

Manual de

Estadística Básica

Guía para Observatorios
Provinciales de Drogas

Dirección del Observatorio
Argentino de Drogas



SEDRONAR
Presidencia de la Nación

Contenido

FUNDAMENTACIÓN	4
OBJETIVO DEL MANUAL	6
I. Definición de Estadística	7
II. La Estadística y su relación con la investigación científica	8
III. Clasificación de la estadística	9
Estadística Descriptiva	9
Estadística Inferencial	9
IV. Fuentes de información estadística	10
V. Métodos de obtención de datos de fuentes primarias	11
Censos	11
Estudios muestrales	11
Registros Permanentes	11
VI. Estudio estadístico	13
Unidades estadísticas	13
Variables estadísticas	13
Tipos de variables	14
VII. Procesamiento de la información estadística	15
Nociones de Tablas, Cuadros y Gráficos	15
Distribuciones de Frecuencias	16
Presentación de datos en gráficos	18
Principales Tipos de Gráficos	19
VIII. Interpretación de los datos	23
Sistematización de los Datos	23

Medidas de resumen	23
Medidas de posición	24
Medidas de dispersión	26
Indicadores	27
Tipo de Relaciones entre Variables o Cifras	28
IX. Nociones de estadística inferencial	34
Población.....	35
Muestra	35
Parámetros	35
Estimadores.....	36
X. Nociones de muestreo	37
Muestreo No Aleatorio.....	37
Muestreo Aleatorio o Probabilístico.....	37
Técnicas Muestrales.....	39
Muestreo Aleatorio Simple (MAS).....	39
Muestreo Sistemático (MS).....	40
Muestreo Replicado.....	40
Muestreo Estratificado (ME).....	40
Muestreo por Conglomerados.....	40
Paneles.....	41
BIBLIOGRAFÍA	42

Fundamentación

El Manual de Estadística, conjuntamente con el de Epidemiología y el Manual de Procedimientos en SIG (Sistemas de Información Geográfica) constituyen una serie de herramientas de trabajo que el Observatorio Argentino de Drogas (OAD) de la Secretaría de Programación para la Prevención de la Drogadicción y la Lucha contra el Narcotráfico (SEDRONAR), pone a disposición de los observatorios de drogas provinciales y municipales, así como a todo profesional o técnico de la salud que requiera del desarrollo de competencias en investigación en estos temas de trabajo.

El Observatorio Argentino de Drogas (OAD) se ocupa de la generación y recopilación de información oportuna, válida y confiable sobre los distintos aspectos del consumo de sustancias psicoactivas. Esta información es necesaria para la definición de políticas públicas integrales, inclusivas, participativas, acordes a las necesidades de toda la población y basadas en evidencia científica. La producción empírica que aporta el OAD es un insumo central para la definición de las estrategias de prevención y la ampliación eficaz de la oferta de tratamiento para la problemática del consumo en todo el territorio nacional.

¿Quiénes conformamos el OAD?

El equipo del observatorio es interdisciplinario y está conformado por profesionales con extensa trayectoria en investigación; pertenecen a las siguientes disciplinas: Psicología, Sociología, Geografía, Estadística y Ciencias Políticas.

Desde este un enfoque social y participativo, la SEDRONAR se encuentra abocada a la redefinición de su perfil institucional en base a ejes prioritarios de gestión, entre ellos la elaboración de un Plan Nacional de Drogas, el trabajo conjunto con el COFEDRO (Consejo Federal de Drogas) como instancia nodal del fortalecimiento de la perspectiva federal. En este marco, el Observatorio de Drogas tiene como uno de sus ejes prioritarios de trabajo, la vinculación con las provincias y municipios con miras a la producción de información epidemiológica. El fortalecimiento de los observatorios provinciales tiene por finalidad asegurar la producción y disponibilidad de datos válidos y confiables para la evaluación de la situación local sobre el consumo problemático de sustancias.

Para ello, se requiere trabajar de forma intensa y sostenida en la capacitación y desarrollo de competencias de los equipos de los observatorios provinciales para la realización de actividades tales como el diseño de instrumentos y metodologías para la recolección de datos, ejecución de programas de investigación, profundización del análisis de las muestras provinciales, entre otras. A partir de las acciones del OAD y del fortalecimiento del COFEDRO, se busca que las provincias y municipios desarrollen sustentabilidad en el registro, análisis y seguimiento de indicadores clave para la formulación de sus políticas públicas locales, acordes a las especificidades de sus poblaciones (según género, edad, etnia y otras particularidades). Esta estrategia de trabajo contribuye en el desarrollo del espíritu federal y en la articulación entre las diversas jurisdicciones (nacional, provincial y municipio) con la Sociedad Civil, aspirando a consolidar la gestión de la SEDRONAR.

Los principios que orientan nuestro trabajo incluyen la articulación con diferentes actores -universidades, organismos gubernamentales, agencias de ciencia y técnica, comunidades, organizaciones de la sociedad civil, servicios de salud-; la integración de distintos métodos y niveles de análisis, acordes a los distintos objetivos y propósitos de las investigaciones; la atención a las múltiples dimensiones que conforman la problemática del consumo de sustancias y el abordaje analítico desde un enfoque interdisciplinario.

Los contenidos del Manual han sido elaborados a partir de la consulta de múltiples fuentes bibliográficas y de la experiencia del equipo del OAD. Con la finalidad de facilitar y dinamizar la lectura de este material, decidimos citar las fuentes hacia el final del Manual, exceptuando aquellas situaciones en las cuales se realicen citas textuales.

Objetivo del Manual

El presente documento tiene como objetivo ser una guía para los equipos profesionales y técnicos de los Observatorios Provinciales de Drogas que como parte de su quehacer profesional requieran del uso de herramientas estadísticas.

El enfoque del material no se centra en el cultivo de habilidades de cálculo y uso de fórmulas, sino más bien en la comprensión de los procedimientos, procesos y aportaciones de la Estadística, obviando un excesivo desarrollo teórico.

Por ello, con este manual se espera que a partir de los conocimientos teóricos básicos (particularmente de la inferencia estadística) los profesionales puedan entender el cómo y el porqué de las herramientas de la inferencia estadística como así también interpretar los resultados como parte de un razonamiento crítico basado en evidencias objetivas obtenidas de una población bajo estudio.

Objetivos Específicos

- Aportar recursos cognitivos básicos para el desarrollo de informes a los equipos técnicos y/u Observatorios Provinciales de drogas.
- Describir los procedimientos estadísticos elementales siendo éstas herramientas necesarias para la continuidad o implementación de estudios sobre la temática de consumo de sustancias.
- Ofrecer un material de consulta como aporte para la mejora de los procesos de planificación, ejecución y evaluación de los estudios que desarrollen los Observatorios Provinciales de drogas.

I. Definición de estadística

Diversas definiciones de Estadística, coinciden en señalarla como la Ciencia formada por un conjunto de teorías y técnicas mayoritariamente cuantitativas, que tienen por objeto la organización, presentación, descripción, resumen y comparación de conjunto de datos numéricos, obtenidos de poblaciones en su conjunto de individuos o fenómenos o bien de muestras que representan las poblaciones estudiadas, así como el estudio de su variación, propiedades, relaciones, comportamiento probabilístico de dichos datos y la estimación, inferencia o generalización de los resultados obtenidos de muestras, respecto a las poblaciones que aquéllas representan. En síntesis, el uso de la Estadística nos permite tener información lo más próxima a la realidad bajo estudio en base a información válida y confiable.

II. La estadística y su relación con la investigación científica

Todo proceso de investigación y/o estudio estadístico implica la realización de una serie de pasos tales como determinar qué estudiar y para qué, qué datos recolectar, dónde y cómo recolectarlos, cómo procesarlos, cómo interpretarlos, etc., valiéndose además de distintos tipos de diseños.

La elección de un determinado diseño depende fundamentalmente del objetivo que se busque alcanzar y de las posibilidades que existan para su ejecución.

El poder explicitar estos aspectos previamente al diseño del estudio, ayuda mucho para la selección de la metodología más apropiada. Es decir, el diseño metodológico del estudio, la ejecución de cada una de las etapas y el posterior análisis de la información estarán subordinados a la definición de “qué estudiar” y “para qué”.

Qué estudiar significa definir el problema que deseamos investigar. Para ello generalmente nos basamos en diversas fuentes de información y las confrontamos. Por otra parte, definir el problema permite focalizar en un campo de estudio. A su vez, como se pueden perseguir distintos objetivos hay que plantearse “¿Para qué investigar el problema en cuestión?”. De acuerdo entonces a los objetivos planteados, se definirá el tipo de estudio que se requiere realizar.

La elección del tipo de estudio, además del objetivo de su investigación, depende, entre otros factores, de los datos con que se cuente para la investigación y de la disponibilidad de tiempo y recursos.

Una vez establecido el objeto de estudio en base a los conocimientos teóricos, se inicia la etapa de Diseño Metodológico donde se define el proceso de recolección de datos, delimitando las unidades bajo estudio y las variables a medirse, que permitan contestar las preguntas formuladas, en el proyecto de investigación.

Los datos obtenidos, de la realidad investigada, se analizan aplicando métodos y técnicas estadísticas.

III. Clasificación de la estadística

Dependiendo de cómo se analizan los datos, la Estadística se clasifica como:

Estadística descriptiva Rama de la Estadística que trata sobre la descripción y análisis estadístico de una población, que resumen y presenta datos obtenidos de la población o de una muestra, mediante métodos adecuados. Tiene como objetivo, caracterizar los datos, de manera gráfica o analítica, para resaltar las propiedades de los elementos bajo estudio.

Estadística inferencial Rama de la Estadística que estudia el comportamiento y propiedades de las muestras y la posibilidad, y límites, de la generalización de los resultados obtenidos a partir de aquellas a las poblaciones que representan. Esta generalización de tipo inductivo, se basa en la probabilidad. También se le llama también Estadística Matemática, por su complejidad matemática en relación a la Estadística Descriptiva. Tiene como objetivo, generalizar las propiedades de la población bajo estudio, basado en los resultados de una muestra representativa de la población.

IMPORTANTE: Ningún método estadístico puede corregir los defectos por una inadecuada selección del problema que se investiga, o por una mala recolección de datos. Una investigación que empieza mal, con seguridad termina mal. Con datos de mala calidad, no será posible dar respuesta adecuada a un problema científico.

IV. Fuentes de información estadística

Puede decirse a grandes rasgos que las fuentes de información de datos estadísticos son dos:

- Fuentes Primarias: aquellas en las que los datos se obtienen directamente de las unidades originales o iniciales (personas, empresas, organismos, etc.), sin pasar previamente por ninguna entidad o persona encargada de la obtención, procesamiento y/o divulgación. Ejemplos: los datos logrados en un Censo de Población; los resultados de una Encuesta de Opinión; etc.
- Fuentes Secundarias: las que no cumplen la condición anterior; es decir que los datos se obtienen de informes, publicaciones, registros, la utilización de datos proporcionados por el Registro Civil (nacimientos, defunciones, etc.), etc.; efectuados por otras entidades. Ejemplos: el uso de los datos relativos a la Población extraídos de un Anuario; las cifras de una Encuesta Política publicadas en un diario; etc.

Ambas fuentes presentan diferencias incuestionables:

- » Las fuentes secundarias resultan mucho más económicas, ya que usualmente sólo se requiere la consulta a publicaciones, registros, bases de datos, etc. elaborados y/o difundidos por otras entidades.
- » Como contrapartida, esas fuentes pueden contener errores de transcripción, agrupamientos excesivos o incomparables, definiciones insuficientes, y otras limitaciones originadas en un enfoque diferente entre quien genera la información, y el usuario que la reproduce.
- » Las fuentes primarias usualmente incluyen aclaraciones y definiciones detalladas sobre la información que se difunde, mención de la metodología utilizada para obtener la información, las limitaciones u omisiones cometidas, etc.
- » Los principales inconvenientes de éstas son: el costo de obtención y los conocimientos que se requieren para posibilitar una recopilación, procesamiento y análisis eficientes.

Reconociendo sus ventajas y limitaciones, ambas fuentes son útiles para usos estadísticos, aunque debido a exigencias de confianza y seguridad, siempre son preferibles las fuentes primarias.

V. Métodos de obtención de datos de fuentes primarias

Las formas de obtener información estadística, dependen en gran medida de los objetivos que se tengan para su uso, y los recursos que se disponen para ese emprendimiento.

Dejando de lado los casos en que la información se obtiene de fuentes secundarias (en los cuales el método consistirá en copiar en parte o totalmente, los datos existentes en alguna publicación o entidad), los métodos de obtención de datos de fuentes primarias, pueden ser resumidos de la siguiente forma:

- » **Censos:** constituyen una indagación completa sobre todas las unidades o elementos cubiertos en el estudio. Desechando posibles errores en las etapas de recopilación o procesamiento, tienen la ventaja de que el conocimiento censal de una población, asegura la obtención de datos fehacientes. La principal desventaja según los casos, es su costo, el tiempo de obtención de los resultados y/o la imposibilidad de realización. Ejemplos: el Censo Nacional de Población (se realizan cada 10 años), el Censo Nacional Económico, el Censo de Jubilados y Pensionados, etc.

- » **Estudios Muestrales:** como su nombre lo indica, están constituidos por una indagación parcial de las unidades o elementos que conforman la población en estudio. Tiene la ventaja de su accesibilidad económica (y de tiempo para la disponibilidad de los resultados), además de constituir en muchos casos la única posibilidad de aplicación. Ejemplos: la Encuesta Permanente de Hogares (se la utiliza para medir el desempleo), las Encuestas Preelectorales, los estudios del Control de la Calidad de una producción en serie, etc.

- » **Registros permanentes:** una recopilación continua de datos que se van actualizando en forma permanente. En este caso la desventaja está representada en que usualmente su cobertura depende de la buena voluntad o del grado de respeto a las normas de aquellos que deben suministrar la información. Sus grandes ventajas son: el bajo costo de obtención de los datos y la permanente actualización de las cifras que se cubren. Ejemplos: el Registro Nacional de las Personas, el Registro Nacional del Parque Automotor, el Registro Municipal de Inmuebles, etc.

Además, existen combinaciones de los métodos anteriores y formas particulares de obtención de datos que no pertenecen exactamen-

te a los casos previamente descritos. Entre ellos puede mencionarse especialmente el método de simulación o experimentación, que consiste en recrear las condiciones reales de algún problema, y experimentar mediante la ayuda de algún sistema aleatorio.

VI. Estudio estadístico

Como ya se mencionó anteriormente, un Estudio Estadístico, constituye “una serie de pasos o etapas cuya finalidad principal es la observación de los resultados que se obtienen”. Ejemplos extremos de esta definición son:

- Arrojar un dado al aire para observar los puntos que salen en la cara de arriba (su realización toma unos pocos segundos).
- Realizar un Censo Nacional de Población para procesar y difundir sus resultados (desde que se comienza a programar hasta que se publican los últimos resultados pasan varios años).

Estos Estudios Estadísticos se realizan sobre Unidades Estadísticas.

1. Unidades Estadísticas

Se entiende por tales “al objeto de un Estudio Estadístico”. Son aquellas sobre las que se realizan las observaciones o mediciones, o que responden las consultas que se realizan. Por ejemplo:

- Las Personas: en un censo de población, en una encuesta de opinión, etc.
- Las Empresas: en un censo económico, en una encuesta sobre el sistema de riesgos del trabajo, en un estudio sobre insumos de producción, etc.
- Los Países: en un estudio sobre exportaciones agropecuarias, en una encuesta sobre la natalidad de la población, etc.

En las Unidades Estadísticas se miden **Variables Estadísticas**.

2. Variables Estadísticas

Puede definirse como Variable Estadística a “todo aquello pasible de tomar valores diferentes de unidad en unidad”. Por ejemplo:

Unidades Estadísticas	Variables
Personas	» Edad » Estado Civil » Cantidad de Hermanos » Nivel de Estudios alcanzado

Unidades Estadísticas	VARIABLES
Empresas	<ul style="list-style-type: none"> » Rama de Actividad » Tamaño » Cantidad de Personal Ocupado

Unidades Estadísticas	VARIABLES
Países	<ul style="list-style-type: none"> » Exportaciones (en millones de U\$S o Tn) » Cantidad de habitantes alfabetizados

Como puede observarse en los ejemplos de la tabla anterior, estas Variables pueden ser de características muy diversas, y de acuerdo al tipo de resultado que pueden tomar, se las suele clasificar de la siguiente manera:

Tipos de Variables		
Primera Clasificación	Segunda Clasificación	
Cuantitativas Aquellas cuyos resultados se expresan en forma numérica (con cantidades)	Discretas Aquellas cuyos resultados numéricos solo toman algunos valores del campo de definición de la variable.	Ejemplos: Cantidad de hermanos, Personal Ocupado, Cantidad de Países Acreedores, etc.
	Continuas Aquellas que, teóricamente, pueden tomar como resultado cualquier valor dentro del campo de definición de la variable.	Ejemplos: Edad, Producción, Exportaciones, etc.
Cualitativas Aquellas cuyos resultados se expresan en palabras o no numéricamente (con cualidades)	Nominales Aquellas cuyos resultados no expresan dimensión ni orden entre las variables.	Ejemplos: Estado Civil, Rama de Actividad, Rubro principal de Exportaciones, etc.
	Ordinales Aquellas cuyos resultados expresan dimensión u orden entre las variables.	Ejemplos: Nivel de Estudios Alcanzado, Tamaño, Nivel de "Riesgo País".

VII. Procesamiento de la información estadística

Cualquiera haya sido la fuente de información y el método de obtención de los datos, básicamente el procesamiento estadístico de datos tiene por objeto principal resumir la información obtenida.

A partir de que se dispone de un conjunto de datos recopilados, la tarea de clasificar y tabular información estadística implica: “Disponer los datos en forma ordenada por medio de tablas, de modo que se facilite su análisis e interpretación, y de ser necesario se agilicen los cálculos”.

Esta “disposición ordenada de los datos”, toma formas diferentes según el tipo y complejidad de la información con la que se trabaja. Pueden tenerse: tablas simples (de una variable), tablas complejas (con dos o más variables), series de tiempo, tablas de asociación entre variables, etc.

1. Nociones de Tablas, Cuadros y Gráficos

Dentro de la Estadística Descriptiva, se denominan “Tablas” a:

“Los resultados resumidos de procesar un conjunto de datos, que quedan dispuestos en forma ordenada a partir de una clasificación previamente determinada”.

Estas “tablas” son una etapa dentro del procedimiento de trabajo, y no necesariamente deben tener “buena presentación”, ni contener todos los elementos que permitan identificar a qué corresponden, y hasta es probable que algunos de ellos sean representados por símbolos. Esto se debe a que el objetivo de las tablas es presentar los datos en estudio en forma resumida y ordenada para su análisis por parte de quién realiza el estudio estadístico.

Un Cuadro estadístico, debe poseer todos esos elementos que no se requieren en las Tablas, es decir: una precisa identificación de a qué tipo de información se refieren, cómo se la clasifica, de dónde provienen, a qué momento del tiempo representan, etc. La causa de esto, es que los cuadros tienen por **objeto** presentar la misma información que las tablas, pero en forma “autosuficiente”, es decir que por sí solos, deben ser comprensibles y explicativos.

Finalmente los **Gráficos** estadísticos deben ser precisamente eso: “gráficos”, entendiéndose por tales, que mediante una simple visualización de la figura representada, se puede tener una idea general

de la información que se presenta. Los gráficos también deben ser “autosuficientes”, aunque no deben contener demasiados detalles numéricos ni explicativos.

Por lo expuesto, queda claro que el objeto de presentar la información estadística en forma de tablas, cuadros y gráficos, no es otro que el de facilitar el análisis y la interpretación de los datos.

2. Distribuciones de Frecuencias

Agrupamiento de Datos

Aunque los métodos estadísticos pueden ser aplicados a datos dispuestos en forma aislada (es decir, sin clasificarlos ni agruparlos previamente), en general al tener que manejarse conjuntos de datos relativamente numerosos, la primera etapa de un estudio estadístico sugiere la conveniencia o necesidad de agruparlos. La forma de hacerlo depende en gran medida del tipo de variable que se analiza, la unidad de medida y la precisión utilizada durante la etapa de recopilación para efectuar la medición.

Una vez lograda la agrupación de los datos en diferentes categorías o intervalos, el paso siguiente es determinar la frecuencia de observaciones en cada categoría o intervalo de la variable y construir la Tabla de Distribución de Frecuencias.

Esta tabla de distribución de frecuencias permite ordenar el número de individuos que pertenecen a cada categoría de la variable. La frecuencia de observaciones en cada categoría puede expresarse como una frecuencia absoluta (total de observaciones en una determinada categoría) o como una frecuencia relativa (proporción o porcentaje en que un atributo o cualidad se manifiesta dentro de una determinada categoría).

Por ejemplo, en un estudio se realizó una encuesta a 120 pacientes con consumo problemático de cocaína y se determinó el “número de internaciones que refería haber presentado en el último año”. Los datos fueron volcados en la tabla que se presenta a continuación:

Tabla de frecuencias¹ del “Número de internaciones por consumo de cocaína en el último año”.

¹ Frecuencia absoluta de un valor observado. Constituye la cantidad de veces que se presenta o repite el valor de la variable.

Frecuencia absoluta acumulada de un valor observado. Constituye la cantidad de casos que se tienen hasta el valor de la variable.

Frecuencia relativa de un valor observado. Constituye la proporción de veces que se presenta o repite el valor de la variable. Es la frecuencia en relación al total de unidades consideradas. Generalmente se expresa multiplicada por 100 (%).

Frecuencia relativa acumulada de un valor observado. Constituye la proporción de casos que se tienen hasta el valor de la variable. Generalmente se expresa multiplicada por 100 (%).

Número de internaciones	Nº de individuos	Frecuencia relativa (%)	Frecuencia acumulada (%)
0	22	18,4	18,4
1	39	32,5	50,9
2	31	25,8	76,7
3	16	13,4	90,1
4	8	6,6	96,7
5	3	2,5	99,2
6	1	0,8	100,0
Total	120	100,0	

Analizando e interpretando la tabla anterior se puede establecer por ejemplo, que de los 120 pacientes con consumo problemático de cocaína encuestados 8 (primera columna, frecuencia absoluta) referían haber presentado 4 internaciones en el último año, o bien que el 2,5% tuvo 5 internaciones (frecuencia relativa %) y también se sabría por la tabla, que 76,7% tuvo a lo sumo 2 hospitalizaciones (frecuencia acumulada%).

Cuando la variable cuyos datos se pretende organizar fue medida con una escala numérica continua o con una escala numérica discreta pero con un rango amplio de valores deben construirse intervalos y presentar esta información en una tabla de frecuencias con intervalos.

Por ejemplo, en el caso de la variable “nivel de cocaína en sangre” utilizaríamos una escala numérica continua para su medición y los datos deberían agruparse en intervalos para facilitar su interpretación.

Intervalos	Nº de individuos	Frecuencia relativa (fr %)	Frecuencia acumulada
3.0 - 3.9	17	6,3	6,3
4.0 - 4.9	73	27,2	33,5
5.0 - 5.9	101	37,7	71,2
6.0 - 6.9	54	20,1	91,3
7.0 - 7.9	18	6,7	98,0
8.0 - 8.9	4	1,5	99,5
9.0 - 9.9	1	0,5	100,0
Total	268	100,0	

Los extremos de estos intervalos son llamados límites, distinguiéndose el límite inferior de cada intervalo (3,0; 4,0... en el ejemplo) y el límite superior (3,9; 4,9...). La diferencia entre el límite superior y el límite inferior se denomina amplitud del intervalo.

Con la agrupación de los datos en intervalos se gana en comodidad, pero se pierde información, ya que una vez agrupados, todos los datos pertenecientes a un intervalo serán identificados para el análisis

posterior con un punto medio o marca de clase de ese intervalo. Para calcular dichas marcas de clase se realiza la siguiente fórmula:

$$\text{Marca de clase del Intervalo} = \frac{\text{Límite superior} + \text{límite inferior}}{2}$$

En el ejemplo de los niveles de cocaína en sangre, la marca de clase para el primer intervalo sería:

$$\text{Marca de clase} = \frac{3,99 + 3,0}{2} = 3,5$$

El número de intervalos recomendable es aquel que permita presentar los datos de manera resumida sin distorsionar los datos. La utilización de demasiados intervalos difiere poco de la tabulación original de los datos, y un número demasiado reducido puede enmascarar información esencial.

Por otra parte, los intervalos que se construyan deben ser mutuamente excluyentes, de modo de evitar la mala clasificación de los datos. También es conveniente para la determinación de los intervalos:

- Identificar el dato correspondiente al valor máximo y el correspondiente al valor mínimo.
- Determinar la diferencia entre estos dos valores (rango).
- Dividir esta diferencia por el número de intervalos que se desea obtener para identificar así la amplitud de cada intervalo.

3. Presentación de datos en gráficos

En ocasiones, se prefiere representar gráficamente los datos, con el objeto de obtener una rápida impresión visual del conjunto. Para ello se pueden utilizar diferentes tipos de gráficos, contemplando las siguientes premisas:

El gráfico debe ser sencillo y explicarse por sí mismo.

- Es preferible que el gráfico no contenga demasiada información y su lectura sea fácil.
- El diseño debe ser atractivo, pero sin deformar los hechos que está describiendo.
- Se debe seleccionar el gráfico más apropiado de acuerdo al tipo de variable y la escala de medición utilizada para medirla.

El tipo de gráfico está condicionado por el tipo de escala utilizada para medir la variable que desea graficar.

3.1. Principales Tipos de Gráficos

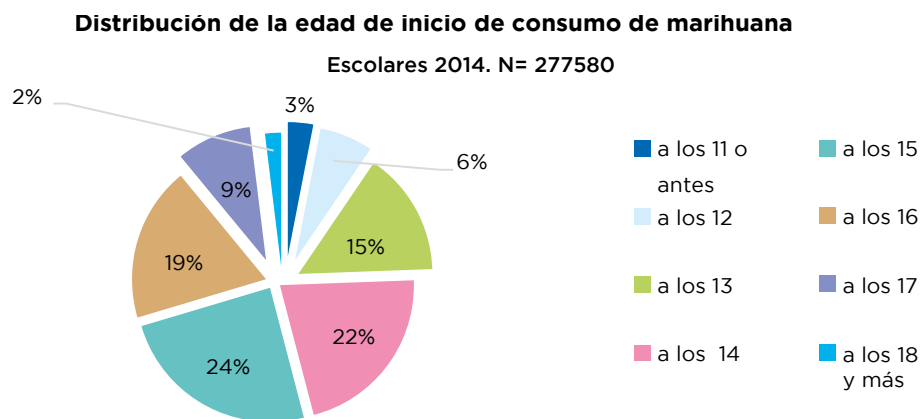
Gráfico de Sectores Circulares

Consiste en un círculo dividido en sectores con áreas proporcionales a la frecuencia con que se presenta cada categoría.

Habitualmente se utiliza este tipo de gráfico para presentar datos que se miden con una escala nominal, pero también puede utilizarse para variables ordinales o numéricas discretas con un rango pequeño de valores.

Es muy importante no omitir el número total de individuos estudiados (n) porque la interpretación del gráfico es muy diferente de acuerdo al tamaño de la población cuyos datos se están representando.

Ejemplo de gráfico circular



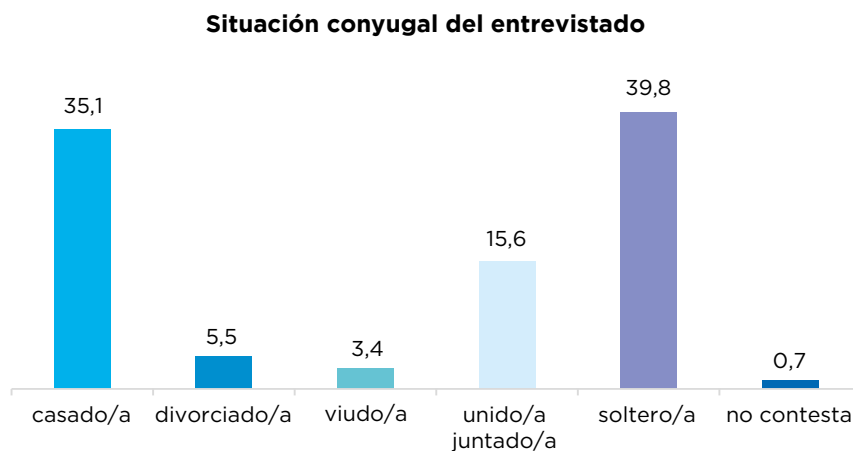
Fuente: Sexto Estudio Nacional sobre Consumo de Sustancias Psicoactivas en Estudiantes de Enseñanza Media 2014.

Gráfico de Barras

Este tipo de gráficos se construye dibujando barras o rectángulos en los cuales la longitud indica la magnitud o la frecuencia de cada categoría de la variable. Se utiliza para representar la distribución de frecuencias de variables en escala nominal, así como para variables en escala ordinal y numérica discreta cuando el rango de valores de la variable es reducido.

En el eje horizontal se indican los valores de la variable y en el vertical sus correspondientes frecuencias.

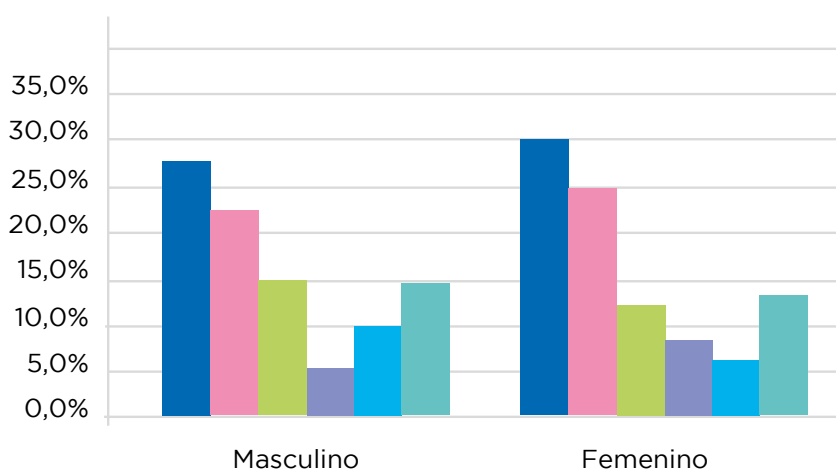
Ejemplo:



El gráfico de barras es muy útil para comparar no sólo la distribución de frecuencias de las diferentes categorías de una variable, sino también la distribución de acuerdo a más de una variable.

Ejemplo:

Distribución de la frecuencia de consumo de pasta base/paco. Escolares 2014

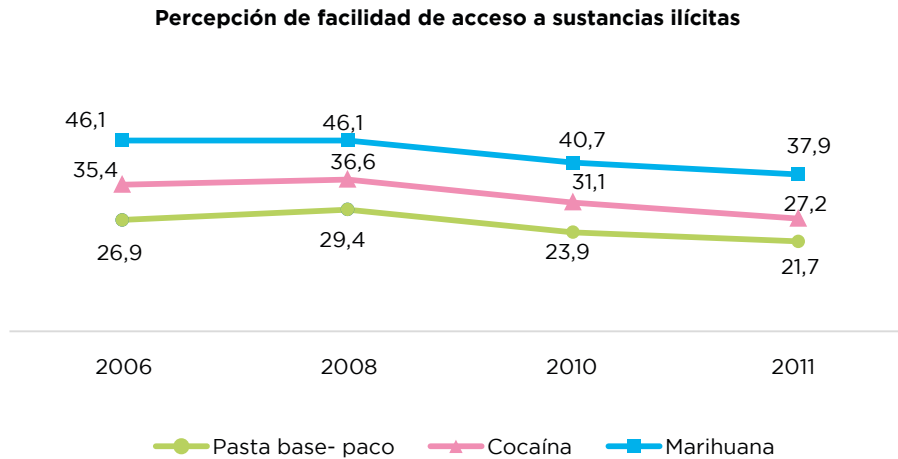


- Una sola vez
- Algunas veces durante los últimos 12 meses
- Algunas veces mensualmente
- Algunas veces semanalmente
- Diariamente
- Ns / Nc

Gráfico de líneas

En este tipo de gráfico se representan los valores de los datos en dos ejes cartesianos ortogonales entre sí. Se pueden usar para representar una serie, dos o más.

Ejemplo:



Histograma

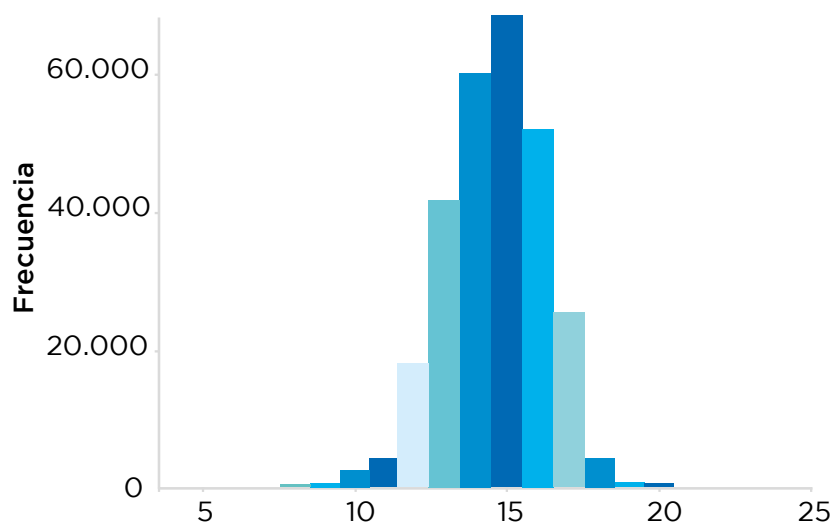
Los histogramas se utilizan para representar una distribución de frecuencias de una variable medida con una escala numérica continua.

Es una representación gráfica análoga al gráfico de barras (en este caso contiguas), pero la información que un histograma comunica es del área de las barras que lo constituyen, y no ya sólo de su altura.

Por eso, si la amplitud de los intervalos no varía, las áreas son proporcionales a las alturas y las frecuencias están representadas tanto por las alturas como por las áreas. Si los intervalos fueran de diferente amplitud, la altura de las barras debe modificarse para conservar el área correcta.

Ejemplo:

Edad de inicio de consumo de marihuana. Escolares 2014.



VIII. Interpretación de los datos

Como ya se mencionó, básicamente el procesamiento estadístico de datos tiene por objeto principal resumir la información obtenida.

Esto contempla (además de la tarea de clasificar y tabular la información), la realización de diferentes medidas de resumen, que en general, “son los valores que identifican o caracterizan a ciertas particularidades de las Distribuciones de Frecuencias”.

Estos valores son usualmente funciones algebraicas de las cifras que toma la variable analizada al ser medida u observada, o bien cumplen ciertas condiciones estipuladas para la misma y deben ser seleccionadas de acuerdo a la escala de medición utilizada para medir la variable.

Sistematización de los Datos

Toda vez que se desee procesar los datos obtenidos, se debe sistematizar el trabajo de alguna manera. Una forma conveniente es la siguiente:

- Clasificar cada una de las variables del estudio.
- Organizar y presentar los datos en el gráfico o tabla que se considere más apropiado.
- Resumir los datos. Para hacerlo adecuadamente se debe tener en cuenta el tipo de variable y analizar en primer lugar los datos de una sola variable por vez (Análisis Univariado). Luego se podrá pasar a analizar la relación entre dos variables (análisis bivariado) o entre tres o más variables: análisis multivariado.

Medidas de resumen

Las medidas de resumen más utilizadas pueden clasificarse de la siguiente forma:

- Medidas de Posición: son aquellas que sugieren o indican el lugar de ubicación o posición en que se encuentra una Distribución de Frecuencias, dentro del campo de definición de la variable.
- Medidas de Dispersión: son aquellas que determinan el grado de

dispersión (o bien la concentración), de una Distribución de Frecuencias.

- **Indicadores:** en forma genérica, puede denominarse así a “todo tipo de información que muestra (o “indica”) la situación en un momento dado, o la evolución o modificaciones que registra una unidad a través del tiempo”.

Medidas de posición

De acuerdo con su definición, todas las Medidas de Posición toman valores que se encuentran dentro del campo de variabilidad de la variable en estudio, y quedan expresadas en la misma unidad de medida que ésta.

Es decir que, si denominamos con MP a cualquier Medida de Posición, deberá cumplirse que: $x_{\text{Mín}} < MP < x_{\text{Máx}}$

Existe una gran variedad de Medidas de Posición, y cada una de ellas mide una particularidad diferente de las Distribuciones de Frecuencias.

De acuerdo con el problema que se enfrente deberán calcularse algunas de ellas (aquellas que tiendan a responder las incógnitas que se planteen). Las principales son:

Modo: es el valor de variable más frecuente o al que le corresponde la mayor frecuencia.

Media aritmética (\bar{X}): es también conocida como promedio. Se calcula como la suma de las observaciones dividida por el número total de observaciones (n). Supongamos que se tienen n datos que notaremos como $X_1, X_2, X_3, \dots, X_n$.

La media aritmética de estos valores se define como la suma de todos ellos, dividida por n y se simboliza con (\bar{X}). Se puede encontrar ese valor realizando el siguiente cálculo:

$$\bar{X} = (X_1 + X_2 + X_3 \dots + X_n) / n$$

Mediana: es la observación que deja la mitad de los datos (ordenados de menor a mayor) a cada lado. Se puede encontrar fácilmente ese valor realizando el siguiente cálculo:

$$\text{Posición de la Md en un conjunto de datos ordenados (de menor a mayor)} = \frac{n + 1}{2}$$

Si el número total de observaciones (n) es impar existirá una única mediana y será el valor que ocupe la posición central. Si el número de observaciones es par, no existe un único valor central y la mediana corresponde a la media de los dos valores centrales.

La mediana es la medida de centralización que se utiliza cuando la variable se mide con una escala numérica u ordinal. Es la medida más apropiada en caso de que la distribución de frecuencia de la variable sea asimétrica, ya que es menos sensible que la media aritmética a valores extremos.

Medidas de Orden: también encontradas en la bibliografía como percentiles, cuantiles o cuantiles, son valores que acumulan hasta sí una cierta cantidad de unidades cuyos valores de variable son menores o iguales que ellos, y son superados (o igualados) por los restantes.

Por ejemplo, el percentil 10 deja por debajo al 10% de los valores observados. La mediana definida como el valor que separa el conjunto de los datos en dos mitades iguales, es también el percentil 50.

De acuerdo con la cantidad de partes en que se divida a la Distribución, el nombre genérico de Cuantiles puede tomar las siguientes formas:

Cuartilos (Q_i): aquellos que dividen a la Distribución en cuartas partes. Se tiene:

Q0: el valor que es superado (o igualado) por la totalidad de los n valores de la Distribución. Naturalmente $Q0 = x_{\text{Mín}}$.

Q1: el valor que supera o iguala a una cuarta parte de la Distribución, y es superado (o igualado) por las tres cuartas partes restantes.

Q2: el valor que supera o iguala a dos cuartas partes de la Distribución, y es superado (o igualado) por las dos cuartas partes restantes. Naturalmente $Q2 = \text{Me}$.

Q3: el valor que supera o iguala a las tres cuartas partes de la Distribución, y es superado (o igualado) por la cuarta parte restante.

Q4: el valor que supera o iguala a la totalidad de los n valores de la Distribución. Naturalmente $Q4 = x_{\text{Máx}}$.

Decilos (D_i): Aquellos valores que dividen a la Distribución en décimas partes. Se tiene por ejemplo:

D1: el valor que supera o iguala a una décima parte de la Distribución, y es superado (o igualado) por las nueve décimas partes restantes.

D7: el valor que supera o iguala a siete décimas partes de la Distribución, y es superado (o igualado) por las tres décimas partes restantes.

Centilos (C_i): Aquellos valores que dividen a la Distribución en centésimas partes. Por ejemplo:

C7: valor que supera o iguala a siete centésimas par-

tes de la Distribución, y es superado (o igualado) por las noventa y tres centésimas partes restantes.

C48: valor que supera o iguala a cuarenta y ocho centésimas partes de la Distribución, y es superado (o igualado) por las cincuenta y dos centésimas partes restantes.

Medidas de dispersión

El concepto de las Medidas de Dispersión, es que a partir del valor cero (cuando no existen diferencias entre los valores de la variable, es decir que no hay dispersión), toman mayores resultados (positivos) cuanto más dispersos se encuentran los valores que toma la variable. La unidad de medida resultante, depende de la forma en que se las define, aunque la mayoría queda expresada en la misma unidad de medida que la variable.

Si se identifica genéricamente a las Medidas de Dispersión con MD, se deberá cumplir que: $0 < MD < xMáx$.

Donde: si $MD = 0$ significa que la Distribución de Frecuencias no tiene dispersión, o sea que todos los valores de la variable de todas las n unidades en estudio, son iguales entre sí.

si $MD = xMáx$ es la máxima dispersión que puede darse en la Distribución de Frecuencias.

Como en el caso de las Medidas de Posición, existe una gran variedad de Medidas de Dispersión, las que deberán ser utilizadas de acuerdo con el problema que se esté enfrentando. Las principales son:

Amplitud (o Rango) (Am): es la diferencia entre los valores máximo y mínimo que toma la variable.

Expresado algebraicamente: $Am = xMáx - xMín$

Es la más simple de las Medidas de Dispersión, y por su simplicidad casi siempre se la calcula.

Sin embargo tiene como desventaja que su cálculo se basa en sólo dos valores que, por ser los extremos, pueden ser muy atípicos. No da información acerca de cómo se dispersan los datos dentro del intervalo limitado por el menor y el mayor valor.

Rango Intercuartílico (RI): para solucionar en parte el inconveniente que presenta el rango, el RI surge como una medida simple que no se basa en los valores extremos, sino calculando el intervalo entre el primer y el tercer cuartil.

Esta medida de dispersión se define como la distancia entre los valores entre los cuales se encuentra el 50% central de los datos.

Expresado algebraicamente: $Q = \frac{1}{2} (Q3 - Q1)$

Es conveniente utilizar el rango intercuartílico como medida de dispersión que complementa la información suministrada por la mediana.

Desvío estándar (S o DE): Es la medida de dispersión más utilizada aunque su significado y su cálculo pueden parecer algo complejos. El DE mide la dispersión de los datos alrededor de la media aritmética. Por lo tanto, su uso es correcto siempre que la media sea la medida de centralización adecuada.

Matemáticamente se lo define como la raíz cuadrada del promedio del cuadrado de los desvíos de la variable respecto a la media aritmética de los valores. Simplemente para tener una idea de cómo se calcula el valor del DE (ya que con cualquier software sencillo se puede conocer su valor) se expresa la definición en fórmulas como:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

El símbolo \sum significa sumatoria en este caso de las diferencias entre cada valor individual y la media (\bar{x}) elevado al cuadrado. Una vez hecho este cálculo, se divide por el número de individuos de la población (n) menos 1 y finalmente se calcula la raíz cuadrada de este resultado.

Coficiente de Variación (CV) (o **Dispersión Relativa**): es la relación entre el Desvío Estándar y el valor absoluto de la Media Aritmética (es decir, el cociente entre la principal Medida de Dispersión y la principal Medida de Posición expresada con signo positivo).

Expresándolo en fórmulas:

$$CV = \frac{\sigma}{|x|}$$

La principal ventaja de esta medida, es que no queda expresada en ninguna unidad de medida, es un número neutro, abstracto. Debido a ello, puede ser utilizado para comparar dispersiones entre distribuciones de diversa índole.

Indicadores

Dentro de los Estudios Estadísticos, lo que es Unidad y lo que es Variable depende de cada problema en particular, en tanto que lo que es "unidad" en un caso, puede ser "variable" en otro (o viceversa). Por ejemplo:

En un caso en el que se analice la cantidad de personas residente en los Partidos del Gran Buenos Aires, la población constituye la “variable” en estudio (las “unidades” serían los Partidos).

Esas mismas personas constituyen las “unidades” de análisis, si lo que se debe hacer es una encuesta de opinión para determinar la orientación del voto en cada Partido del GBA.

En la práctica es muy común que un mismo estudio combine más de una variable o cifras pertenecientes a las mismas unidades, o bien se combinen en diversas relaciones varios conjuntos de unidades, y luego, se confeccionen con ellos, “Indicadores” o “Relaciones”, con las que se condensa la información básica y se enriquece su análisis.

Concepto de Indicador

En forma genérica, puede denominarse así a “todo tipo de información que muestra (o “indica”) la situación en un momento dado, o la evolución o modificaciones que registra una unidad a través del tiempo”. Si bien la información que permite realizar esta descripción puede ser de cualquier tipo, lo usual es que se denomine Indicador a:

“Las relaciones lógicas entre variables o cifras que resumen información diversa en unos pocos guarismos, que se utilizan para mostrar el estado de una unidad de análisis o su evolución a través del tiempo, o la comparación entre varias unidades”.

De este modo, cualquier dato básico que permita describir la situación o evolución de una unidad, puede ser considerado un **Indicador**, por ejemplo:

La Población (cantidad de personas que habitan en una región), de un país, es un **indicador** de su tamaño o importancia con relación a otros países o regiones.

El Producto Bruto Interno (PBI) (suma de los resultados de las actividades económicas de una región), es otro **indicador** de su tamaño o importancia en relación a otros países.

Relacionando a esos datos básicos, se tiene otro Indicador que cumple con la definición específica:

“El PBI por Habitante de un país: es un Indicador del grado de desarrollo o del nivel económico de las personas que lo habitan”

Tipo de Relaciones entre Variables o Cifras

Algunos de los indicadores o relaciones, son coincidentes con medidas o técnicas utilizadas en el análisis de conjuntos de datos de una sola variable (presentados en el apartado “Medidas de resumen”), otros constituyen simplificaciones o modificaciones de otras técnicas más elaboradas.

En general, los Indicadores o Relaciones entre Variables o Cifras que se utilizan con más frecuencia, pueden ser clasificados de la siguiente forma:

- » Proporciones
- » Promedios
- » Índices
- » Tasas
- » Razón

A continuación, se realiza un esbozo conceptual de cada una de ellas.

Proporciones

Son las relaciones o cocientes entre cifras que expresan la importancia de los valores de una variable (o de un conjunto de ellas), en el total del cual forman parte.

Particularmente, las Frecuencias Relativas mencionadas en el apartado sobre las Distribuciones de Frecuencias, constituyen una Proporción, ya que representan la importancia de cada uno de los valores de la variable (a través de su frecuencia absoluta), en relación al total de las unidades consideradas.

En el ejemplo de los 120 pacientes con consumo problemático de cocaína 3 referían haber 5 internaciones. En este caso la proporción es $3/120 = 0,025$ personas (frecuencia relativa).

Lo usual en la práctica, es que en las Proporciones, las relaciones entre la parte y el total se encuentren multiplicadas por algún número redondo (10, 100, 1.000, etc.), a fin de obtener resultados mayores que uno más fáciles de interpretar. Para el caso del ejemplo anterior, $0,025 \cdot 100$ se lee: el 2,5% de los 120 pacientes con consumo problemático de cocaína estuvieron 4 veces internados.

La expresión algebraica con que se las debe calcular es la siguiente:

$$P = (C/T) \cdot N^{\circ}$$

Donde:

C = cifra correspondiente a la cantidad de veces que se repite un valor de la variable.

T = total de unidades consideradas, dentro de las que forma parte C.

N° = 10, 100, 1.000, etc.

Por ejemplo, si en una población de 35.000 habitantes 800 han consumido alguna vez medicamentos para adelgazar, la proporción en esa población es de $800/35000 = 0,02$ o bien del 2 % expresada en porcentaje. El valor de una proporción puede variar así de 0 a 1, mientras que el porcentaje, varía entre 1 y 100.

Promedios

En el caso ya analizado de las Distribuciones de Frecuencias, un promedio (o media aritmética), es la relación o cociente entre la suma de todos los valores que la misma toma y el total de unidades consideradas. En el caso en que se relacionan variables o cifras, se entiende por Promedio a:

“Las relaciones entre las cifras de dos variables medidas sobre una misma unidad o conjunto de unidades”.

Al definírsele de esta forma, un Promedio queda siempre expresado en unidades de medidas compuestas, dependiendo de las unidades de medida de las variables que se relacionan.

La expresión algebraica con que se los debe calcular es la siguiente:

$$PR_i = (CV1_i / C2_i)$$

Donde: $i = 1 - n$ y:

CV1_i = cifra correspondiente a la variable 1 en la unidad i-ésima

CV2_i = cifra correspondiente a la variable 2 en la unidad i-ésima

n = cantidad total de unidades consideradas.

Ejemplos:

Según el Censo de Centros de Tratamientos 2011, en los 131 centros generales de atención que también atienden a personas con problemas de abuso de drogas, existen unos 1.209 profesionales dedicados a este tipo de atención.

Es decir: en **promedio** hay $1209/131 = 9,2$ profesionales por centro.

Los datos aportados por los dispositivos censados revelan un total de 524 camas disponibles para la atención de pacientes para 24 centros de tratamiento en la región patagónica.

Es decir: en **promedio** hay $524/24 = 20$ camas por institución.

Índices (Simples)

Aunque intrínsecamente constituyan el mismo tipo de Indicador, desde el punto de vista de la aplicación práctica, se los puede clasificar en:

- **Índices Cronológicos:** aquellos en los que se relacionan cifras de una variable medidas sobre una misma unidad (o conjunto de ellas), a través del tiempo, tomando un cierto período como comparación o base de referencia.

Ejemplo:

Índice de mano de obra ocupada de una ciudad, comparando la población ocupada según las Encuestas Permanentes de Hogares, con respecto a la registrada en la toma realizada en 2014.

- Índices Sectoriales/Geográficos/etc.: aquellos en los que se relacionan cifras de una misma variable en una mismo período de tiempo pertenecientes a distintas unidades (sectoriales, regionales etc.), tomando una de estas como comparación o base de referencia.

Ejemplo:

Índice de las ventas de alcohol de los municipios de la provincia de Bs As durante el último mes, tomando una de ellos como punto de comparación o referencia.

La expresión algebraica con que se los debe calcular es la siguiente:

$$I_{i/o} = (C_i / C_0) * 100$$

Donde: $i = 1 - n$ y:

I_i = Número índice (Simple), correspondiente al momento (en el caso de los Cronológicos), sector, región, etc. (en el caso de los restantes), i-ésimo, en base o referencia al momento, sector, región, etc. 0 (el que se toma como referencia o punto de comparación).

C_i = cifra correspondiente al momento, sector, región, etc. i-ésimo.

C_0 = cifra correspondiente al momento, sector, región, etc. "0" (el que se toma como base o referencia).

A cualquiera de ambos tipos de Índices se los suele presentar en "porcentajes".

Tasas

También este Indicador puede ser clasificado en dos tipos conceptualmente diferentes:

- Tasas Simples: "Son relaciones entre dos variables medidas sobre una misma unidad, con la particularidad de que el denominador es mayor que el numerador, y por lo tanto para facilitar su lectura y comprensión, su resultado se multiplica por algún número redondo (100, 1.000, etc.).

Técnicamente son Promedios o Proporciones que cumplen las condiciones particulares enunciadas en la definición.

El caso de los Promedios es el caso más frecuente entre las Tasas, y son aquellos cuyos resultados directos dan menor que uno, y al multiplicárselos por algún número redondo quedan expresados "en porcentajes", "por miles", etc.

Ejemplos:

Tasa de Médicos por Habitantes (usualmente se expresa por 10.000).

Tasa de Mortalidad (muertos cada 1.000 habitantes).

También hay Tasas que son confundidas con las Proporciones, como ser:

- Tasa de Desempleo: es el total de Desocupados sobre el total de Población Económicamente Activas (se lo presenta en porcentajes)
- Tasa de Analfabetismo: Cantidad de analfabetos sobre el total de la población en edad escolar (se lo expresa en porcentajes).

La expresión algebraica con que se las debe calcular es la siguiente:

$$Ti = (Yi / Xi) * N^{\circ}$$

Donde: $i = 1 - n$ y:

Ti = Tasa correspondiente al momento o lugar i-ésimo.

Yi = Valor de la variable yi en el momento o lugar i-ésimo.

Xi = Valor de la variable xi en el momento o lugar i-ésimo.

N° = Número “redondo” en el cual quedan expresados los resultados

- Tasas de Incremento o Evolución: “Es la relación entre el incremento producido en una variable entre dos momentos del tiempo, y la cifra correspondiente al momento base o inicial”. Este es un tipo de relación muy particular, que puede ser asociada a los Índices Cronológicos (ya que ambos miden cambios a través del tiempo).

Ejemplos:

Incremento (o disminución) del consumo de marihuana de un grupo poblacional de un año al siguiente (expresado en relación al año anterior).

La expresión algebraica con que se las debe calcular es la siguiente:

$$Ti = [(Vi - Vi-1) / Vi-1] . N^{\circ}$$

Donde:

Vi = Valor de la variable en análisis en el momento i-ésimo.

Vi-1 = Valor de la variable en el momento i-1 éximo.

N° = Número “redondo” en el cual quedan expresados los resultados.

Razón

Esta medida de frecuencia se emplea para variables de tipo nominales.

Es una división que no implica ninguna relación específica entre el numerador y el denominador. Esto quiere decir que el numerador y el denominador llevan unidades diferentes y el numerador no se encuentra incluido en el denominador.

Por ejemplo: si en un grupo de 300 personas con consumo frecuente de cocaína 200 son varones y 100 son mujeres, puede decirse que la razón hombre/mujer ($200/100$) es en este caso de 2 a 1. Es decir: por cada 2 varones consumidores de cocaína hay una mujer en iguales condiciones.

IX. Nociones de estadística inferencial

Por medio de la Estadística Descriptiva, se puede clasificar, organizar y resumir los datos obtenidos, logrando un resumen completo y una visión general del fenómeno que se está investigando. Por otra parte, la Estadística Inferencial, da la posibilidad de que, a partir del estudio de sólo un subgrupo de la población (muestra), se pueda estimar o inferir lo que ocurre en la totalidad de la población. Asimismo, permite también determinar si la presencia de un evento en particular fue simplemente mera casualidad. La Estadística Inferencial permite conocer lo que ocurre en toda una población a partir del estudio de una muestra, logrando una visión integradora de lo que se desea investigar.

Como se expresó en el apartado **Métodos de obtención de datos de fuentes primarias**, se tiene una enumeración completa o también información censal “cuando se observa en forma exhaustiva a todos los elementos de la población en estudio”. En el caso de “observar sólo a una parte de la población”, la información es de carácter parcial o muestral. El hecho de obtener estimaciones sin observar a todas las unidades, es un proceso de tipo inductivo (paso de lo particular a lo general), el cual produce un **error** debido precisamente al procedimiento de trabajar con una muestra en lugar de la población completa. Este tipo de error (aleatorio) es inevitable a la hora de sacar conclusiones para una población a partir de una muestra. Sin embargo, es un **error cuantificable** ya que uno de los objetivos del muestreo estadístico es conocer el grado de incertidumbre que tiene lo que se está diciendo y la precisión de la inferencia será mayor cuanto más pequeño sea el margen de error.

Es decir este **error debido al muestreo** no es posible anular a menos que la muestra sea exhaustiva (es decir que se transforme en censo), no obstante la teoría del muestreo con el aporte de las probabilidades hace posible su medición e inclusive el análisis de las alternativas para reducirlo.

Por otra parte, existe el llamado **error no debido al muestreo** el cual procede de defectos en los instrumentos de medida y de las condiciones en las que se establece la medida, así como de otro conjunto de errores en la transmisión de datos. Entre los errores no estadísticos se encuentra el **sesgo** y se refiere a una desviación sistemática de las observaciones respecto a lo que se está midiendo. En investigación mediante muestras el sesgo se produce al obtener muestra que no se adecúan a la población, es decir, por el desfase que existe entre la población objetivo y la población de la que obtenemos la muestra. Por esa razón este tipo de errores se pueden (y deben) ser evitados.

Población

Una definición específica de Población, ligada con la teoría y práctica de las técnicas del muestreo es la siguiente: “Población es un conjunto de elementos definidos en el tiempo y en el espacio, sobre los cuales se realizarán las observaciones en el caso de una encuesta exhaustiva o censo, o a los cuales se referirán los resultados de la investigación en el caso de un estudio por muestreo”.

Muestra

En toda investigación por muestreo, la muestra se selecciona de la población, con lo cual se la define como: “Subconjunto de unidades seleccionadas de la población definida. En esta recae la realización de las observaciones, mediciones, etc.”.

Parámetros

Las actividades de estimación, comprobación, etc. que se realizan con los resultados de una muestra, tienen por objetivo estudiar características específicas de una población denominadas **Parámetros**, por lo que **se adopta entonces que un Parámetro, es una característica medible de los elementos de una Población, también llamado valor poblacional, que la caracteriza e identifica con relación a otras poblaciones.** Definidos de esta forma, **los Parámetros son constantes**, ya que no se modifican ni varían dentro de una población determinada, y en un momento y espacio fijados.

Dentro de los estudios por muestreo, lo más frecuente es que el interés se centre en los siguientes parámetros de una población:

- **Promedios:** por ej.: ingreso promedio por habitante, promedio de edad de inicio en el consumo de tabaco. (corresponde a la suma total de los valores de una variable dividida por la cantidad de unidades consideradas).
- **Totales:** por ej.: cantidad total de población de una ciudad. Corresponde a la suma total de los valores de una variable a lo largo de todas las unidades de la Población.
- **Proporciones:** por ej.: porcentaje de viviendas con NBI, proporción de jóvenes con consumo problemático de alcohol. (casos favorables a un determinado valor de variable respecto al total de unidades).
- **Cantidad de Casos Favorables:** por ej.: población total en nivel de indigencia, total de establecimientos de salud que ofrecen tratamiento por consumo problemático de sustancias. Corresponde al conteo de casos o unidades que cumplen una cierta condición en

la población. Se la puede analizar como una combinación de Proporciones y Totales.

En realidad, estos no son los únicos Parámetros que identifican o caracterizan poblaciones, ya que existen otros como: la Mediana, el Modo, la Amplitud, el Desvío Medio, etc., y algunos menos frecuentes aún, que aparecen dentro de la Inferencia Estadística (diferencia de dos promedios, cocientes de variancias, etc.).

Al encararse un estudio por Muestreo, lo usual es que los Parámetros (por lo menos algunos de ellos), sean desconocidos, y que justamente uno de los objetivos de la muestra (sino el único o principal), sea estimarlos o comprobar sus valores.

Estimadores

Los parámetros se estiman o comprueban utilizando los denominados Estimadores o “estadísticos”, que pueden ser definidos de la siguiente forma: “Estimador a una expresión matemática, que se construye con los valores correspondientes a los elementos que integran una muestra, y que como su nombre lo indica, sirven para obtener estimaciones del correspondiente valor poblacional o parámetro desconocido que se intenta estimar o verificar”

En definitiva, los Estimadores son características muestrales, calculadas con las observaciones de una muestra aleatoria, y a partir del valor del Estadístico se puede inferir el valor del Parámetro, es decir partiendo de la muestra generalizar a nivel de la población. Por otra parte en muchos casos no sólo se debe estimar el Parámetro propiamente dicho, sino que también se debe calcular una estimación del error debido al muestreo o el error estándar del estimador.

Al estar definidos como fórmulas matemáticas o condiciones que cumplen los resultados de una muestra, los Estimadores deben ser considerados variables, ya que pueden tomar resultados diferentes de muestra en muestra. Este concepto permite remarcar una diferencia sustancial dentro de la Inferencia Estadística: “En el muestreo probabilístico los parámetros deben ser considerados constantes usualmente desconocidas, en cambio los estimadores son siempre variables aleatorias”.

Al describir la ocurrencia de un determinado evento en una población en particular y mediante la cuantificación de las observaciones que se realizan, se puede extraer conclusiones acerca del fenómeno observado. Se puede además, querer comparar los resultados con observaciones acerca del mismo evento en otras poblaciones, o bien en la misma población en diferentes momentos en el tiempo.

Es allí donde se necesita de la estadística como **herramienta fundamental** para la cuantificación de los fenómenos, es decir para expresar numéricamente las observaciones realizadas.

X. Nociones de muestreo

Todos los métodos comprendidos dentro de la Inferencia Estadística, llevan implícitos la realización de muestras aleatorias o probabilísticas, que son aquellas en las que las unidades que integran las muestras son seleccionadas mediante procedimientos en los que interviene el azar. La selección de muestras aleatorias² permite el cálculo de las probabilidades de cada una de ellas, y por extensión las probabilidades de cada uno de los resultados de los estimadores que con ellas se calculan.

Si bien no es objetivo de este manual el entrar en detalles relativos a las técnicas de muestreo, se realizará un bosquejo amplio sobre el Muestreo y las técnicas que incluye considerando que es el área que identifica con mayor precisión la esencia de lo que es la Estadística.

1. Tipos de Muestreo

Los Tipos o Técnicas de Muestreo pueden ser clasificados inicialmente en Aleatorios o Probabilísticos y los No Aleatorios. Si bien estos últimos no permiten la realización de inferencias con los fundamentos teóricos que brinda la aleatoriedad, lo cierto es que se encuentran ampliamente difundidas en ciertos tipos de aplicaciones, por lo se presentan someramente sus principios, bondades y defectos.

Muestreo No Aleatorio: es aquel en el cual las unidades que integran la muestra no se eligen mediante procedimientos aleatorios, y por lo tanto no es posible conocer y asignar a cada unidad una probabilidad de ser seleccionada.

Estos métodos pueden ser clasificados en:

- *Circunstancial, Casual o Fortuito:* quien realiza la elección de las unidades, lo hace de modo de cubrir la mayor cantidad de casos en el menor tiempo posible, y lo hace seleccionando a aquellas unidades que tiene a su alcance y le resultan más accesibles. La muestra se elige entonces por comodidad o circunstancialmente. En los casos de poblaciones homogéneas, este tipo de muestras brinda resultados satisfactorios, aunque los riesgos de sesgos o errores sistemáticos son elevados. Por ejemplo: en arqueología o en historia se obtienen conclusiones a partir de los elementos que se pueden conseguir a mano; algo similar ocurre a veces en las ciencias de la salud y astronomía.

² En realidad, los que son aleatorios son los métodos o procedimientos para obtener las muestras, y no las muestras propiamente dichas.

- **Selección Experta:** es el caso en que se recurre a expertos para seleccionar unidades o grupos que a su juicio sean considerados representativos de la población. Una variante de este método es componer una muestra integrada por todos los respondentes que se suponen calificados.
- **Intencional u Opinable:** en una derivación de la anterior. En este caso los encuestadores o seleccionadores eligen expresamente las unidades que participarán de la muestra, recurriendo a su intención u opinión. En general la elección de las unidades intenta obtener una “muestra representativa” de la población bajo estudio, pero esta representatividad queda sujeta a la subjetividad y preferencias de quien elige.
- **Por Cuotas:** de difundida aplicación en los estudios de mercado o los de opinión. A los encuestadores se los instruye para obtener cuotas preestablecidas, de clases de elementos que presentan alguna característica común. La selección de las unidades dentro de la población, queda a cargo del encuestador, quien únicamente debe respetar los tamaños fijados para las cuotas. Por ejemplo, las cuotas asignadas a los encuestadores pueden consistir en entrevistar un número determinado de propietarios de vehículos de diferentes marcas, por ejemplo: 7 propietarios de Renault; 6 de Ford; 4 de Volkswagen; 5 de Fiat, etc., pero es el encuestador quien elige según su criterio a los propietarios de cada tipo de vehículo hasta agotar la correspondiente cuota.
- **Por Cuotas “cuasi” Probabilístico:** últimamente en las grandes ciudades, se agudizaron los inconvenientes de los encuestadores para entrar en los hogares particulares. Esto trajo como consecuencia cierta popularización del muestreo por cuota, pero cumpliendo algunas condiciones que lo acercan al muestreo probabilístico. Este tipo de muestreo es común en las encuestas de opinión, no obstante no es un modelo probabilístico ya que en la última etapa el encuestador elige a conveniencia, lo que lleva a tomar las respuestas de aquel que está dispuesto a responder.
- **De Poblaciones en Movimiento:** es también un tipo de muestra casual o fortuita. La población está constituida por seres vivos, tales como peces, insectos, animales salvajes, etc., que se capturan para su análisis y sacar conclusiones y se los vuelve a dejar en libertad con alguna marca identificatoria así en caso de recaptura se obtendrían datos con relación a su movilidad, cambios experimentados, etc. Este método es conocido como “Muestreo de Captura y Recaptura”. Podría ser Probabilístico o “cuasi”, si las zonas de “contacto” se seleccionaran al azar.
- **Grupos Focales:** se define como “Grupo Focal” a un grupo de personas que son reunidas en un mismo lugar o espacio físico y que poseen conocimientos para informar sobre un tema especificado a requerimiento de un conductor o coordinador que hace preguntas y obtiene las respuestas individuales o en conjunto, usualmente a través de un diálogo. Si bien puede concebirse que el conjunto de personas que conforman el grupo podrían ser elegidas aleatoria-

mente, en la práctica es un caso raro que lo sea.

- *Otros*: existen algunas otras formas de muestreo no probabilístico, pero todas tienen en común, que dependen del criterio individual de una o más personas, o de lo que se pueda obtener, contactar o capturar. En condiciones adecuadas, cualquiera de estos métodos podría ser útil, sin embargo, no es lícito medir el error de muestreo de las estimaciones.

La única forma de evaluar una muestra no probabilística, es comparándola con una situación en que los resultados sean conocidos. Esto resulta raro en la práctica, pues si se dispone de los mismos no tiene fundamento realizar la muestra.

Muestreo Aleatorio o Probabilístico: como ya se lo adelantó, es aquel en el cual las unidades que integran la muestra se eligen mediante procedimientos aleatorios, y por lo tanto es posible conocer y asignar a cada unidad de la población una probabilidad de ser seleccionada. La operación de selección aleatoria es indispensable en el muestreo probabilístico. Consiste en obtener “n” números aleatorios del total “N”, donde cada uno identifica a una unidad de muestreo seleccionada, y su conjunto constituye la muestra probabilística.

El sistema de selección de las unidades de la población puede ser cualquiera que garantice que todas y cada una de ellas tenga una probabilidad conocida y distinta de cero de participar de la muestra. En la práctica, esta probabilidad es asignada a cada unidad de muestreo en forma automática, a través de la selección aleatoria.

Por otra parte **sólo** si la muestra es probabilística se justifica medir el error debido al muestreo de los estimadores.

Técnicas Muestrales

Dependiendo de las características de la población y del conocimiento que se tenga de ellas, existen diversas Técnicas muestrales que se pueden aplicar, existiendo para cada una de ellas extensos desarrollos teóricos que fundamentan y permiten el cálculo de los estimadores y de sus dispersiones.

Sin pretender entrar en mayores detalles, pueden enunciarse las siguientes **Técnicas de Muestreo**:

- *Muestreo Aleatorio Simple* (MAS) o *Muestreo Simple al Azar*: como su nombre lo sugiere, es la técnica de muestreo más simple y para su aplicación se requiere tener identificadas a todas y cada una de las N unidades que integran la población (por ej.: con un número individual y unívoco), y aplicando algún sistema aleatorio de selección, elegir a partir de ellos cada una de las n unidades que integrarán la muestra. Si bien lo usual en el MAS es que todas las unidades de la población tengan la misma probabilidad de ser elegidas en la muestra, esta no es una condición indispensable, aunque de ser

así se facilitan las fórmulas que se deben aplicar para realizar las estimaciones y medir sus dispersiones.

- **Muestreo Sistemático** (MS): con esta técnica de selección, se debe identificar (por ej.: numerar), una parte de la población (las primeras k unidades), elegir de ellas una al azar por MAS, y a partir de la elegida seleccionar las $(n - 1)$ restantes en forma sistemática, una cada k unidades [es decir, saltar $(k - 1)$ unidades y tomar para la muestra a la k -ésima]. El MS se muestra más eficiente que el MAS en el caso que previo a la selección las unidades se encuentran o son ordenadas en base a algún aspecto o variable que tiene que ver con la causa del muestreo. También en este caso las unidades suelen tener la misma probabilidad de selección, aunque de ser necesario es posible asignar probabilidades diferentes a cada una (por ej.: probabilidades proporcionales al tamaño), aspecto que debe ser tenido en cuenta al momento de realizar las estimaciones (cada unidad de la muestra debe ser ponderada por la inversa de su probabilidad de selección).
- **Muestreo Replicado**: La muestra total está compuesta por un número de submuestras elegidas de la misma población. Por ejemplo por este método en lugar de seleccionar una única muestra de $n = 500$, seleccionaríamos cinco (u otra cantidad) muestras independientes de 100 elementos cada una de la misma población de “N”.
- **Muestreo Estratificado** (ME): consiste en subdividir a la población en partes o “estratos”, dentro de cada uno de los cuales las unidades deben ser lo más parecidas posibles (en cuanto al aspecto o variable por las que se muestrean), pero entre ellos las unidades contenidas deben ser lo más diferentes posible. La técnica prevé que dentro de cada estrato se obtenga una parte de la muestra total, haciéndoselo mediante MAS o MS. De construirse bien los estratos, usualmente el ME logra mejores resultados que el MAS o el MS. Su principal desventaja es el conocimiento que se debe tener de la población, y su acondicionamiento para que se puedan extraer muestras independientes de cada estrato. Para la determinación de la muestra a extraer dentro de cada estrato, se puede recurrir a diversos criterios alternativos: Igual tamaño de muestra en cada estrato, Proporcional al tamaño de cada uno en la población, Asignación Óptima (proporcional a la dispersión que presentan las unidades dentro de cada estrato), o Asignación por Costo (inversamente proporcional al costo de seleccionar cada unidad en cada estrato).
- **Muestreo por Conglomerados**: esta técnica requiere disponer la población en grupos o partes que cumplen características totalmente inversas a los estratos. En este caso, los “conglomerados” deben contener dentro de sí a todo tipo de unidades, y representar cada uno un fiel reflejo de la población total; y en ese caso todos los conglomerados deberían ser parecidos entre sí en lo que concierne a su constitución o estructura. Cumplido ese requisito, la técnica consiste en elegir uno o varios de esos conglomerados, para constituir con todas sus unidades la muestra a observar. De-

bido a que uno de los campos de aplicación más frecuente es en los muestreos de zonas geográficas, se lo conoce también como “Muestreo por Áreas”.

- *Paneles:* se define como “Panel” a un grupo de unidades (personas, hogares, etc.), que son seguidas a través del tiempo para medirles un conjunto de variables similares en cada una de las tomas o muestras, y analizar su evolución. En realidad los “Paneles” no constituyen una técnica de muestreo diferente, ya que para constituir la muestra de unidades que se analizarán cronológicamente, puede recurrirse a cualquiera de las otras técnicas conocidas. Representa más bien una forma de análisis de la población, que contiene dentro de sí una serie de particularidades que le han dado un desarrollo teórico especial, para por ej.: determinar la forma de tratar a las unidades que “desaparecen” con el correr de las tomas, los cambios que esas unidades van teniendo en el tiempo, etc. (reemplazarlas o no).

Bibliografía

Los contenidos que aquí presentados pretenden constituir un material de consulta útil aunque no es suficiente si se desea o necesita profundizar los temas estadísticos de interés. Se recomienda por ello, la lectura y comprensión de bibliografía especializada. Por otro lado para la elaboración de este manual se ha tenido en cuenta y/o adaptado principalmente la siguiente bibliografía:

- Aguirre, C.; Niño, F.; Simonetti, E. (2005). Estadística Aplicada a las Ciencias Sociales y Humanas. Posadas: Ed. Univ. Nac. de Misiones.
- Hildebrand, D. & Lyman Ott, R. (1997). Estadística Aplicada a la Administración y la Economía. Estados Unidos: Addison Wesley.
- Kelmansky, D. (2009) Estadística para Todos. Buenos Aires: Instituto Nacional de Educación Tecnológica - Ministerio de Educación de la Nación.
- Cappelletti, C. (1983). Cesarini Elementos de Estadística. Buenos Aires: Cesarini Hnos.

Dirección del Observatorio
Argentino de Drogas



SEDRONAR
Presidencia de la Nación