SUMA Y SIGUE

MATEMÁTICA EN LÍNEA

MATERIAL PEDAGÓGICO COMPLEMENTARIO

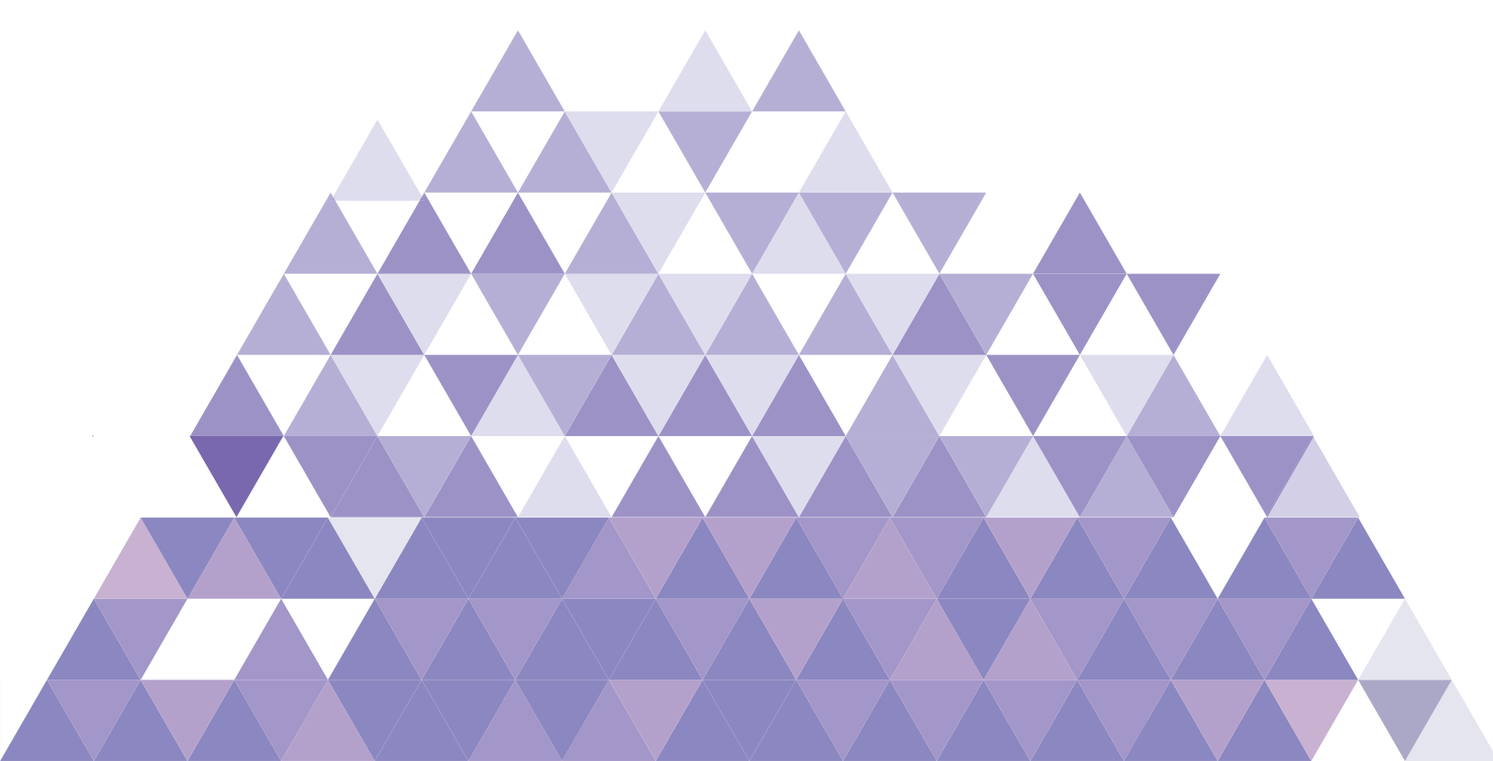
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

CMM Centro de Modelamiento Matemático

Ministerio de Educación Gobierno de Chile

MATERIAL PEDAGÓGICO COMPLEMENTARIO

FICHAS TALLER 2:
RECOLECCIÓN Y ORGANIZACIÓN DE DATOS



INTRODUCCIÓN

El foco de este taller fue identificar los aspectos que caracterizan los procesos de recolección, registro y organización de los datos. Para ello, se propuso un problema de investigación que planteó la necesidad de estudiar cierta característica de interés de la población a través de los resultados de una muestra. El problema permitió abordar las características deseables de una muestra, los métodos de muestreo, los sesgos que se pueden generar en la recolección y registro de los datos, la importancia de la limpieza o depuración de ellos, y la necesidad de organizarlos convenientemente para su análisis.

Los temas abordados en las fichas son los siguientes:

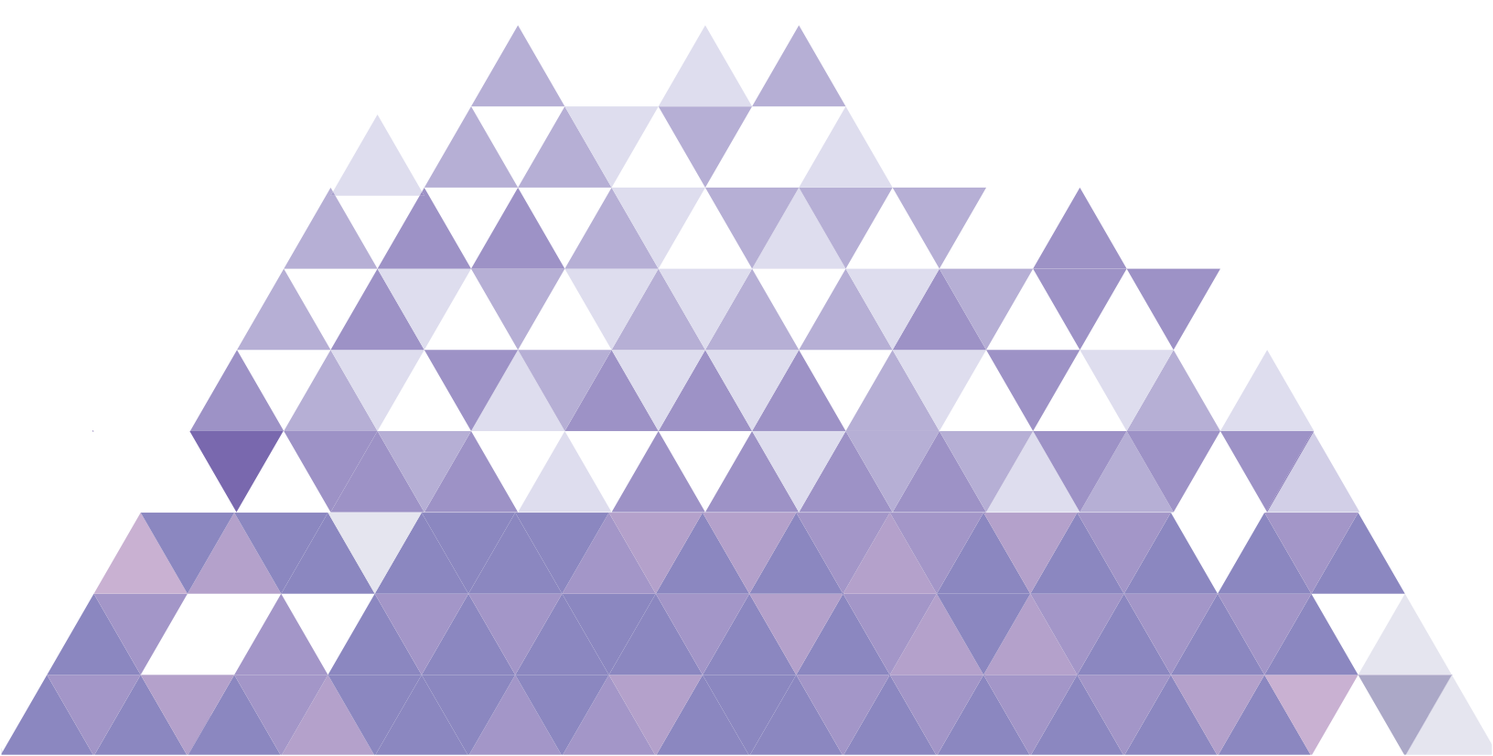
- Clasificación de variables estadísticas: cualitativas y cuantitativas

- Censo

- Muestra:
    - Definición
    - Características de una muestra: representativa, no sesgada
    - Métodos de muestreo
    - Tamaño de la muestra

- Recolección de datos:
    - Sesgos en la recolección y registro
    - Limpieza y depuración
    - Base de datos

- Organización de datos:
    - Manipulación de los datos
    - Rango
    - Tablas de frecuencias



1. Clasificación de las variables estadísticas

La variables estadísticas se pueden clasificar en dos tipos, de acuerdo al valor que ellas toman:

Ejemplos:

1. Color de ojos: variable cualitativa nominal.
2. Nivel de escolaridad: variable cualitativa ordinal.
3. Número de años de escolaridad: variable cuantitativa.

Comentarios

- Una variable no se puede identificar como cuantitativa solo porque sus valores están expresados a través de números. En efecto, cuando en una variable se utilizan números solo como etiquetas para identificar las categorías, pero no se tiene la posibilidad de operar con ellos, estamos en presencia de una variable cualitativa. Esto se observa, por ejemplo, en la denominación de las regiones del país o en los números de las camisetas de los jugadores de un equipo de fútbol.

Ubicación: Módulo 1

Taller: Recolección y organización de datos.
Actividad: ¿Cómo armamos la selección de básquetbol?

2. Censo

Un estudio estadístico en el que se recolectan los datos de la totalidad de la población se denomina **censo**. Algunos ejemplos de censos son:

- El censo de población y vivienda, que tiene como finalidad caracterizar a los habitantes de un país y el lugar donde habitan.
- Pruebas del SIMCE, en las que la población corresponde a todos los estudiantes de determinados niveles.

Comentarios

- Los censos pueden ser tanto de personas como de viviendas, ganado, celulares, o cualquier otro conjunto de individuos u objetos definidos como población de interés.
- El propósito del análisis estadístico es conocer lo que ocurre con determinadas características de la población. Si es posible asumir el costo que involucra tomar datos a cada miembro de la población, la mejor opción es un censo.
- En la práctica, para poblaciones grandes, puede resultar muy difícil obtener datos de la totalidad de la población, por lo que es usual también considerar como censo a estudios que abarcan un porcentaje cercano al 100% de los datos de la población.

Ubicación: Módulo 1

Taller: Recolección y organización de datos.
Actividad: ¿Cómo armamos la selección de básquetbol?

3. Muestra

Cuando se está interesado en determinar ciertas características de una población y resulta imposible o demasiado costoso recolectar la totalidad de los datos, se opta por estudiar una parte o subconjunto de ella, denominada **muestra**, cuyos resultados luego se generalizan a la población.

Al trabajar con muestras es deseable que estas sean:

- **Representativas**: se busca que las muestras representen a la población, esto es, que la forma en que se distribuye la característica de interés en la población y en la muestra sean similares, de modo que los resultados obtenidos a partir de la muestra sean extrapolables a la población.
- **No sesgadas**: se desea que la muestra sea representativa, es decir, que la distribución de la característica de interés en la muestra se asemeje a la de la población. Luego, en la elección de los elementos que componen la muestra no deben intervenir criterios que afecten esto.

Comentarios

- Dado que, en general, la distribución de la característica de interés de la población es desconocida, resulta complejo evaluar la representatividad de una muestra. Sin embargo, existen métodos que ayudan a eliminar posibles sesgos en su elección y otorgan de esa manera mayor validez a las inferencias que se hacen de la población basadas en los resultados de la muestra.

Ubicación: Módulo 1

Taller: Recolección y organización de datos.
Actividad: Escogiendo la muestra.

4. Métodos de muestreo

Los **métodos de muestreo** son técnicas que permiten seleccionar los elementos que conforman las muestras. Se dividen en dos grandes grupos, dependiendo de si los elementos de la muestra se seleccionan o no al azar:

- **Muestreos aleatorios**: el más utilizado es el **muestreo aleatorio simple**, en que cada elemento se elige al azar, cuidando de que todos tengan las mismas posibilidades de ser seleccionados. Por ejemplo, seleccionar una muestra de niños sorteando los nombres en una tómbola.
- **Muestreos no aleatorios**: son útiles cuando solo se dispone de un subgrupo específico de la población. Por ejemplo, una muestra compuesta solo por los que desean participar en una encuesta.

Comentarios

- Un muestreo aleatorio ayuda a evitar sesgos, mientras que en un muestreo no aleatorio es casi seguro que la muestra resultará sesgada. Por ello, las técnicas de muestreo aleatorio resultan más útiles cuando se desea extraer conclusiones sobre la población a través de una muestra.
- Existen herramientas computacionales que permiten seleccionar aleatoriamente los elementos de una muestra, lo que resulta práctico para poblaciones grandes.

Ubicación: Módulo 1

Taller: Recolección y organización de datos.
Actividad: Escogiendo la muestra.

5. Tamaño de la muestra

La posibilidad de hacer inferencias válidas para la población a partir de una muestra depende de su tamaño.

Se debe tener en cuenta que si tomamos distintas muestras del mismo tamaño, habrá variabilidad entre ellas. Esa variabilidad:

- es mayor para muestras pequeñas con respecto al tamaño de la población, por lo que la información que se extrae de ellas es variable, lo que afecta la validez de las inferencias que se hacen para la población.
- se reduce en la medida en que el tamaño de las muestras se acerca al de la población, por lo que muestras de mayor tamaño entregan mejor información.

En la práctica, existen consideraciones operativas y de costos que limitan el tamaño de la muestra que se puede seleccionar, por lo que es necesario encontrar un balance entre una muestra suficientemente grande que permita representatividad y que a su vez sea factible de estudiar.

Comentarios

- Existen técnicas estadísticas que hacen posible estimar el tamaño adecuado para una muestra, dependiendo del tamaño de la población y del grado de confiabilidad que se espera que tengan los resultados.
- La representatividad de una muestra depende tanto de su tamaño como de la forma en que se elige. Se deben utilizar procedimientos que eviten sesgos en la muestra, y al mismo tiempo establecer un tamaño que permita tener un grado de confiabilidad aceptable en los resultados que se obtengan a partir de la muestra.

Ubicación: Módulo 1

Taller: Recolección y organización de datos.
Actividad: Escogiendo la muestra.

6. Sesgos en la recolección y registro de los datos

Para procurar validez en los resultados de una investigación estadística es necesario revisar y evitar posibles sesgos en el proceso de recolección de datos. Algunos de estos sesgos ocurren cuando:

- se generan errores asociados al instrumento que se utiliza para recolectar los datos o al manejo de dicho instrumento.
- las condiciones en que se realiza la recolección no son adecuadas.
- las condiciones en que se recolectan los datos, cuando esto se hace en distintas instancias o momentos, no son similares.

Es importante tratar de anticiparse a la ocurrencia de estos errores revisando la exactitud de los instrumentos y estableciendo procedimientos claros para la recolección de los datos.

Comentarios

- En ocasiones, los datos que componen una muestra se deben recolectar en distintos momentos o lugares. En estos casos es importante generar condiciones similares en la recolección de los datos para que la variabilidad que existe entre las distintas submuestras obtenidas de ese modo no se incremente por factores ajenos a la naturaleza aleatoria de los datos.

Ubicación: Módulo 1

Taller: Recolección y organización de datos.
Actividad: Llegó el momento de aplicar el test.

7. Limpieza o depuración de los datos

Los errores en la recolección o registro de los datos generan sesgos en la muestra que pueden afectar la validez de los resultados. Para corregir estos errores, es necesario realizar un proceso de **limpieza** o **depuración** de datos que consiste en:

1. identificar los registros incorrectos, inexactos, no pertinentes o incompletos.
2. evaluar la posibilidad de sustituir, rectificar o excluir del análisis los datos que presentan problemas.

Comentarios

- Aun cuando se considere que la recolección de los datos fue realizada de forma correcta, es común que, al revisar los datos, aparezcan errores. Esta revisión es necesaria para que los resultados sean lo más confiables posible.

Ubicación: Módulo 1

Taller: Recolección y organización de datos.
Actividad: Llegó el momento de aplicar el test.

8. Base de datos

Una **base de datos** es un conjunto de datos, pertenecientes a un mismo contexto, que se almacenan con la intención de acceder a ellos cuando se requiera. Por lo general, una base de datos está estructurada de manera de facilitar la consulta y el tratamiento de los datos.

Ejemplos de bases de datos:

1. Portal de datos públicos (http://datos.gob.cl/), que contiene bases de datos con la información pública del gobierno de Chile en diversos ámbitos: finanzas, salud, medio ambiente, etc.

2. Datos de libre acceso del Banco Mundial (http://datos.bancomundial.org/), que contiene bases de datos relacionados con indicadores de desarrollo, tales como género, pobreza, cambio climático, desarrollo social, etc.

Comentarios

- Hoy en día existe una gran cantidad de bases de datos, disponibles en plataformas virtuales, que son una fuente importante de información para investigaciones de todo tipo. También constituyen un recurso muy útil para la enseñanza de la estadística, ya que entregan al docente una fuente de fácil acceso a una gran cantidad de datos reales que puede utilizar para que sus estudiantes realicen análisis estadísticos.

Ubicación: Módulo 1

Taller: Recolección y organización de datos.
Actividad: Llegó el momento de aplicar el test.

9. Organización de los datos

Para organizar los datos en tablas y gráficos se requiere ordenarlos y contarlos, procesos que pueden resultar complejos cuando se trabaja con muchos datos.

Afortunadamente, existen herramientas computacionales que pueden ayudar a manipular gran cantidad de datos, como las planillas de cálculo, las que presentan funciones tales como *ordenar* o *contar.si,* que facilitan estos procesos.

Comentarios

- Utilizar herramientas computacionales para manipular datos permite evitar, identificar y corregir errores.

Ubicación: Módulo 1

Taller: Recolección y organización de datos.
Actividad: ¿Y qué hacemos con los datos?

10. Rango de los datos

Entenderemos por **rango** de los datos al conjunto que contiene todos los posibles valores de una variable cuantitativa, que van desde el mínimo al máximo observado. Por ejemplo:

- Si se está midiendo la altura de los hombres mayores de 18 años, la altura máxima observada fue 193 cm y la altura mínima 150 cm, entonces el rango es el intervalo de valores que va de 150 a 193 cm.
- Si se está observando cuántos hijos tiene un grupo de hombres antes de llegar a los 30 años, siendo la máxima cantidad 3 y la mínima 0, el rango sería {0, 1, 2, 3}.

En ocasiones, el **rango** también se define como la diferencia entre el máximo y el mínimo valor observado.

Comentarios

- La determinación del rango es relevante para establecer la manera en que se organizan los datos. El rango permite, por ejemplo, definir las categorías que se registran en la tabla de frecuencias.

Ubicación: Módulo 1

Taller: Recolección y organización de datos.
Actividad: ¿Y qué hacemos con los datos?

11. Tabla de frecuencias: absolutas y acumuladas

Una **tabla de frecuencias** es un tipo de tabla que permite observar el número de veces que se presenta cada categoría en los datos. Para cada categoría se pueden registrar las frecuencias que se describen a continuación:

- **Frecuencia absoluta**: número de observaciones de la categoría.
- **Frecuencia relativa**: razón entre el número de observaciones de la categoría y el total de datos. Se calcula como:

$$\textit{Frecuencia relativa} = \frac{\textit{Frecuencia absoluta}}{\textit{total de observaciones}}$$

- **Frecuencia relativa porcentual**: porcentaje de observaciones en la categoría respecto del total de datos. Se calcula como:

$$\textit{Frecuencia relativa porcentual} = \frac{\textit{Frecuencia absoluta}}{\textit{total de observaciones}} \cdot 100$$

Ejemplo: Tabla de frecuencias de los puntajes que obtuvieron 20 estudiantes en un test.

Puntaje	Frecuencia absoluta	Frecuencia acumulada	Frecuencia relativa porcentual
1	2	0,1	10%
2	4	0,2	20%
3	5	0,25	25%
4	8	0,4	40%
5	1	0,05	5%
Total	20	1	100%

$\frac{4}{20} = 0{,}2$

$\left(\frac{5}{20}\right) \cdot 100 = 25$

Adicionalmente, es posible agregar a esta tabla la **frecuencias acumuladas**, ya sean absolutas, relativas o relativas porcentuales, las que describen el número, razón o porcentaje, respectivamente, de observaciones que hay desde la primera categoría hasta la categoría correspondiente.

Comentarios

- Las tablas de frecuencias permiten describir la distribución de los datos y dar cuenta de su variabilidad al entregar información respecto de los valores de los datos y la frecuencia con que estos se repiten.
- En el caso de variables cualitativas nominales, en las que no hay un orden en las categorías, las frecuencias acumuladas carecen de sentido.

Ubicación: Módulo 1

Taller: Recolección y organización de datos.
Actividad: ¿Y qué hacemos con los datos?

12. Conclusiones basadas en inferencias a partir de muestras

El propósito de analizar los datos de una muestra es hacer inferencias respecto de la población en estudio. Las inferencias que se realizan a partir de un análisis descriptivo de los datos es un proceso inductivo, en que el usuario extiende a toda la población las propiedades que se observan en los datos.

Dichas inferencias son solo estimaciones de lo que ocurre en la población, puesto que los resultados están sujetos a la variabilidad de las muestras.

El grado de validez de las conclusiones que se obtienen a partir de una muestra dependen de la ausencia de sesgos en el proceso de selección de la muestra y de la recolección de los datos.

Comentarios

- Es importante tener en cuenta que no es posible asegurar que las conclusiones obtenidas a partir de la muestra respecto de la población sean 100% fidedignas, aun cuando la muestra se haya escogido bajo criterios que minimicen el sesgo.

Ubicación: Módulo 1

Taller: Recolección y organización de datos.
Actividad: El criterio para preseleccionar al equipo de básquetbol.