población que se pretende investigar. Como no todas las muestras van a ser igualmente representativas tendremos que investigar cómo pueden variar todas las posibles muestras entre sí. Otra rama de la Estadística, el Diseño de Experimentos, nos indica cómo deben diseñarse las muestras para extraer la mayor cantidad posible de información minimizando el esfuerzo requerido en la extracción de la muestra.

En resumen: La Estadística es la ciencia cuyo objetivo es reunir una información cuantitativa o cualitativa referente a individuos, grupos, series de hechos, etc. analizar los datos obtenidos y deducir, a partir de este análisis y mediante técnicas propias, conclusiones generales o previsiones para el futuro con un cierto grado de incertidumbre.

1.1.2. Aplicaciones de la estadística

Una de las utilidades de la estadística es dar servicio a los estados. La Estadística influye en las decisiones de los gobiernos y las administraciones estatales. Con sus técnicas, los estados pueden conseguir un conocimiento claro de la población con la que cuenta, permitiéndole, por ejemplo, establecer su política fiscal. En relación con este hecho el estado organiza numerosas encuestas para el conocimiento de la población como, por ejemplo, el Censo de población. En España, como en otros paises, se realizan numerosas encuestas a nivel nacional: la Encuesta de población activa (EPA), la Encuesta de Presupuestos Familiares, las encuestas del CIS...

La práctica del recuento de la población y de algunas características de ésta por los Estados es muy antigua (los faraones egipcios lograron recopilar, hacia el año 3050 antes de Cristo, muchos datos sobre la población y la riqueza de su país). En concreto, la palabra estadística deriva de la palabra estado.

La Estadística colabora en el conocimiento de las necesidades de la sociedad, permitiendo planificar los servicios sociales de ésta, como los hospitales, las subvenciones, las necesidades asistenciales, etc. Para estudiar estas necesidades sociales normalmente se recurre a encuestas, ya que investigar a toda la población sería lento y caro. Las técnicas estadísticas de Muestreo, permiten obtener muestras válidas para que las conclusiones sean extrapolables a toda la población.

Las técnicas de Investigación de Mercados permiten planificar la producción y saber si un producto nuevo, o un nuevo centro comercial va a ser viable económicamente, estudiando el numero de personas que prefieren ese tipo de productos o estarían gustosos de comprar en este tipo de establecimiento. También se puede conocer la audiencia en Televisión y Radio, para ver el impacto esperado de las campañas publicitarias.

En Medicina se emplea la estadística para seleccionar, entre un conjunto de tratamientos para una enfermedad, el mejor posible.

Con el estudio de las Series Temporales se puede tener una mejor comprensión del comportamiento aleatorio de los fenómenos meteorológicos que pueden ayudar en la previsión del tiempo, hacer pronósticos sobre el comportamiento en bolsa de ciertas acciones, de las fluctuaciones de las ventas, etc.

En cuanto a la Industria, el Control Estadístico de Calidad permite seguir la calidad de un producto en todas las fases de la cadena de producción dentro de la fábrica, tomando decisiones correctivas si procede. Esto permite conseguir un producto, no sólo de mejor

calidad, sino incluso más barato. También son útiles en la Industria los estudios estadísticos sobre la duración sin fallos de los productos una vez que están siendo usados por el consumidor, para lo cual se emplean las técnicas estadísticas de Fiabilidad, que permiten, entre otras cosas, establecer los periodos ofertados de garantía para el producto, evaluando el costo total esperado para la fábrica por este concepto. Asimismo, en Agricultura se aplican técnicas estadísticas para estimar los rendimientos obtenidos en una cosecha, o seleccionar qué producto será más rentable económicamente en el mercado.

La Estadística es en la actualidad una herramienta auxiliar para todas las ramas del saber; inclusive en Lingüística se aplican técnicas estadísticas, para atribuir un escrito a un cierto autor o para establecer las características propias de un idioma. Su utilidad se entiende mejor si tenemos en cuenta que los quehaceres y decisiones diarias embargan cierto grado de incertidumbre y que la Estadística ayuda a tomar las decisiones más adecuadas en cada situación reduciendo el grado de incertidumbre.

1.1.3. Notas Históricas

Los comienzos de la Estadística pueden situarse en el antiguo Egipto, cuyos faraones recopilaron, hacia el año 3050 antes de Cristo, una gran cantidad de datos sobre la población y la riqueza del país. Así que podemos decir que la Estadística es más antigua que las pirámides de Egipto. Se cree que este registro de la riqueza y de la población se hizo, precisamente, con el objetivo de preparar la construcción de las pirámides.

El libro bíblico de *Números* da referencias de dos censos de la población de Israel y el de *Crónicas* describe el bienestar material de las diversas tribus judías. En China existían registros numéricos similares con anterioridad al año 2000 A.C.

Los griegos clásicos realizaban censos cuya información se utilizaba para recabar información tributaria y estimar los elementos de la población susceptibles de ser militarizados en caso de guerra.

Los romanos fueron, entre los pueblos antiguos, quienes mejor supieron emplear los recursos de la Estadística. Cada cinco años realizaban un censo de la población y sus funcionarios públicos anotaban nacimientos, defunciones y matrimonios y realizaban registros periódicos del número de cabezas de ganado y de las riquezas contenidas en las tierras conquistadas. Para el nacimiento de Cristo sucedía uno de estos empadronamientos de la población bajo la autoridad del imperio, como queda registrado en los evangelios.

Aunque Carlomagno, en Francia, y Guillermo el Conquistador, en Inglaterra, intentaron recobrar estas costumbres romanas, los métodos estadísticos permanecieron casi olvidados durante la Edad Media.

Durante los siglos XV, XVI, y XVII, hombres como Leonardo de Vinci, Nicolás Copérnico, Galileo, Neper, William Harvey, Sir Francis Bacon y René Descartes, hicieron grandes aportaciones al método científico, de tal forma que cuando se crearon los Estados Nacionales y surgió el comercio internacional existía ya un método capaz de aplicarse a los datos económicos.

En la primera mitad del siglo XVI en Francia se exigió a los clérigos registrar los bautismos, fallecimientos y matrimonios. Durante un brote de peste que apareció a fines de la década de 1500, el gobierno inglés comenzó a publicar estadísticas semanales del número de muertes y sus causas. Esa costumbre continuó muchos años, y en 1632 estos *Bills of Mortality* (Cuentas de Mortalidad) contenían los nacimientos y fallecimientos por sexo. En 1662, el capitán John Graunt usó estos documentos, que abarcaban treinta años, y efectuó predicciones sobre el número de personas que morirían de varias enfermedades y

sobre las proporciones de nacimientos de varones y mujeres que cabía esperar. El trabajo de Graunt, condensado en su obra *Natural and Political Observations...Made upon the Bills of Mortality* (Observaciones Políticas y Naturales ... Hechas a partir de las Cuentas de Mortalidad), fue un esfuerzo innovador en el análisis estadístico. Los eruditos del siglo XVII cultivaron la Estadística Demográfica para responder a la cuestión de saber si la población aumentaba, decrecía o permanecía estática.

El primer empleo de los datos estadísticos para fines ajenos a la política tuvo lugar en 1691 y estuvo a cargo de Gaspar Neumann. Este investigador se propuso acabar con la creencia popular de que en los años terminados en siete moría más gente que en los restantes. Para ello investigó los archivos parroquiales de la ciudad. Después de revisar miles de partidas de defunción pudo demostrar que en tales años no fallecían más personas que en los demás. Los procedimientos de Neumann fueron conocidos por el astrónomo inglés Halley, descubridor del cometa que lleva su nombre, quien los aplicó al estudio de la duración de la vida humana. Sus cálculos sirvieron de base para las tablas de mortalidad que hoy utilizan todas las compañías de seguros.

A mediados del siglo XVII los juegos de azar eran frecuentes en los salones europeos. El caballero De Méré, jugador empedernido, consultó al famoso matemático y filósofo Blaise Pascal (1623-1662) para que le revelara las leyes que controlan el juego de los dados, el cual, interesado en el tema, sostuvo una correspondencia epistolar con Pierre de Fermat (1601-1665) dando origen a la teoría de la probabilidad, que llegaría a constituir la base primordial de la Estadística.

Durante el siglo XVII y principios del XVIII, matemáticos como Bernoulli, Francis Maseres, Lagrange y Laplace desarrollaron la teoría de probabilidades. No obstante durante cierto tiempo, la teoría de las probabilidades limitó su aplicación a los juegos de azar y hasta el siglo XVIII no comenzó a aplicarse a los grandes problemas científicos.

Jacques Quételet (1796-1874) interpretó la teoría de la probabilidad para su uso en las ciencias sociales. Quételet fue el primero en realizar la aplicación práctica de todo el método Estadístico entonces conocido, a las diversas ramas de la ciencia.

En el período del 1800 al 1820 se desarrollaron dos conceptos matemáticos fundamentales para la teoría Estadística; la teoría de los errores de observación, aportada por Laplace y Gauss; y la teoría de los mínimos cuadrados desarrollada por Laplace, Gauss y Legendre. En 1840 Sir Francis Galton partió de una distribución discreta y la fue refinando hasta llegar a una continua similar a la normal. Incluso inventó una máquina que permite ilustrar la distribución normal.

El cálculo de probabilidades se comenzó a usar en demografía y en la matemática actuarial. La mecánica estadística, que introdujeron Maxwell (1831-1879) y Boltzman dio una justificación a la distribución normal en la teoría de los gases. A finales del siglo XIX, Quételet aplicó por primera vez análisis estadísticos en biología humana y Sir Francis Galton, primo de Darwin, estudió la variación genética humana usando métodos de regresión y correlación. De aquí partió, ya en el siglo XX, el desarrollo del coeficiente de correlación creado por Karl Pearson y otros cultivadores de la ciencia biométrica inglesa.

Durante los siglos XVIII y XIX la estadística se desarrolló muchísimo, a pesar del estancamiento de la teoría de probabilidades debido a que no se disponía de una definición general de probabilidad. Esta definición no se lograría hasta que Andrey Nickolaevich Kolmogorov (1903-1987) definiera axiomáticamente la probabilidad, hecho que marca el nacimiento de la Estadística Matemática moderna y convierte a la Estadística en una ciencia independiente, ya que hasta este momento su estudio estaba ligado muy íntimamente con sus aplicaciones a las otras ciencias.

En la primera mitad de este siglo Gossett, con el seudónimo de Student, estudió el problema del tratamiento de pequeñas muestras desarrollando el test de Student y Fisher desarrolló el Análisis de la Varianza, de gran interés en el diseño de experimentos.

La segunda mitad del siglo XX destacan los trabajos de Wilcoxon, que estudió los pesticidas desarrollando un test no paramétrico para comparar dos muestras, de Kruskal Wallis, que aportó un test no paramétrico para comparar más de dos muestras, de Spearman y Kendall que desarrollaron sendos coeficientes de correlación no paramétricos, de Tukey, que desarrolló procedimientos de comparación múltiple...

La llegada de los ordenadores revolucionó el desarrollo de la estadística, propiciando la aparición de nuevas técnicas. Benzecri, en Francia, y Tukey, en Estados Unidos, fueron pioneros en repensar la estadística en función de los ordenadores, adaptando, mejorando y creando nuevos instrumentos, técnicas analíticas y gráficas para estudiar una gran cantidad de datos.

1.1.4. Paquetes estadísticos

La sociedad genera una gran cantidad de información que necesita dar a conocer, resumir, interpretar y emplear para tomar decisiones. Con el avance de la Informática y la vinculación de ésta con la Estadística se ha conseguido manejar de manera rápida, fiable y relativamente sencilla estos volúmenes de información, y obtener conclusiones a partir de esta información. En la actualidad, y con la ayuda de la informática, la estadística ha dejado de ser patrimonio exclusivo del estado y de científicos brillantes, pasando a impregnar la sociedad en las vertientes económica, social, industrial, sanitaria, etc. En esta extensión de la estadística, la informática ha jugado un papel fundamental, propiciando el uso de paquetes estadísticos. En el mercado existen paquetes estadísticos muy completos. Destacamos algunos de los más utilizados: STATGRAPHICS, SPSS, SAS, STATISTICA, EXCEL SOLVER, S-Plus y R (éste último gratuito).

1.2. Estadística Descriptiva Unidimensional

1.2.1. Conceptos básicos

Damos, en primer lugar, algunas definiciones básicas de interés general y que nos ayudarán a clasificar los tipos datos que se nos presenten.

Población: Conjunto sobre el cual se va a realizar la investigación. Está compuesta por elementos. Puede ser de tamaño finito o infinito.

Muestra: Subconjunto de la población del que se dispone de información necesaria para realizar el estudio.

Caracteres: Cualidades o propiedades de los elementos de una población que son objeto del estudio. Atendiendo a que sean o no medibles, los caracteres se pueden clasificar en **cuantitativos** (o variables) y **cualitativos** (o atributos). Las variables cuantitativas pueden ser a su vez discretas o continuas.

$$\text{caracteres} \begin{cases} \text{cuantitativos} \begin{cases} \text{variables discretas} \\ \text{variables continuas} \end{cases} \\ \text{cualitativos} \end{cases}$$

Ejemplo: Supongamos que se desea investigar ciertas características de los alumnos de un instituto. Se han seleccionado al azar 50 de ellos para realizar una encuesta.

Se ha registrado para cada uno de los alumnos seleccionados: su talla, el tipo de estudios realizados por su padre, el número de personas que conviven en su domicilio y su peso.

Los datos están registrados en la tabla 1.1 de la página 9. En este caso la población está formada por todos los alumnos del instituto, la muestra por los 50 alumnos seleccionados. Los caracteres seleccionados son: talla, estudios del padre, número de personas que viven en su domicilio y el peso. La talla y el peso son caracteres cuantitativos continuos, los habitantes de la casa es un carácter cuantitativo discreto y los estudios del padre es un carácter cualitativo o atributo.

1.2.2. Tipos de muestreo

A la hora de decidir sobre la forma de recoger la información de la muestra se utilizan distintos criterios, originando distintos tipos de muestreo.

Muestreos aleatorios

Se seleccionan los elementos de la muestra por un procedimiento de azar (un sorteo). El investigador no decide qué elementos van a tomar parte de la muestra, aunque debe conocer la probabilidad de selección de cada elemento. Estos tipos de muestreo permiten aplicar las técnicas de inferencia estadística. Entre ellos se usan los siguientes:

Muestreo aleatorio simple con y sin reemplazamiento: Todos los elementos de la Población tienen la misma probabilidad de ser incluido en la muestra y la selección de cada uno de los elementos es independiente de la selección de otro. Si cuando se extrae un elemento de la Población para formar parte de la muestra, ya no puede extraerse de nuevo (no se reemplaza en la Población) el muestreo se llama *Muestreo aleatorio simple sin reemplazamiento.* Si por el contrario se devuelve a la Población y puede formar de nuevo parte de la muestra, el muestreo se dice *Muestreo aleatorio simple con reemplazamiento.*

Muestreo estratificado: Este muestreo requiere que la Población esté dividida en grupos más o menos homogéneos con respecto a la característica que se investiga. A cada uno de estos grupos se le llama *clase o estrato.* Dentro de cada uno de estos estratos se selecciona la muestra con un muestreo aleatorio simple. La muestra que resulta se llama una *muestra estratificada.*

Muestreo por conglomerados o Agrupado: Consiste en dividir la población en grupos parecidos entre sí y seleccionar aleatoriamente un conjunto de estos grupos. Para que sea eficiente los grupos han de ser bastante parecidos entre sí, ya que todos ellos han de ser modelos en miniatura de la población. La diferencia de un grupo con un estrato consiste en que los estratos han de ser diferentes entre sí, aunque homogéneos interiormente. Sin embargo, los grupos son parecidos entre sí, pero interiormente reflejan la variabilidad de la población de la que proceden.

Muestreo Sistemático: Se supone que los elementos de la población están ordenados de alguna forma. Se selecciona sucesivamente los elementos de k en k, comenzando por un elemento seleccionado aleatoriamente.

TALLA	ESTUDIOS	HABITANTES	PESO
1.63	bachiller	2	54.88
1.44	fp	3	33.5
1.9	bachiller	3	89.38
1.58	bachiller	3	53.28
1.38	bachiller	5	36.62
1.8	bachiller	6	79.47
1.64	superior	3	72.47
1.93	bachiller	5	85.26
1.95	bachiller	2	102
1.59	bachiller	2	62.28
1.78	primario	4	81.17
1.96	primario	4	100.61
1.89	bachiller	5	94.43
1.52	bachiller	4	57.96
1.74	diplomado	4	67.86
1.68	diplomado	5	63.2
2	diplomado	6	105.01
1.67	diplomado	3	66.57
1.46	primario	5	43.46
1.98	primario	4	96.4
1.47	primario	2	42.38
1.74	primario	3	80.41
1.9	diplomado	4	92.7
1.65	fp	5	62.81
1.34	fp	8	39.707
1.75	diplomado	2	78.951
1.68	diplomado	3	67.15
1.72	bachiller	4	71.94
1.65	bachiller	4	64.1
1.8	bachiller	6	76.68
1.94	bachiller	4	96.18
1.99	primario	4	103.029
1.36	primario	3	44.105
1.69	primario	5	70.16
1.69	bachiller	4	66.79
1.51	bachiller	3	57.27
1.98	bachiller	5	94.71
1.84	bachiller	3	81.25
2.02	bachiller	6	101.3
1.76	primario	4	73.68
1.96	fp	3	90.7
1.78	fp	3	84.59
1.54	superior	4	53.97
1.8	diplomado	5	82.13
1.53	diplomado	4	52.8
1.74	diplomado	5	77.22
1.7	primario	3	74.7
1.66	primario	4	69.34
1.83	primario	5	92.24
1.52	bachiller	4	61.05

Cuadro 1.1: Encuesta entre 50 alumnos de un instituto

Muestreo Doble, Múltiple y Secuencial: Este tipo de muestreo se usa principalmente en Control de Calidad.

El muestreo Doble es un procedimiento mediante el cual se selecciona en primer lugar una muestra pequeña. Si la información obtenida con esta muestra nos parece suficiente, hemos terminado. Si esto no fuera así se procede a tomar una segunda muestra, normalmente más grande con la que completaremos la información. En el muestreo múltiple este procedimiento se repite sucesivamente un número finito de veces. Una modificación de este tipo de muestreo múltiple consiste en decidir para cada elemento que se incorpora a la muestra si tomamos un siguiente elemento o ya la muestra extraída es suficiente para nuestro propósito. El número de elementos de la muestra no es conocido a priori, ya que dependerá de la propia muestra ya extraída y de la regla de decisión empleada para cerrar la muestra o seguir muestreando.

Muestreos no Aleatorios

Este tipo de muestreo no permite, rigurosamente hablando, aplicar técnicas de inferencia estadística, ya que la formulación de estas técnicas se realiza bajo la hipótesis de la aleatoriedad de las muestras.

Muestreo Dirigido o Adaptado: Se seleccionan para formar parte de la muestra elementos, que según la opinión de los encuestadores, sean representativos. Se suele emplear en las primeras fases del estudio para construir una muestra piloto.

Muestreo por cuotas: Cada encuestador debe entrevistar a un cierto número de personas de unas características definidas. Por ejemplo: 15 hombres solteros con edades comprendidas entre 25 y 30 años, 22 mujeres casadas de edades comprendidas entre 30 y 50 años, 20 personas con hijos en edad escolar, etc..

Muestro deliberado: Se selecciona la muestra en un sector de la Población por comodidad de acceso. Por ejemplo cuando se dispone fácilmente de una lista de personas, como la guía de teléfono, las matrículas de los automóviles, etc..

1.2.3. Presentación de los datos: tablas y representaciones gráficas

Una primera manera en que pueden presentarse los datos es mediante una relación exhaustiva de todas las ocurrencias de la variable. Esto es lo que se conoce como una *tabla de tipo I*. Por ejemplo, si estamos estudiando el número de hermanos que tienen los alumnos de un colegio, se nos podrían presentar los siguientes datos:

$$1, 2, 1, 0, 0, 1, 3, 1, 0, 2, 1, 2, 0$$

Este manera de presentar los datos solo es factible cuando se tiene un número muy pequeño de observaciones. Si el número de observaciones es grande, lo que se hace es agrupar los datos, indicando a continuación el número de ocurrencias de cada uno. Esto es lo que se llama una *tabla de tipo II*. En el ejemplo anterior, la tabla tipo II correspondiente sería:

x_i	n_i
0	4
1	5
2	3
3	1

Donde n_i representa el número de veces que se presenta cada una de las observaciones. Para la variable *estudios* de los datos de la página 9 la tabla de tipo II es la siguiente:

x_i	n_i
bachiller	20
diplomado	10
fp	5
primario	13
superior	2

Podemos considerar que una tabla tipo I es una tabla tipo II en la que todos los n_i valen 1. En el caso de que sean muchos los posibles valores que pueda tomar la variable, agrupamos los datos en intervalos. Obtendremos entonces una *tabla de tipo III*. Por ejemplo, al estudiar la talla de los alumnos del instituto de la tabla 1.1 podemos agrupar a los que tengan una talla parecida. En la siguiente tabla se han tomado 8 clases. Cada una de ellas tiene una amplitud de 10 cm. Es decir, que hemos agrupado los datos de la variable talla en intervalos de idéntico tamaño.

Intervalo	marca de clase	frecuencia absoluta
[1.3, 1.4]	1.35	3
(1.4, 1.5]	1.45	3
(1.5, 1.6]	1.55	7
(1.6, 1.7)	1.65	11
(1.7, 1.8]	1.75	11
(1.8, 1.9]	1.85	5
(1.9, 2.0]	1.95	9
(2.0, 2.1]	2.05	1

A cada uno de los intervalos se le denomina *intervalo de clase*, y al punto medio de cada uno lo llamaremos *marca de clase*. La longitud de los intervalos de clase no tiene que ser siempre la misma, aunque es preferible que así sea. Para realizar determinados cálculos (por ejemplo la media), nos será útil pasar de una tabla tipo III a una tabla tipo II, considerando que todas las ocurrencias corresponden a la marca de clase.

Daremos a continuación algunas definiciones de interés. A cada una de las n_i se le llama *frecuencia absoluta* de la observación x_i. Si tenemos en total n observaciones y se presentan k casos distintos $x_1, x_2 \dots x_k$, entonces se cumple que:

$$\sum_{i=1}^{k} n_i = n$$

donde n_i es la frecuencia absoluta de cada dato de diferente valor contenido en la muestra.

Es decir que la suma de las frecuencias absolutas de todas las observaciones es, como es natural, el número total de observaciones realizadas (número de elementos de la muestra).

Puede comprobarse que la suma de las frecuencias de la tabla anterior es 50, que era el número de alumnos entrevistados.

Llamaremos *frecuencia relativa* de la observación x_i a

$$f_i = \frac{n_i}{n} \qquad \forall i = 1, 2, \ldots, k.$$

La frecuencia relativa es por tanto un número comprendido entre 0 y 1. Se cumple que

$$0 \leq f_i \leq 1 \quad \forall i = 1, 2, \ldots, k.$$

Del mismo modo pueden definirse N_j, *frecuencia absoluta acumulada* correspondiente a la observación x_j, como la suma de las frecuencias absolutas correspondientes a observaciones menores o iguales a la observación j. Es decir :

$$N_j = \sum_{i \leq j} n_i \qquad j = 1, 2, \ldots, k$$

Y la *frecuencia relativa acumulada* como

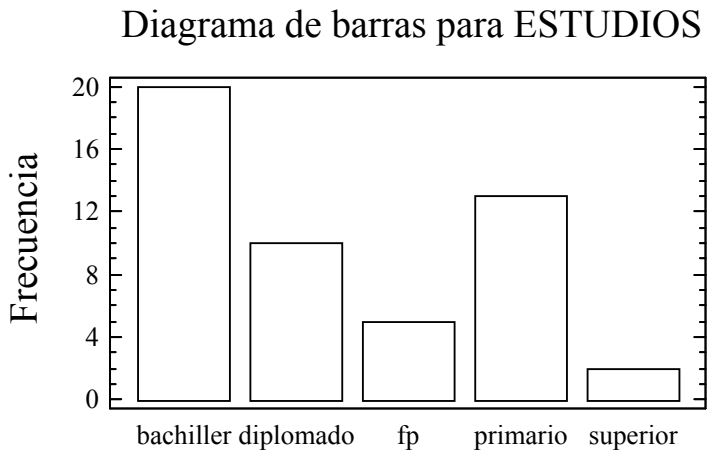
$$F_j = \sum_{i \leq j} f_i = \frac{N_j}{n} \qquad j = 1, 2, \ldots, k$$

En la siguiente tabla se dan las distintas frecuencias correspondientes a la tabla de tipo II para las tallas de los alumnos del instituto:

Intervalo	Marca de clase	Frecuencia absoluta (n_i)	Frecuencia absoluta acumulada (N_i)	Frecuencia relativa (f_i)	Frecuencia relativa acumulada (F_i)
[1.3, 1.4)	1.35	3	3	0.06	0.06
[1.4, 1.5)	1.45	3	6	0.06	0.12
[1.5, 1.6)	1.55	7	13	0.14	0.26
[1.6, 1.7)	1.65	11	24	0.22	0.48
[1.7, 1.8)	1.75	11	35	0.22	0.70
[1.8, 1.9)	1.85	5	40	0.10	0.80
[1.9, 2.0)	1.95	9	49	0.18	0.98
[2.0, 2.1]	2.05	1	50	0.02	1.00

1.2.4. Representaciones gráficas

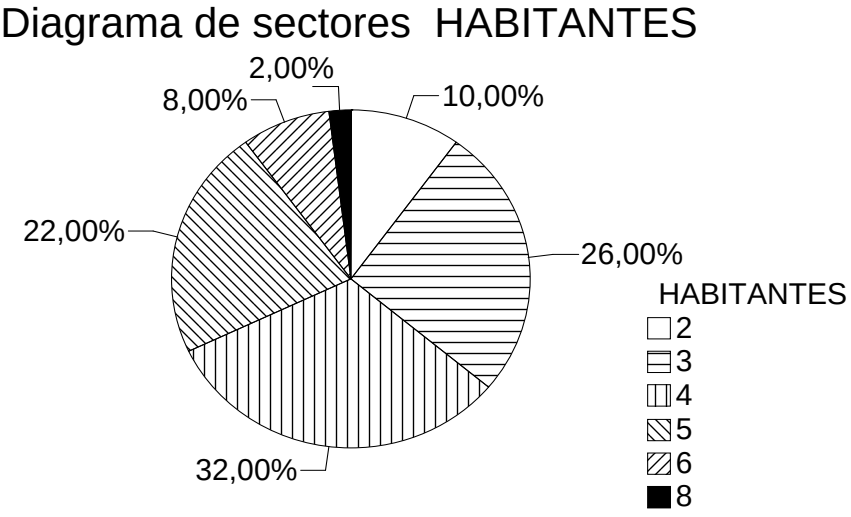
Diagrama de barras: Se construye un gráfico poniendo en el eje horizontal los valores observados y elevando sobre cada valor una barra de altura proporcional a su frecuencia. El diagrama de barras correspondiente a la tabla tipo II para la variable estudios es:



Representar la frecuencia absoluta o relativa como altura de la barra no influye en su forma, ya que sólo se realiza un cambio de escala.

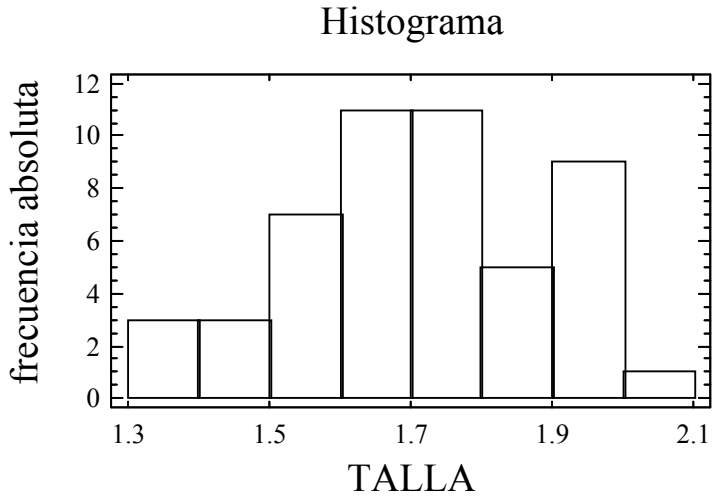
Este tipo de diagramas, y los de sectores que describiremos inmediatamente, son adecuados para variables cualitativas o cuantitativa discreta si el número de datos diferentes es pequeño, como ocurre en el caso de la variable habitantes.

Diagrama de sectores: Se dibuja un círculo y se divide en sectores circulares, de modo que cada uno represente la frecuencia de aparición en la muestra de un valor observado. Cada sector debe tener un área proporcional a su frecuencia, que suele venir indicada en la tabla en tanto por ciento. Si optamos por indicar la frecuencia relativa, el gráfico presenta el mismo aspecto, pero la frecuencia relativa viene indicada en tanto por uno.



Histograma: Se utiliza en el caso de las tablas de tipo III. Se construye el gráfico representando en el eje horizontal los intervalos de clase y elevando sobre cada uno de ellos un rectángulo de área proporcional a su frecuencia. El histograma correspondiente a la tabla tipo III mostrada anteriormente sería:

Nota: Si los intervalos no tienen la misma amplitud, habrá que representar las alturas del histograma $h_i = \frac{n_i}{a_i}$ siendo a_i la amplitud del intervalo correspondiente.



Polígono de frecuencias: Se construye una curva uniendo los puntos medios de los lados superiores de cada rectángulo del histograma. El polígono de frecuencias correspondiente al anterior histograma de la variable Talla es:

Diagrama de tallo y hojas: A continuación mostramos un diagrama de tallo y hojas para la variable Talla.

3	1.3 | 468
6	1.4 | 467
13	1.5 | 1223489
23	1.6 | 3455678899
(9)	1.7 | 024445688
18	1.8 | 000349
12	1.9 | 0034566889
2	2.0 | 02

El recorrido de la variable se ha dividido en 8 partes (los tallos), que vienen representados por los valores 1.3, 1.4, 1.5, etc. Los valores que le siguen, tras la línea vertical, son las hojas que corresponden a cada tallo. Así en el primer tallo tenemos las hojas 4, 6, 8. Esta rama corresponde a los datos más pequeños de la variable talla 1.34, 1.36, 1.38. La frecuencia acumulada de cada rama esta especificada a su izquierda. Así la frecuencia de la primera rama es 3, la de la segunda también es 3, pero la acumulada es 6. En este caso la acumulación de las frecuencias se hace por ambos

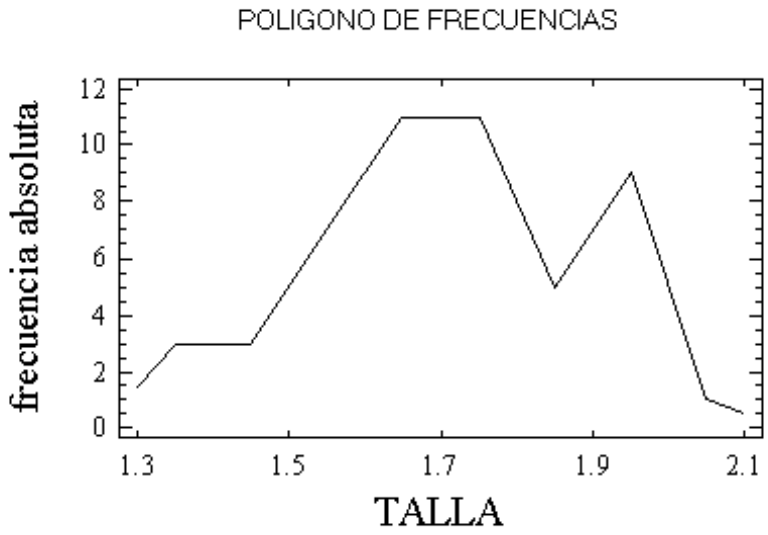
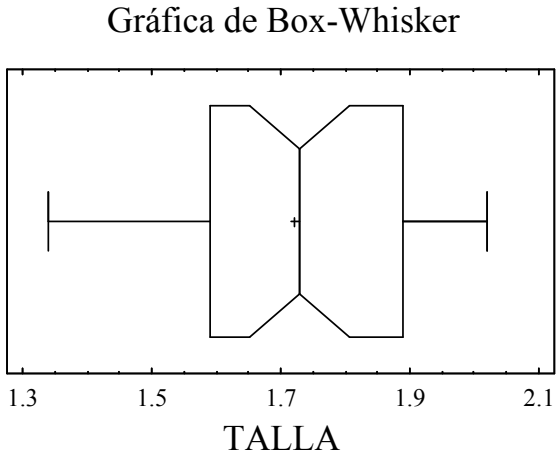


Figura 1.1:

lados de la tabla hasta llegar al tallo que contiene a la mediana. Este tallo contiene 9 elementos como está indicado entre paréntesis.

Esta representación tiene la ventaja de que superpone una tabla de frecuencias y una representación gráfica dada por la forma que toman los números, y que es similar al histograma de frecuencias. Además no hay pérdida de información, ya que se pueden reconstruir todos los datos de la variable primitiva contenida en la muestra a partir de esta representación.

Gráfica de caja y bigotes (Box and Whisker): En está gráfica los datos se dividen en cuatro intervalos de igual frecuencia. La parte ancha, llamada *Caja*, contiene el 50 % central de los datos de la variable. Comienza en el primer cuartil y termina en el tercer cuartil. La muesca de la caja marca la mediana (la definición de mediana y de cuartil se verá más adelante en el apartado de medidas de posición). En el gráfico de Box-Whisker correspondiente a la variable Talla, que aparece a continuación, se ha marcado además un punto, con un signo +, que corresponde a la media aritmética de los valores muestrales.



Las dos líneas horizontales se llaman *Bigotes* y se extienden a derecha e izquierda de la Caja. El bigote de la izquierda comienza por el dato más pequeño que dista del primer cuartil menos que 1.5 veces el rango intercuartílico (distancia entre el primer y tercer cuartil). En este caso corresponde al valor 1.34 El bigote de la derecha acaba en el mayor valor de la variable talla que diste del tercer cuartil menos que 1.5 veces el rango intercuartílico. Corresponde en este caso al valor mayor de la variable talla que es 2.02. A veces hay valores de la variable que sobresalen de los bigotes. Estos valores se clasifican como valores extraños (Outliers).

Las tablas y las gráficas pretenden ordenar y clarificar la información contenida en la muestra. En los casos tratados, excepto en el caso del diagrama de tallo y hojas, siempre se hace perdiendo parte de información. En el siguiente apartado se darán algunas definiciones que pretenden reducir la información contenida en la muestra de una forma aún más drástica: a sólo unos cuantos valores, los parámetros estadísticos de la muestra. Entre ellos destacamos las medidas de posición y las de dispersión, aunque también se podrían calcular parámetros de simetría, curtosis, concentración, desigualdad, etc.

1.2.5. Medidas de posición

Suponemos los datos ordenados de menor a mayor. Las medidas de posición caracterizan ciertos datos por la posición que ocupan en esta serie. Entre las medidas de posición tenemos las siguientes:

Mediana: Definimos la mediana como aquel valor que hace que el 50 % de las observaciones sean menores o iguales a él y otro 50 % mayor o igual que él. Si el número total de observaciones es n, y ordenamos los datos de menor a mayor, la mediana será la que ocupe el lugar $\frac{n+1}{2}$ si n es impar, o estará entre los valores $\frac{n}{2}$ y $\frac{n}{2}+1$ si n es par. En este caso la mediana se obtiene como la semisuma de estos dos valores centrales.

Si partimos de una tabla de tipo III, y no conocemos los valores primitivos de la variable, calculamos en primer lugar en qué intervalo se encuentra la mediana. Dicho intervalo, al que denominaremos intervalo mediano y denotaremos por $(L_i, L_{i+1}]$, será aquel en el que la frecuencia absoluta acumulada sea igual o supere a $\frac{n}{2}$.

Entonces se aplicará la fórmula:

$$Me = L_i + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i$$

siendo a_i la amplitud del intervalo mediano. Esta fórmula puede obtenerse aplicando el procedimiento de interpolación en la frecuencia usando como datos los extremos del intervalo mediano y su frecuencia acumulada.

Medidas de posición no central

La mediana se conoce como una medida de posición central, ya que divide las observaciones en dos partes de igual frecuencia. Definimos ahora otras medidas de posición que dividen la muestra en partes de distinta frecuencia. Reciben el nombre genérico de **cuantiles**. Destacaremos los cuartiles, deciles y percentiles.

El *cuartil* j, Q_j $(j = 1, 2, 3)$ es aquel valor que hace que las j cuartas partes de las observaciones sean menores o iguales a él y el resto mayores o iguales. El segundo cuartil coincide con la mediana.

El *decil* j, D_j $(j = 1, 2 \ldots 9)$ es aquel valor que hace que las j décimas partes de las observaciones sean menores o iguales a él y el resto mayores o iguales.

El *percentil* j, P_j $(j = 1, 2 \ldots 99)$ es aquel valor que hace que las j centésimas partes de las observaciones sean menores o iguales a él y el resto mayores o iguales.

Medidas de posición central

Las medidas de posición central pretenden ser representantes o ejemplos ilustrativos del tamaño de los datos contenidos en la muestra. La mediana es la única medida de posición central propiamente dicha. No obstante la media y la moda, toman frecuentemente valores parecidos a la mediana y se suelen conocer también como medidas de posición central.

Media aritmética La media se define como el cociente entre la suma de todos los valores y el número total de elementos de la muestra.

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

No obstante, si los datos están repetidos, hay n elementos en la muestra pero sólo hay k elementos diferentes cada uno de los cuales aparece con una frecuencia n_i, se puede obtener también la media por medio de las expresiones siguientes:

$$\overline{X} = \sum_{i=1}^{k} \frac{n_i x_i}{n}, \ \overline{X} = \sum_{i=1}^{k} \frac{n_i}{n} x_i = \sum_{i=1}^{k} f_i x_i \tag{1.1}$$

Como hemos observado que los alumnos suelen tener dificultades en ver la igualdad de estas expresiones, ilustramos esta igualdad con el siguiente ejemplo.

Si la muestra está formada por los datos {6, 8, 8, 8, 9, 9, 9, 9} la tabla de frecuencias tipo II es:

x_i	n_i	f_i
6	1	$\frac{1}{8} = 0.125$
8	3	$\frac{3}{8} = 0.375$
9	4	$\frac{4}{8} = 0.5$

La media puede obtenerse de las formas siguientes:

$$\begin{aligned}\overline{X} &= \frac{6+8+8+8+9+9+9+9}{8} = \frac{1\times 6+3\times 8+4\times 9}{8} = \frac{1}{8}\times 6+\frac{3}{8}\times 8+\frac{4}{8}\times 9 = \\ &= 0.125\times 6+0.375\times 8+0.5\times 9 = 8.25\end{aligned}$$

Los cálculos realizados corresponden sucesivamente con las distintas expresiones dadas previamente para la media en 1.1.

Moda Es el valor que presenta una mayor frecuencia. Si es único, se dice que la distribución es unimodal, si no lo es, se dice que es multimodal.

En el caso de tablas tipo III, siendo $a_i = L_{i+1} - L_i,\ h_i = \dfrac{n_i}{a_i}$ es la altura que debe tener el histograma en el intervalo correspondiente. Llamaremos *intervalo modal*, $(L_i, L_{i+1}]$, al intervalo que presenta mayor altura en el histograma.

Aunque hay distintas expresiones para seleccionar un valor concreto para la moda dentro del intervalo modal optamos por la siguiente:

$$Mo = L_i + \frac{h_{i+1}}{h_{i-1}+h_{i+1}}a_i$$

1.2.6. Medidas de dispersión

Pretenden dar una idea sobre si los datos son muy parecidos entre sí o por el contrario están dispersos, es decir bastante separados unos de otros. Para aclarar este concepto consideremos las dos muestras siguientes que suponemos que son las calificaciones obtenidas por dos alumnos en las cuatro preguntas de un examen:

Notas del alumno A = {5, 4, 6, 5}
Notas del alumno B = {1, 9, 10, 0}

Si usáramos la media para obtener la calificación del examen ambos alumnos recibirían la misma calificación, un cinco, pero percibimos que los exámenes de estos alumnos tienen características bien diferentes. Intentamos describir esta diferencia con ciertos parámetros que llamamos medidas de dispersión. Entre los más usados destacamos los siguientes:

Recorrido o Rango Es la diferencia entre el valor máximo y el mínimo de la variable.

Recorrido alumno A = 6 - 4 = 2
Recorrido alumno B = 10 - 0 = 10

Rango intercuartílico Es la distancia entre el primer y tercer cuartil.

Por ejemplo, para obtener los cuartiles de las *Notas del alumno A* = {5, 4, 6, 5}, ordenamos éstas y obtenemos {4 , 5 , 5 , 6}. Dividiendo en cuatro partes la frecuencia obtenemos 4 clases con frecuencia 1. Como la separación de las clases no está en ningún elemento, tomando la media de los valores más cercanos obtenemos:

Primer cuartil = 4.5

Segundo cuartil = 5.5

Así que el rango intercuartílico es 5,5 − 4,5 = 1

A veces se define el Rango semintercuartílico, que es la mitad del Rango intercuartílico. En este caso su valor sería 0.5.

Desviación media Se define de la siguiente forma:

$$\text{des}(X) = \frac{1}{n}\sum_{i=1}^{k} n_i \left|x_i - \overline{x}\right|$$

Mide, por promedio, la distancias entre los datos y la media de la muestra.

La desviación media para el primer alumno es:

$$\text{des}(Notas\ de\ A) = \frac{1}{4}\left(2 \times |5-5| + 1 \times |4-5| + 1 \times |6-5|\right) = 0{,}5$$

$$\text{des}(Notas\ de\ B) = \frac{1}{4}\left(1 \times |1-5| + 1 \times |9-5| + 1 \times |10-5| + 1 \times |0-5|\right) = 4.5$$

Varianza y Cuasivarianza Aunque la desviación media nos da una definición que representa la dispersión de una forma muy intuitiva, a menudo se usa la varianza como medida de dispersión debido, entre otros motivos, a que el valor absoluto presenta ciertos inconvenientes en el cálculo, por no ser una función derivable. La varianza se define como

$$\text{var}(X) = S^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (x_i - \overline{x})^2$$

Puede demostrarse que esta expresión es equivalente a $\frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \quad \overline{x}^2$, que presenta algunas ventajas de cálculo.

La varianza de las notas del alumno A es

$\text{Var}(Notas\ de\ A) = \frac{1}{4}\left(2 \times (5-5)^2 + 1 \times (4-5)^2 + 1 \times (6-5)^2\right) = 0.5$

La cuasivarianza se define como:

$$\text{cuasivar}(X) = s^2 = \frac{1}{n-1}\sum_{i=1}^{k} n_i (x_i - \overline{x})^2$$

La cuasivarianza de las notas del alumno A es

$$s^2 = \frac{1}{4-1}\left(2 \times (5-5)^2 + 1 \times (4-5)^2 + 1 \times (6-5)^2\right) = 0.666\,67$$

Desviación típica y cuasidesviación Se define la desviación típica como la raíz cuadrada positiva de la varianza, $S = +\sqrt{\text{var}(X)}$. Tiene la ventaja sobre la varianza de venir expresada en la misma unidad que los datos. La desviación típica de la variable notas del alumno A es

$$S = \sqrt{\text{var}(Notas\ de\ A)} = \sqrt{0{,}5} = 0.707\,11$$

Definimos la cuasi desviación como la raíz cuadrada de la cuasivarianza. La cuasidesviación de la variable notas del alumno A es

$$s = \sqrt{\text{cuasivar}(Notas\ de\ A)} = \sqrt{0.666\,67} = 0.816\,5$$

Coeficiente de Variación de Pearson Se denomina Coeficiente de Variación de Pearson al cociente:

$$CV = \frac{S}{|\ \overline{x}|}$$

que es una medida relativa de variabilidad y que permite comparar la dispersión de dos conjuntos de datos de diferentes escalas o diferentes unidades de medida. Este parámetro es invariante frente al cambio de escala.