UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE PSICOLOGÍA

ESTADÍSTICA CÁT I

Prof. a cargo: Dra. María Silvia Galibert

EXPLICACIONES COMPLEMENTARIAS PARA LA UNIDAD 1

Por Lic. Alcira García Díaz

Marzo 2020

Capítulo 1. Los datos estadísticos

Definiciones básicas

La Estadística es una rama de la Matemática que, utilizada por distintas áreas de conocimiento, entre ellas las Ciencias Sociales, Humanas y de la Salud, se ocupa de dos aspectos centrales en el tratamiento de datos. Uno de ellos consiste en la organización, descripción, resumen de información cuali y cuantitativa generalmente correspondiente a un número reducido de datos observados, aspecto correspondiente a la que se denomina Estadística Descriptiva. El otro, vinculado a la realización de inferencias, generalizaciones, extrapolaciones de los datos observados a un conjunto generalmente mayor, no observado, de donde los primeros fueron extraídos, corresponde a lo que se denomina Estadística Inferencial. La investigación científica requiere de ambos aspectos de esta disciplina que provee de técnicas y procedimientos para el análisis de datos y para la toma de decisiones.

El presente texto se ocupará de introducir algunas nociones básicas de Estadística Descriptiva.

La aplicación de métodos estadísticos se realiza sobre un conjunto de individuos o entidades que tienen uno o varios atributos en común y que interesa estudiar. Así es que se define

Población: es el conjunto de elementos, individuos o unidades de análisis que presentan una o más características o atributos que son objeto de un estudio particular.

Las unidades de análisis o individuos que constituyen la población pueden ser seres humanos, escuelas, familias, animales o números que representen edades de personas encuestadas.

La población debe estar definida en tiempo y espacio cuando se realiza una investigación, ya que constituye el marco de referencia sobre el cual se van a efectuar interpretaciones y conclusiones y éstas no pueden exceder dicho marco.

Dado que las poblaciones suelen ser bastante numerosas y suele ser difícil trabajar con tanta información se suele considerar sólo una parte de ellas. Así es que se define

Muestra: es un subconjunto de elementos de una población.

Individuo: es un elemento o unidad de análisis de la población.

De los individuos que conforman la muestra o la población obtenemos información a procesar, es decir, estos individuos poseen características o a tributos a ser estudiados.

A continuación, se muestra un fragmento del Cuestionario sobre Impuntualidad de Hoyos et al (2019) con algunos datos sociodemográficos:

Número de Cuestionario: …….

P1. Edad: (expresada en años enteros cumplidos)

P2. Sexo: 1. Mujer 2. Varón

P3. Nivel Socioeconómico: 1. Bajo 2. Medio-Bajo 3. Medio 4. Medio-Alto 5. Alto

P4. Ocupación:
1. Sin ocupación
2. Obrero
3. Empleado
4. Docente
5. Trabajador Independiente
6. Otros

P5. Lugar de residencia:
1. CABA
2. Gran Buenos Aires
3. Otro

P6. Nacionalidad
1. Argentina
2. Boliviana
3. Paraguaya
4. Venezolana
5. Chilena
6. Peruana
7. Otro

A continuación, se exhibe un par de afirmaciones sobre hábitos u opiniones acerca de la puntualidad. Después de cada oración hay una escala que expresa distintos grados en que esa afirmación lo representa. Marque con una cruz X la opción que mejor lo representa.

P7. Antes de salir de casa doy tantas vueltas que termino llegando tarde.

1.Bastante parecido a mí ( ) 2. Algo parecido a mí ( ) 3. Algo diferente de mí ( ) 4. Bastante diferente de mí ( )

P8. Saber que estoy llegando tarde me genera un estado de ansiedad muy desagradable.

1.Bastante parecido a mí ( ) 2. Algo parecido a mí ( ) 3. Algo diferente de mí ( ) 4. Bastante diferente de mí ( )

P9. Tiempo en minutos que tardó en completar el cuestionario.

El cuestionario es un instrumento de medición que genera datos y está dirigido a una población definida, por ejemplo, estudiantes de la carrera de Psicología en la UBA que cursan la materia Estadística durante el primer cuatrimestre de 2020. Si entre

todos dichos alumnos se sortearan, por ejemplo, 40 para administrarles el cuestionario, estos 40 alumnos constituirían una muestra.

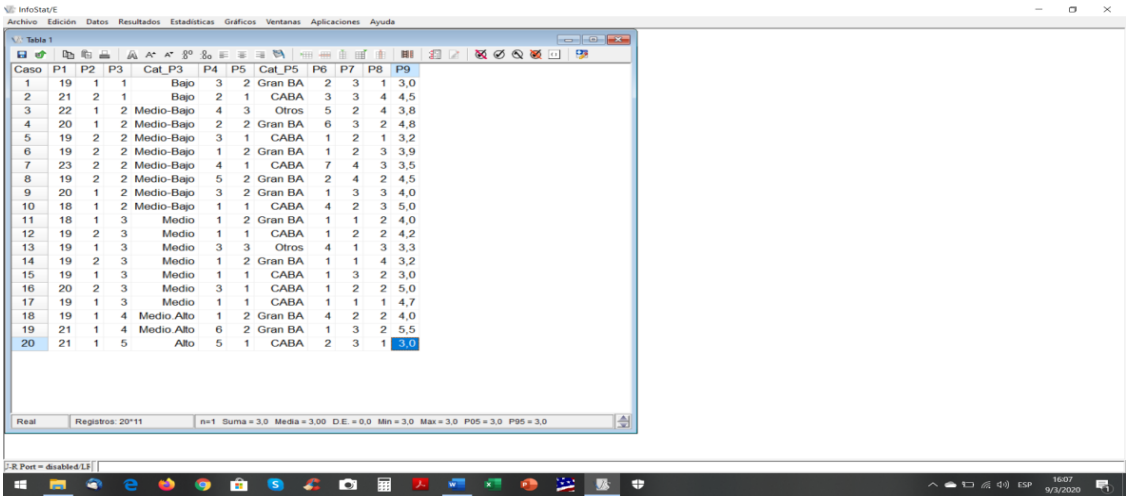
Además, siguiendo con este ejemplo, se advierte que se está solicitando información sobre determinadas características de interés a estudiar en dicha población tales como la edad, el sexo, el nivel socioeconómico, la ocupación, el lugar de residencia, la nacionalidad, un par de ítems referidos a la actitud del encuestado frente a la impuntualidad y el tiempo que tardó en contestar el cuestionario.

Este cuestionario tiene 9 items denominados P1, P2, ……, P9. Cada entrevistado que responde al mismo marca una sola de las respuestas indicadas en cada ítem. Una vez que han respondido todos los entrevistados se obtiene un bloque de información "en bruto" que es necesario ordenar para poder interpretar. Así es que con los datos organizados se forma la "matriz de datos" que se muestra a continuación.

Cuestionario	P1	P2	P3	P4	P5	P6	P7	P8	P9
1	19	1	3	1	1	1	3	2	3
2	20	1	2	3	2	1	3	3	4
3	19	1	3	1	1	1	1	1	4.7
4	21	2	1	2	1	3	3	4	4.5
5	18	1	2	1	1	4	2	3	5
…	…	…	…	…	…	…	…	…	…
20	19	1	4	1	2	4	2	2	4

Este arreglo de datos que constituyen la matriz de datos presenta información en las filas horizontales y en las columnas verticales. En la primera fila se escriben los nombres de los ítems y en las filas siguientes los números que corresponden a las respuestas dadas por los encuestados. La primera columna indica el número de cuestionario o del encuestado, por lo cual el encuestado que respondió el cuestionario 4, según los datos de esa fila tiene 21 años, es un varón que se considera de nivel socioeconómico bajo, es un obrero que vive en la ciudad de Buenos Aires, de nacionalidad paraguaya. Respecto de su actitud frente a la impuntualidad manifiesta no demorarse para salir de su casa y que no le genera ansiedad el saber que llega tarde. Respondió el cuestionario en cuatro minutos y medio.

Esta información puede introducirse en en EXCEL o en un programa estadístico, como InfoStat o Statistix, entre otros.



La matriz de datos es una disposición de números donde cada fila representa a un individuo que posee la información de interés, y cada columna es un aspecto del individuo que se ha seleccionado para estudiar y cada celda es la modalidad que tiene el individuo de la fila en el aspecto de la columna correspondiente.

Se advierte que cada columna de la matriz de datos es un ítem del cuestionario que indica una característica o atributo del individuo que interesa observar. Estas características se transforman en variables mediante el proceso de medición. Ejemplos de variables son la edad, el sexo, el nivel socioeconómico, la ocupación, el lugar de residencia, etc. Cada una de ellas se presenta en diferentes modalidades. Si se piensa en el nivel socioeconómico se ve que puede adoptar distintos valores tales como medio, bajo, medio-alto, etc.. Si se hace referencia a la nacionalidad se espera que en las respuestas aparezcan distintos gentilicios tales como argentino, boliviano, brasilero, español, etc. Si en cambio se atiende a las respuestas dadas al tiempo empleado en minutos para responder al cuestionario se espera dar con números enteros o decimales que indiquen el lapso transcurrido para responder. Así es que se define

Variable es una característica de los individuos o unidades de análisis de una población que puede presentar distintos valores o categorías.

Las **categorías** de una variable son los valores que ésta puede asumir.

Para los ejemplos mencionados más arriba se tiene que la variable nivel socioeconómico puede presentarse según las categorías o valores: bajo, medio-bajo, medio, medio-alto o alto. Para la variable nacionalidad se tiene que sus categorías son tantas como los gentilicios posibles, argentino, boliviano, brasilero, español, etc., correspondientes a distintos países. Mientras que para la variable tiempo empleado en minutos para responder al cuestionario sus valores o categorías son números enteros o decimales mayores o iguales que cero y menores que un cierto tiempo máximo otorgado para responder.

Las categorías de una variable deben se **exhaustivas y mutuamente excluyentes**. Esto quiere decir que los valores establecidos de una variable deben ser

tales que todo individuo pueda incluirse en una de sus categorías y sólo una de ellas (principio de exclusión) y conjuntamente no quede ningún valor por fuera de las anteriormente establecidas (exhaustividad). A modo de ejemplo diremos que si queremos registrar el nivel socioeconómico de un individuo a éste tiene que corresponderle según su situación un determinado nivel, supónganse "medio-bajo" y si se le asigna este nivel no puede asignársele otro, por ejemplo "medio", a esto se refiere la mutua exclusión y por otra parte, a todo encuestado le tiene que corresponder alguno de los posibles niveles socioeconómicos sin que quede por fuera de ellos, este último hecho hace referencia a la exhaustividad.

Existen variables más complejas que las mencionadas anteriormente. Si se piensan en conceptos como el de memoria, inteligencia o ansiedad se advierte que se expresan a través de diversas manifestaciones. La cantidad de palabras recordadas en cierto tiempo puede ser un aspecto que da cuenta de la memoria, pero no lo agota, un correcto ordenamiento de imágenes vistas previamente podría ser otra manifestación de la memoria. Estos conceptos son tratados en el campo de la Psicología como estructuras o procesos internos que explican las manifestaciones observables y se denominan constructos lógicos que suelen ponerse a prueba. Es necesario objetivar estos conceptos que constituyen variables por lo que se habla de operacionalizar la variable.

Operacionalizar una variable significa detallar los procedimientos necesarios para obtener los valores de dicha variable de modo que esos resultados puedan ser reproducidos por otros. Así se podría definir operacionalmente "memoria" a través del sencillo procedimiento de contar la cantidad de palabras recordadas en una lista o dar cuenta de los valores del constructo "inteligencia" registrando el tiempo empleado en resolver una serie de problemas. Por supuesto que estos procedimientos reflejan muchas veces aspectos limitados de dichos constructos. Mediante la definición operacional de una variable distintas personas deberían obtener el mismo valor de dicha variable para un mismo individuo.

Clasificación de las variables

Cuando los valores asignados a una variable son símbolos o incluso números, pero sin las propiedades específicas numéricas, es decir números usados como etiquetas, se clasifica a la **variable** como **cualitativa**. Este tipo de variables se refieren a atributos, condiciones o cualidades que poseen los individuos. Ejemplo de este tipo de variables son el sexo, el lugar de residencia, la ocupación, cuyos valores no admiten orden entre ellos, otros ejemplos como el nivel socioeconómico o el nivel de instrucción constituyen variables cualitativas que si admiten un orden entre sus valores.

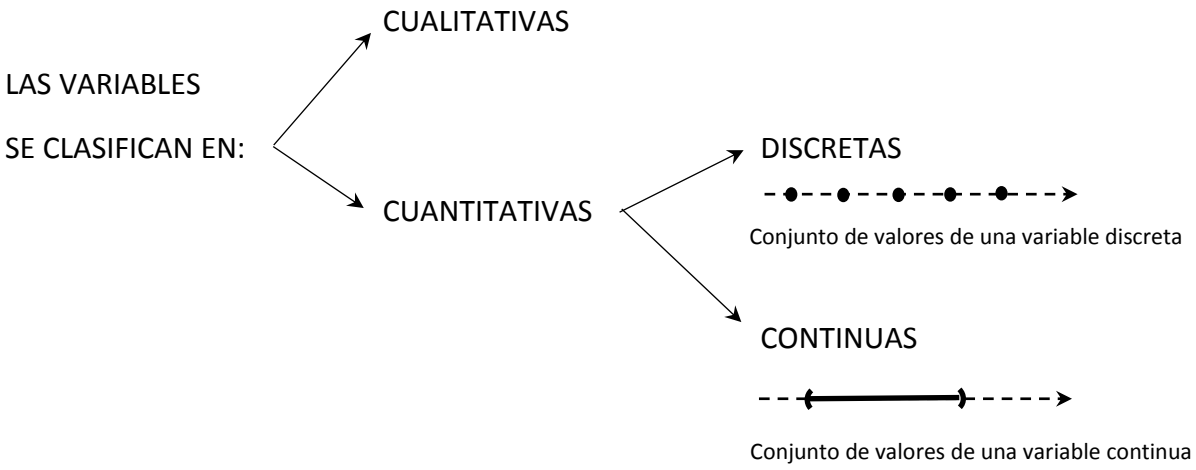
La otra clase de variables conocida como **variables cuantitativas** son aquellas cuyos valores son números con características de tales. A su vez este tipo de variables pueden subdividirse en discretas y continuas.

Los valores que puede tomar una variable cuantitativa discreta pueden enumerarse. Típicamente en las aplicaciones provienen de conteos, aunque no es el único caso. Por ejemplo, la cantidad de palabras recordadas en cierto lapso o la cantidad de miembros de una familia son ejemplos de este tipo de variables cuantitativas discretas.

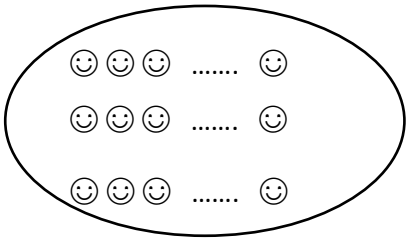
Una variable cuantitativa continua puede tomar valores a lo largo de todo un intervalo de números reales; por lo que no se podrían enumerar. Los valores de una variable cuantitativa continua constituyen un intervalo real. El tiempo empleado en responder a un estímulo (luz, sonido u otra sensación) o en resolver un problema aritmético, la distancia recorrida por una jabalina lanzada por un atleta, son ejemplos de este último tipo de variables.

Es importante distinguir entre variables cuantitativas discretas y continuas dado que los procedimientos específicos para su estudio difieren si se trata de un tipo u otro.

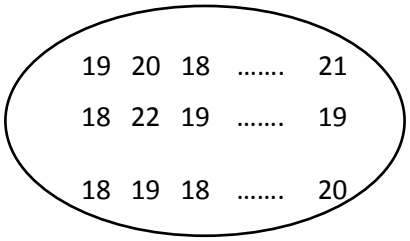
Resumiendo:



La noción de variable introduce nuevas consideraciones acerca de lo que se denomina población y muestra. Se dijo que una población está constituida por individuos (en sentido amplio) que presentan atributos de interés por ser estudiados. Una variable asocia un valor (símbolo o número) a cada individuo. El agrupamiento de estos valores constituye lo que se denomina población de observaciones. A continuación se presenta una representación esquemática de la población de individuos y de la población de observaciones correspondiente a la variable edad, por ejemplo, medida sobre esa población de individuos.



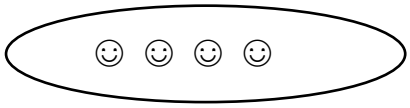
Población de individuos



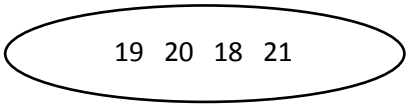
Población de observaciones

Si estamos interesados en otras variables además de la edad, como por ejemplo el lugar de residencia o la ocupación entonces esa misma población de individuos va a generar otras dos poblaciones de observaciones correspondientes a los lugares donde viven cada uno de los encuestados y la ocupación de cada uno de ellos. La misma población de individuos tendrá asociadas tantas poblaciones de observaciones como variables se consideren.

De modo análogo se puede definir muestra de individuos como una parte de la población de individuos y muestra de observaciones como una parte de la población de observaciones de la variable en cuestión, sólo aquellos valores de la variable que exhiben los individuos de la muestra. Siguiendo con la variable edad se ilustran estos últimos conceptos con el esquema siguiente.



Muestra de individuos



Muestra de observaciones

Medición

Las posibilidades de medición en el ámbito de las Ciencias Sociales fueron discutidas argumentándose que los atributos psicológicos no son de la misma naturaleza que los objetos físicos, (Campbell, 1932). Las diferencias fueron sorteadas extendiendo el concepto de medición de modo que se adecue a tales atributos (Stevens, 1951).

Por otra parte, existen reglas establecidas por la American Psychological Asociation que atienden, entre otras, a razones éticas. Al efectuar una medición puede tenerse precaución y evitar extrapolaciones inadecuadas delimitando el alcance de la medición.

Actualmente no se cuestionan ni la posibilidad ni la utilización de la medición en Psicología, aunque se revisen sus instrumentos y se discutan algunas medidas en particular.

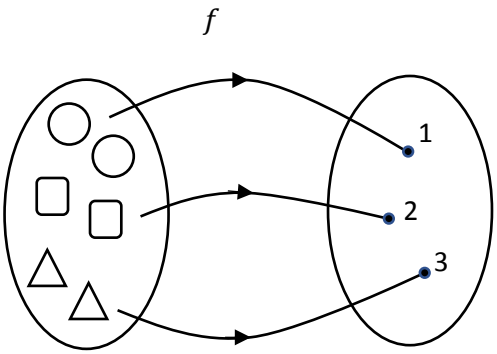
La definición de medición menos restrictiva respecto de las ciencias duras que propuso Stevens puede expresarse en estos términos:

Medir es asignar números a los objetos conforme a ciertas reglas de modo que las propiedades de éstos se vean reflejadas en los números que las representan.

Ejemplo

Propiedad empírica:

Dos objetos son equivalentes si se pueden superponer de modo que coincidan sus bordes



Significado de los números: son rótulos

Regla

A objetos equivalentes se le asignan símbolos iguales y a objetos no equivalentes se le asignan símbolos diferentes.

$$a \sim b \;\Leftrightarrow\; f(a) = f(b)$$

No todas las variables admiten ser medidas con la misma intensidad. No medimos la opinión acerca de un hecho de la misma forma que registramos el tiempo transcurrido para completar una tarea. Las diferencias se explican por el significado que se le adjudica a los símbolos que se asignan a las categorías de las variables, o también por las reglas que vinculan a dichos símbolos con lo que se observa.

Se dice que el **nivel de medición** de una variable está determinado por el significado que tengan los símbolos (numéricos o no) que se asignan a las categorías.

Se distinguen los siguientes niveles de medición:

El nivel nominal

Las variables que corresponden a este nivel de medición o escala cumplen las siguientes propiedades:

- Medir tiene el significado de clasificar.
- En una buena clasificación las clases deben ser exhaustivas (contemplar todas las posibilidades) y excluyentes (no deben compartir elementos para evitar la ambigüedad).
- Los números asignados como valores de la variable deben entenderse sólo como códigos, como tales, pueden ser sustituidos por cualesquiera otros que respeten la regla de asignación.
- La dificultad radica en definir clases verdaderamente exclusivas y exhaustivas.

Ejemplos

Tipos de Personalidad (debida a C. G. Jung) introvertidos que a su vez se subdividen en (pensamiento, sentimentales, sensaciones, intuitivos) y extrovertidos con la misma subdivisión.

País de origen

Área de la Psicología a la que piensa dedicarse en el futuro: clínica, educacional, organizacional, comunitaria, forense, etc.

El nivel ordinal

Las variables que corresponden a este nivel de medición o escala cumplen las siguientes propiedades:

- Se agrega una relación de orden entre los objetos.
- Tiene sentido establecer un orden entre los objetos, aunque no se precise cuantitativamente en qué grado un objeto posee mayor cantidad de un atributo que otro.
- Los números sólo indican rango o jerarquía.
- Pueden ser sustituidos por otros que conserven el mismo orden.

Ejemplos

Nivel Socioeconómico: 1. Bajo; 2. Medio-Bajo; 3. Medio; 4. Medio-Alto; 5. Alto

Grado en que la siguiente afirmación lo representa: “Saber que estoy llegando tarde me genera un estado de ansiedad muy desagradable”.

1.Bastante parecido a mí ( ) 2. Algo parecido a mí ( ) 3. Algo diferente de mí ( ) 4. Bastante diferente de mí ( )

El nivel intervalar

Las variables que corresponden a este nivel de medición o escala cumplen las siguientes propiedades:

- Está definida la distancia entre dos valores de la escala, es decir, existe una unidad de medición, por lo que puede precisarse “cuántas unidades es mayor un valor que otro”.
- Tiene sentido compararlas diferencias entre los números.
- El cero de la escala es arbitrario, convencional. (0° centígrados, año cero en el calendario cristiano occidental).
- Los números pueden sustituirse por otros de modo que conserven la relación entre las distancias. Por esto este nivel queda caracterizado por las transformaciones afines (y = a+bx)

Ejemplos

La temperatura: 25°C, 100°C, 40,3°C, etc.

Año del calendario cristiano occidental: 100, 1778, 2020, etc.

En Psicología las escalas construidas a partir de puntajes de test son tratadas muy frecuentemente como escalas de nivel Intervalar, por ejemplo el puntaje de CI (Coeficiente intelectual).

El nivel de razón o cociente o proporción

A las variables que corresponden a este nivel de medición se les agrega la existencia de un cero absoluto, con significado no convencional, que puede entenderse como ausencia de la característica que se desea medir. Por tanto tiene sentido decir que un valores el doble de otro, es decir tiene sentido interpretar las razones. Por ejemplo las magnitudes físicas usuales como peso, longitud y tiempo transcurrido.

Como se observa la complejidad de cada nivel de medición va en aumento a medida que se pasa del nivel de medición nominal hasta llegar al de razón. Cada una de las escalas mencionadas conserva las propiedades de las anteriores y agrega otras.

Los niveles de medición de Stevens fueron formalizados posteriormente a través de un desarrollo axiomático por Krantz, Suppes, Tversky y Luce en su obra Fundamentos de la medición (1971, 1989 y 1990). Allí abordan problemas de la medición tales como el problema de la representación; el problema de la unicidad; el problema de la significación.

Resumen de los niveles de medición

<table>
<tr><th>Nivel de medición</th><th>Significado de los símbolos numéricos</th><th>Requisito para cambiar los números</th><th>Ubicación del cero</th><th colspan="2">Ejemplos</th></tr>
<tr><td>Nominal</td><td>Designan, distinguen</td><td>Que no se repita el mismo para diferentes categorías</td><td>Sin significado</td><td colspan="2">Ocupación</td></tr>
<tr><td>Ordinal</td><td>Expresan orden</td><td>Que respeten el orden de las categorías</td><td>Sin significado</td><td colspan="2">Nivel de Escolarización alcanzado</td></tr>
<tr><td rowspan="2">Intervalar</td><td rowspan="2">Tienen sentido las distancias entre los valores de la variable</td><td rowspan="2">$y = b_0 + b_1 * x$</td><td rowspan="2">Arbitrario</td><td>Discreta</td><td>Año calendario</td></tr>
<tr><td>Continua</td><td>Puntaje de CI</td></tr>
<tr><td rowspan="2">Razón o Proporcional</td><td rowspan="2">Tienen sentido las razones entre los valores de la variable</td><td rowspan="2">$y = b_1 * x$</td><td rowspan="2">Absoluto (indican ausencia de lo que se mide)</td><td>Discreta</td><td>Número de personas que habitan el hogar</td></tr>
<tr><td>Continua</td><td>Tiempo empleado en responder a un cuestionario</td></tr>
</table>

Capítulo 2: Distribuciones de frecuencias y gráficos

A partir de las definiciones estadísticas básicas del capítulo anterior, se puede plantear que se ha efectuado una recolección de datos, a la manera de la matriz de datos presentada en el capítulo anterior; es decir, se dispone de un conjunto de valores de varias variables que deben ser organizados para extraer más fácilmente información acerca de lo recolectado. Con este fin se construye la distribución de frecuencias y a partir de ella se pueden realizar representaciones gráficas.

Se utilizarán una letra mayúscula para identificar una variable por ejemplo X y dicha variable puede adoptar distintos valores x_1, x_2, x_3,…, x_n ; pero cada uno de esos valores puede aparecer más de una vez en los n elementos que componen la muestra. Por tal razón conviene dar las siguientes definiciones.

Frecuencia absoluta de un valor x_i es el número de veces que ese valor se repite en la muestra y se simboliza con f_i. Vale siempre que $\sum_{i=1}^{k} f_i = n$ donde n es el tamaño de la muestra y k es el número de valores diferentes de la variable (véanse ejemplos más adelante).

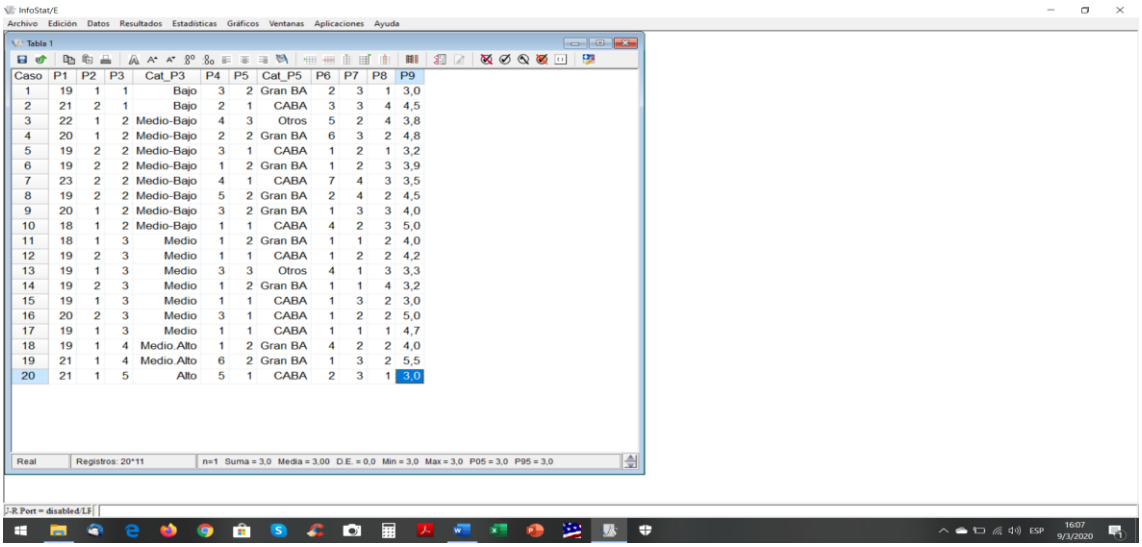
Frecuencia relativa de un valor x_i es el cociente entre la frecuencia absoluta y el tamaño de la muestra y se simboliza con f'_i. Es decir $f'_i = \frac{f_i}{n}$ representa la proporción de casos que asumen ese valor x_i de la variable. Vale siempre que $\sum_{i=1}^{k} f'_i = 1$ si se suman todas las proporciones se obtiene la unidad.

Frecuencia porcentual de un valor X_i es la frecuencia relativa multiplicada por 100 y se simbolizará $f_{\%i}$. Vale decir que $f_{\%i}$ = f'_i*100= $\frac{f_i}{n} * 100$ representa el porcentaje de casos que asumen ese valor x_i de la variable. Vale siempre que $\sum_{i=1}^{k} f_{\%i} = 100$ si se suman todos los porcentajes se obtiene 100%.

Así, vale que

Una **distribución de frecuencias** es un conjunto de valores de una variable con sus correspondientes frecuencias absolutas, relativas o porcentuales.

Esta distribución de frecuencias se define para todo tipo de variables ya sean cuali o cuantitativas. Para ilustrar cómo se construyen las tablas de distribución de frecuencias consideremos la matriz de datos del capítulo anterior.



Caso	P1	P2	P3	Cat_P3	P4	P5	Cat_P5	P6	P7	P8	P9
1	19	1	1	Bajo	3	2	Gran BA	2	3	1	3,0
2	21	2	1	Bajo	2	1	CABA	3	3	4	4,5
3	22	1	2	Medio-Bajo	4	3	Otros	5	2	4	3,8
4	20	1	2	Medio-Bajo	2	2	Gran BA	6	3	2	4,8
5	19	2	2	Medio-Bajo	3	1	CABA	1	2	1	3,2
6	19	2	2	Medio-Bajo	1	2	Gran BA	1	2	3	3,9
7	23	2	2	Medio-Bajo	4	1	CABA	7	4	3	3,5
8	19	2	2	Medio-Bajo	5	2	Gran BA	2	4	2	4,5
9	20	1	2	Medio-Bajo	3	2	Gran BA	1	3	3	4,0
10	18	1	2	Medio-Bajo	1	1	CABA	4	2	3	5,0
11	18	1	3	Medio	1	2	Gran BA	1	1	2	4,0
12	19	2	3	Medio	1	1	CABA	1	2	2	4,2
13	19	1	3	Medio	3	3	Otros	4	1	3	3,3
14	19	2	3	Medio	1	2	Gran BA	1	1	4	3,2
15	19	1	3	Medio	1	1	CABA	1	3	2	3,0
16	20	2	3	Medio	3	1	CABA	1	2	2	5,0
17	19	1	3	Medio	1	1	CABA	1	1	1	4,7
18	19	1	4	Medio Alto	1	2	Gran BA	4	2	2	4,0
19	21	1	4	Medio Alto	6	2	Gran BA	1	3	2	5,5
20	21	1	5	Alto	5	1	CABA	2	3	1	3,0

Recordemos la codificación:

P1. Edad: (expresada en años enteros cumplidos)

P2. Sexo:
1. Mujer
2. Varón

P3. Nivel Socioeconómico:
1. Bajo
2. Medio-Bajo
3. Medio
4. Medio-Alto
5. Alto

P4. Ocupación:
1. Sin ocupación
2. Obrero
3. Empleado
4. Docente
5. Trabajador Independiente
6. Otros

P5. Lugar de residencia:
1. CABA
2. Gran Buenos Aires

P6. Nacionalidad
1. Argentina
2. Boliviana
3. Paraguaya
4. Venezolana
5. Chilena
6. Peruano
7. Otro

A continuación, se exhibe un par de afirmaciones sobre hábitos u opiniones acerca de la puntualidad. Después de cada oración hay una escala que expresa distintos grados en que esa afirmación lo representa. Marque con una cruz X la opción que mejor lo representa.

P7. Antes de salir de casa doy tantas vueltas que termino llegando tarde.
1.Bastante parecido a mí ( ) 2. Algo parecido a mí ( ) 3. Algo diferente de mí ( ) 4. Bastante diferente de mí ( )
P8. Saber que estoy llegando tarde me genera un estado de ansiedad muy desagradable.
1.Bastante parecido a mí ( ) 2. Algo parecido a mí ( ) 3. Algo diferente de mí ( ) 4. Bastante diferente de mí ( )
P9. Tiempo en minutos que tardó en completar el cuestionario.

Si seleccionamos los valores de la columna P5 (Lugar de residencia), tendríamos la siguiente tabla de distribución de frecuencias para la variable Lugar de residencia. En la matriz de datos contamos nueve unos; es decir, nueve personas que viven en la ciudad de Buenos Aires, nueve números dos, que son las personas que viven en el Gran Buenos Aires y dos tres que representan a quienes viven en otras ciudades del interior de la provincia de Buenos Aires u otras Provincias de nuestro país o eventualmente otros países. Estos números 9, 9 y 2 constituyen las frecuencias

absolutas para esta variable. A partir de ellas podemos obtener las frecuencias relativas y las porcentuales para cada valor de la variable, como se muestra a continuación.

Distribución de frecuencias para la variable Lugar de residencia (P5)

$x_i\ (Lugar\ de\ residencia)$	f_i	f'_i	$f_{\%i}$
1 CABA	9	0,45	45
2 Gran Buenos Aires	9	0,45	45
3 Otro	2	0,10	10
Totales	20	1	100

Análogamente si fijamos la atención en los valores de la columna P3 (Nivel Socioeconómico) tendríamos la siguiente tabla

Distribución de frecuencias para la variable Nivel Socioeconómico (P3)

$x_i(Nivel\ socioeconómico)$	f_i	f'_i	$f_{\%i}$
1 Bajo	2	0,10	10
2 Medio-Bajo	8	0,40	40
3 Medio	7	0,35	35
4 Medio-Alto	2	0,10	10
5 Alto	1	0,05	5
Totales	20	1	100

Los dos ejemplos anteriores corresponden a variables cualitativas, del mismo modo se procede para variables cuantitativas discretas. La tabla de distribución de frecuencias para la variable edad se construye inicialmente contando la cantidad de personas con una determinada edad; es decir, calculando las frecuencias absolutas y a partir de ellas las demás frecuencias como se ilustra en el siguiente ejemplo:

$x_i(edad)$	f_i	f'_i	$f_{\%i}$
18	2	0,10	10
19	10	0,50	50
20	3	0,15	15
21	3	0,15	15
22	1	0,05	5
23	1	0,05	5
Totales	20	1	100

Afortunadamente existen programas estadísticos que nos dan estas tablas de distribución de frecuencias, especialmente útiles para muestras de gran tamaño.

A partir de la distribución de frecuencias es posible construir gráficos, que muestran la misma información de las tablas con el impacto de la visualización de los resultados. Éstos difieren según sean las variables cuali o cuantitativas.

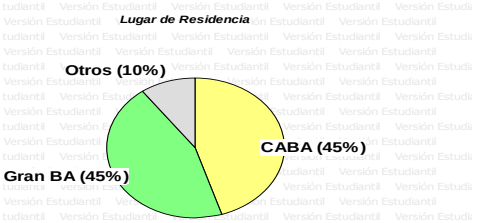
Así, para variables cualitativas medidas a nivel nominal, suelen utilizarse el diagrama circular o diagrama de sectores circulares y para nivel tanto nominal como ordinal el gráfico de barras.

A continuación se exhiben estos gráficos para alguna de las distribuciones de frecuencias antes mencionadas.

Diagrama Circular

En este gráfico los sectores circulares son proporcionales a las frecuencias de los valores de la variable, razón por la cual se pueden expresar las amplitudes angulares de los sectores medidas en grados sexagesimales (llamémoslas α_i); por ejemplo, en función de las frecuencias relativas, como $\alpha_i = f'_i * 360°$.

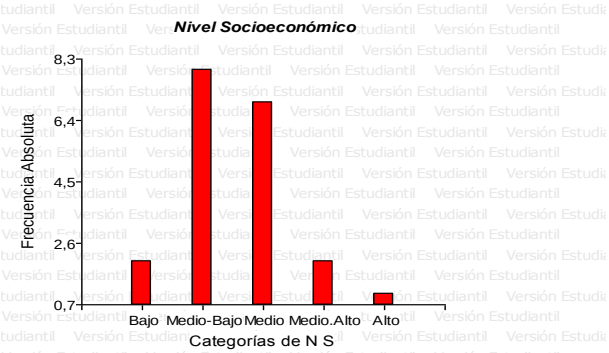
Diagrama Circular para Lugar de residencia



CABA	$\alpha_1 = f'_1 * 360° = 0.45 * 360° = 162°$
Gran Bs As	$\alpha_2 = f'_2 * 360° = 0.45 * 360° = 162°$
Otros	$\alpha_3 = f'_3 * 360° = 0.10 * 360° = 36°$

Gráfico de Barras

En este tipo de gráfico se dibujan dos ejes perpendiculares de los cuales el eje horizontal donde se exhiben rectángulos generalmente del mismo ancho y equiespaciados que representan a los valores de la variable en cuestión, no es numérico en sentido estricto, a lo sumo se le adjudica un orden. El eje vertical en cambio sí es numérico y allí se representan algún tipo de frecuencias. Utilizaremos este tipo de gráfico para la Variable Nivel Socioeconómico, variable cualitativa medida a nivel ordinal, por lo cual sus valores pueden ordenarse, por ejemplo, de menor a mayor nivel socioeconómico (de izquierda a derecha en el gráfico).



Para las variables cuantitativas discretas se utiliza el llamado gráfico de bastones o también el polígono de frecuencias. Como es habitual en las representaciones gráficas matemáticas, hay un sistema de ejes numéricos perpendiculares tales que en el eje horizontal o de abscisas se muestran los valores de la variable y en el vertical o de ordenadas se grafica algún tipo de frecuencia.

Gráfico de bastones

Estos gráficos deberían ser líneas similares a las que se ven a continuación porque indican que toda la frecuencia se concentra en un punto y no en un intervalo alrededor del mismo (como en las variables continuas). Sin embargo, es común que se utilicen barras para visualizarlos mejor o por limitaciones de los programas. De ahí que muchos autores, Bologna entre ellos, también los llaman diagrama de barras.

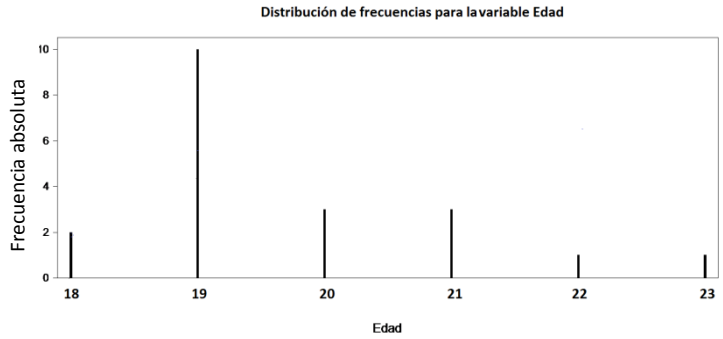


Diagrama de tallo - hoja

Otro modo de mostrar las observaciones de una variable cuantitativa es utilizar un esquema de presentación de los datos que es una combinación de tabla y gráfico como lo es el denominado diagrama de tallo-hoja (Stem and Leaf plot). Esta presentación de los datos consiste en separar cada dato en el último dígito, que se denomina hoja y las cifras delanteras restantes, que forman el tallo.

Utilizando el programa Statistix, para confeccionar este diagrama en el caso de la variable edad de la base de datos se obtiene lo siguiente:

```
Stem and Leaf Plot of Edad

   Leaf Digit Unit = 0.1                    Minimum  18.000
   18  0  represents 18.0                   Median   19.000
                                            Maximum  23.000
         Stem  Leaves
        2   18  00
      (10)  19  0000000000
        8   20  000
        5   21  000
        2   22  0
        1   23  0

20 cases included   1 missing cases
```

En la primera (Leaf Digit Unit = 0,1) se indican las unidades en que están expresadas las hojas. En este caso, la hoja corresponde a la cifra de los décimos que en todos los casos es 0, ya que los números son todos enteros. En la segunda línea a la izquierda se ejemplifica cómo deben leerse los datos así 18 0 representa al 18.0 o también al 18. El diagrama de tallo-hoja que se muestra consta de tres columnas. En la tercera están las hojas (leaves), en la segunda se presentan los tallos (stems). Si se observa la segunda línea se lee que en ella hay 10 observaciones todas iguales a 19.0 o también a 19 por eso aparecen como hojas 10 ceros. El valor entre paréntesis (10) indica que en esa línea o tallo hay 10 observaciones y que contiene un resumen estadístico, llamado mediana y que se estudiará en otro capítulo. Los valores de la primera columna por encima del que está entre paréntesis (en el ejemplo es sólo el 2) son frecuencias acumuladas de menor a mayor valor de la variable hasta el valor que ocupa la posición central, la mediana (concepto que se verá en el próximo capítulo), mientras que los valores de la primera columna que están por debajo del que está entre paréntesis (en el ejemplo el 8, 5, 2 y 1) son frecuencias acumuladas de mayor a menor valor de la variable hasta la mediana.

Veamos otro ejemplo de este tipo de diagrama, utilizando nuevamente el Statistix ahora para la variable de la matriz de datos correspondiente a la columna P9 donde se registra el tiempo empleado en responder al cuestionario, cuyos valores se miden con precisión de un decimal distinto de cero.

```
Stem and Leaf Plot of Tiempo empleado en responder

   Leaf Digit Unit = 0.1                Minimum  3.0000
   3  0  represents 3.0                 Median   4.0000
                                        Maximum  5.5000

         Stem  Leaves
      6    3  000223
      9    3  589
     (4)   4  0002
      6    4  5578
      2    5  00

20 cases included   0 missing cases
```

Nuevamente se indica en la segunda línea a la izquierda cómo deben leerse las observaciones. Así 3 0 representa al dato 3.0, como también el 3 8 representa al dato 3.8. Si se lee la segunda línea debajo de **Stem and Leaves** se observan que hay tres valores de la variable que son el 3.5, 3.8 y el 3.9.

Los tallos 3 y 4 se partieron en dos, según sus hojas van de 0 a 4 o de 5 a 9. La necesidad de partir los tallos se presenta cuando hay poca variabilidad entre los tallos; en este caso hay solamente 3 tallos diferentes.

Observemos que el diagrama de tallo-hoja es una manera de agrupar los datos en intervalos: en cada línea hay tantas hojas como frecuencia del intervalo. Por ejemplo, el primer intervalo va desde 3 hasta 3,4 y tiene frecuencia 6 (6 hojas); el segundo de 3,5 a 3,9 con frecuencia 3; el tercero de 4 a 4,5 (frec 4), el cuarto de 4,5 a 4,9 (frec 4) y el último 5 a 5,5 (frec. 2). La forma del diagrama es similar al de un histograma rotado. Luego, esta disposición de los datos permite, al mismo tiempo que leer los datos, apreciar la forma de la distribución.

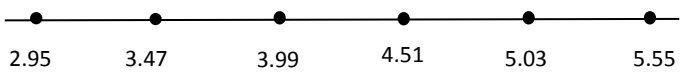
Hasta el momento se han mostrado distribuciones de datos, en el caso de variables cuantitativas, sin ser agrupados en intervalos. Consideremos nuevamente en la matriz de datos los valores de la columna P9, son los valores de una variable cuantitativa continua como lo es el tiempo empleado en responder al cuestionario. Dado el tipo de variable considerada se sabe que no puede hacerse un listado de sus valores y adjudicarle una frecuencia absoluta a cada uno de ellos. Sabemos que los valores observados en realidad representan el centro de un intervalo de números reales por lo tanto las frecuencias deberán ser asignadas a intervalos de valores de la variable. ¿Cómo determinar los intervalos de valores de la variable a los que se le asignarán frecuencias? La respuesta puede variar según los textos. Aquí daremos un criterio bastante general y relativamente simple, aunque no es único.

Consideremos las veinte observaciones de la variable tiempo empleado en contestar el cuestionario (medido en minutos)

3 4.5 3.8 4.8 3.2 3.9 3.5 4.5 4 5

4 4.2 3.3 3.2 3 5 4.7 4 5.5 3

La diferencia entre los valores extremos de la variable es igual a la diferencia entre el valor máximo y el valor mínimo, es decir, 5.5-3=2.5 que llamaremos recorrido de esa variable o amplitud. Con una pérdida no relevante de precisión podemos, de modo que todos los valores sean admitidos en los intervalos a construir, considerar como recorrido a la diferencia 5.55-2.95= 2.6. Nos hemos desplazado 0.05 centésimos hacia la izquierda del mínimo valor observado y 0.05 centésimos hacia la derecha del máximo valor observado. Es decir, hemos aumentado la amplitud en 0.1 = 0.05+0.05 de cada lado que es el orden de precisión de los datos. De este modo nos aseguramos que el aumento de amplitud sea simétrico y los valores extremos estén incluidos en el nuevo recorrido y también que los demás valores de la variable no coincidan con los límites de los intervalos a construir. Elegimos dividir el recorrido de extremos 2.95 y 5.55 en cinco intervalos iguales por lo que la amplitud de cada uno de ellos será de $\frac{2.6}{5} = 0{,}52$. Estos intervalos reciben el nombre de clases. Los cinco intervalos de clase con sus límites correspondientes están indicados a continuación.



El punto medio de cada intervalo se denomina marca de clase. En el ejemplo son:

3.21 3.73 4.25 4.77 5.29

La cantidad de observaciones de la variable en estudio que pertenecen a una clase particular, es la frecuencia absoluta de esa clase y tal frecuencia queda asignada a su correspondiente marca de clase. Los extremos de los intervalos considerados suelen denominarse extremos exactos, para los que el límite superior de uno de ellos coincide con el límite inferior del siguiente. A partir de determinar la frecuencia absoluta de cada intervalo de clase pueden hallarse las frecuencias relativas y porcentuales.

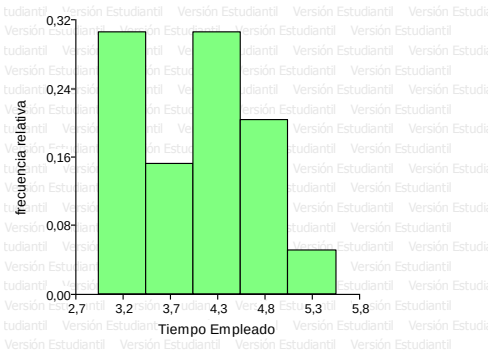
Podemos construir la tabla de distribución de frecuencias para las observaciones de tiempo empleado en contestar un cuestionario, agrupadas en intervalos de clase.

Clase	$x_i^{´}$: Marca de clase	f_i	$f^{´}_i$	$f_{\%i}$
2,95 – 3,47	3,21	6	0,30	30
3,47 – 3,99	3,73	3	0,15	15
3,99 – 4,51	4,25	6	0,30	30
4,51 – 5,03	4,77	4	0,20	20
5,03 – 5,55	5,29	1	0,05	5
Totales		20	1	100

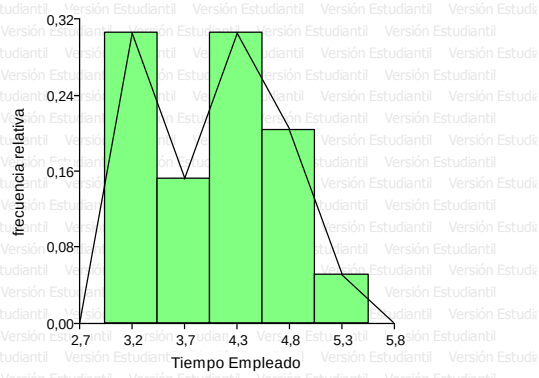
La agrupación de datos en intervalos suele ser el procedimiento adecuado para tratar a las variables cuantitativas continuas, pero puede ser utilizado para las variables cuantitativas discretas generalmente cuando se dispone de muchos datos y hacer la presentación detallada de cada una de las categorías conduciría a una tabla muy grande por lo que agrupar valores en intervalos permite hacer una presentación más sencilla. Usualmente se utilizan no menos de cinco intervalos y no más de veinte, ya que con menos de cinco se acentúa la pérdida de precisión y con más de veinte se pierden las características de resumen.

Para visualizar el comportamiento de las distribuciones de frecuencia de datos agrupados en intervalos, se confeccionan rectángulos verticales y contiguos, cuyas bases son los intervalos de clase y sus alturas son proporcionales a las frecuencias (generalmente relativas) correspondientes a cada clase. Este tipo de gráfico se denomina Histograma.

Histograma para la variable tiempo empleado para terminar el cuestionario.



El polígono cuyos vértices son: el punto medio de un intervalo previo al primero de altura cero, los puntos correspondientes a las marcas de clase y las alturas correspondientes a cada rectángulo y que finaliza en el punto medio de un intervalo posterior al último de altura cero se denomina polígono de frecuencias y se muestra a continuación para la variable considerada.



Aunque las frecuencias se grafican en las alturas de los rectángulos, siendo todos los intervalos de igual longitud, el área resulta proporcional a la frecuencia, por lo que podemos interpretar las áreas como frecuencias. El polígono de frecuencias "suaviza" el histograma ayudando a abstraer una curva que muestra "la forma de la distribución". Más adelante se verá que, al modelizar la distribución de frecuencias de las variables continuas, dicha curva se denomina "función de densidad de probabilidad". Nótese en el gráfico que el área debajo del polígono de frecuencias es igual al área debajo del histograma, ya que por cada triángulo que queda fuera del polígono en un rectángulo, se agrega otro de igual área en el rectángulo del intervalo siguiente.

A partir de estos gráficos podemos apreciar la diferencia esencial que hay entre una variable discreta y una continua:

Mientras que en una variable discreta cada punto se lleva una parte de la frecuencia total, en la continua la frecuencia total "se desparrama" a lo largo de todo

un intervalo de valores de manera tal que a cada punto en particular no le corresponde nada de la frecuencia sino que las frecuencias son de los intervalos de valores. En síntesis: en las variables discretas las frecuencias se concentran en valores individuales mientras que en las continuas reparten en intervalos.

Hasta ahora se presentaron por un lado datos sin agrupar y se definió la distribución de frecuencias para una variable, ya sean frecuencias absolutas, relativas o porcentuales y por otro lado se mencionó el tratamiento de datos agrupados en intervalos. En todos los casos se asignó, siguiendo algún criterio, una frecuencia a cada valor puntual de la variable o cada intervalo de clase. Sin embargo, hay situaciones en que se quiere asignar frecuencia a valores iguales o menores que uno dado (acumulación descendente) es decir a la acumulación de valores de una variable, supongamos cuantitativa. Por tal motivo se definen las frecuencias acumuladas como sigue.

Frecuencia acumulada absoluta de un valor X_i es la cantidad de casos que asumen ese valor y todos los valores menores a él. Se simboliza con una mayúscula F_i.

Frecuencia acumulada relativa de un valor X_i es la proporción de casos que asumen ese valor y todos los menores a él. Se simboliza con una mayúscula F'_i.

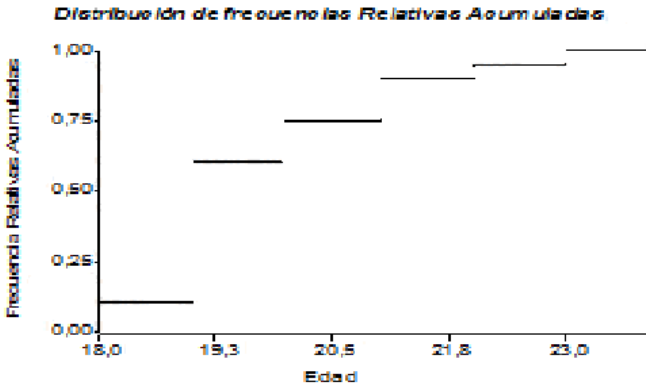
Frecuencia acumulada porcentual de un valor X_i es el pocentaje de casos que asumen ese valor y todos los valores menores a él. Se simboliza con una mayúscula $F\%_i$.

Las frecuencias acumuladas tienen sentido de definirse para variables medidas desde nivel ordinal en adelante, siendo el criterio de acumulación el de menor a mayor grado del atributo. La acumulación puede realizarse también, en sentido descendente, es decir considerar los valores mayores que uno dado o bien si se trata de variables medidas a nivel ordinal acumular de mayor a menor grado del atributo.

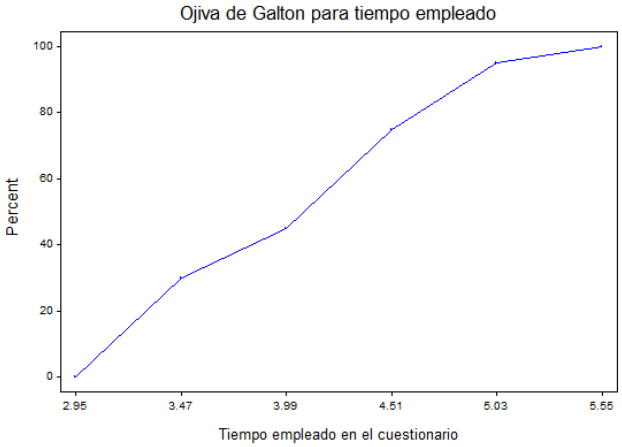
Así para la variable P3 (Nivel socioeconómico) a la distribución de frecuencias simples, ya calculada se le podrían agregar tres columnas más para mostrar las frecuencias acumuladas absolutas, relativas y porcentuales, como se muestra a continuación.

$x_i(nivel\ socioeconómico)$	f_i	f'_i	$f_{\%i}$	F_i	F'_i	$F\%_i$
1 Bajo	2	0,10	10	2	0,10	10
2 Medio-Bajo	8	0,40	40	10	0,50	50
3 Medio	7	0,35	35	17	0,85	85
4 Medio-Alto	2	0,10	10	19	0,95	95
5 Alto	1	0,05	5	20	1	100
Totales	20	1	100			

Si la variable es discreta, cada valor aporta su frecuencia, que “salta” en el valor siguiente por lo que el gráfico de las frecuencias acumuladas (ya sean absolutas, relativas o porcentuales) tiene forma escalonada. A continuación se muestra el gráfico de frecuencias acumuladas absolutas para la variable edad.



Si la variable es continua, las frecuencias se van acumulando de modo que quedan representadas por una línea quebrada ascendente, denominada Ojiva de Galton y permite interpolar valores no observados y que no aparecen en la tabla. Así se puede ilustrar la frecuencia acumulada relativa para la variable P9 (tiempo empleado en responder al cuestionario) en el gráfico que sigue a continuación.



Bibliografía

Bologna, E. (2018) *Métodos estadísticos de Investigación*. Ed. Brujas. Ciudad de Córdoba. Argentina

Botella Usina, J.; Suero Suñe, M. y Ximenez Gómez, C. (2012) *Análisis de datos en Psicología I.* Ed Pirámide. Madrid, España.

Clases Teóricas correspondientes al curso de Estadística dictado por Horacio Attorresi y María Silvia Galibert (1994).

Clases Teóricas desgravadas correspondientes al curso de Estadística dictado por Carlos Pano (1996).

Galibert, M., Hoyos Páez, C. y Alvarez Ponte, L. (2019). Construcción de un cuestionario para detectar la impuntualidad crónica. *XI Congreso Internacional de Investigación y Práctica Profesional en Psicología*, Fac. Psicología, UBA. Libro de resúmenes.