Notas de clase

Proceso de Recolección de datos

Este material está sujeto a correcciones, comentarios y demostraciones adicionales durante el dictado de las clases, no se recomienda su uso a aquellos alumnos que no concurran a las mismas.

Prof. Nora Arnesi

Estadística y Análisis de Datos

Estadística es la ciencia de recolectar, analizar y sacar conclusiones a partir de un conjunto de datos

Recolección / Fuente de datos apropiada

Organizar y resumir la información

- ★ Tablas
- ★ Gráficos
- ★ Medidas resumen

Estadística Descriptiva

Sacar conclusiones o tomar decisiones a partir de los datos: generalizar los resultados

Estadística Inferencial

Se denomina **población** de interés al conjunto completo de individuos u objetos acerca del cual se desea obtener información. Una **muestra** es un subconjunto de la población seleccionada de una determinada manera.

Cuando se generalizan los resultados de una muestra a una población se corre el riesgo de realizar una conclusión incorrecta debido a que dicha conclusión se basará en información incompleta.

Un aspecto importante en el desarrollo de técnicas inferenciales es la cuantificación de la probabilidad de realizar una conclusión incorrecta.

Ejemplo

Un artículo reportó que investigadores en un hospital en Italia compararon los niveles promedio de colesterol para una muestra de 331 pacientes que habían sido admitidos al hospital luego de un intento de suicidio y habían sido diagnosticados con depresión clínica, con el colesterol promedio de 331 pacientes admitidos al hospital por otras razones.

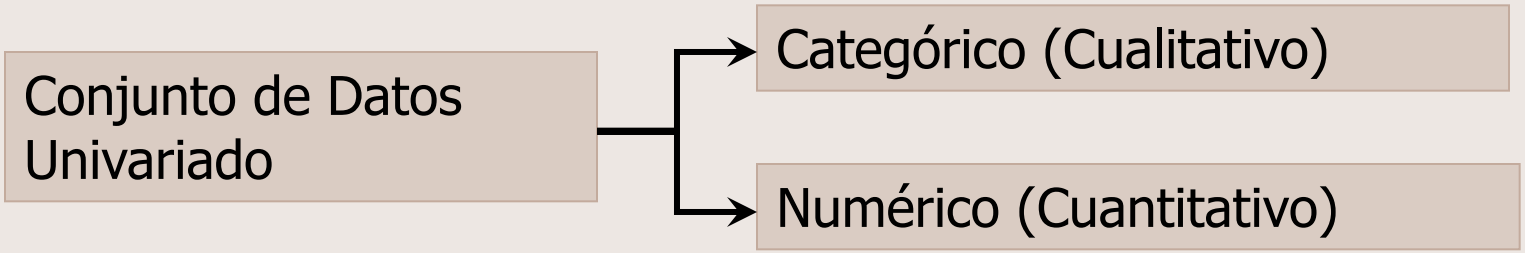
Se utilizaron técnicas estadísticas para analizar el conjunto de datos y para mostrar que el nivel de colesterol promedio fue menor en el grupo de pacientes con depresión.

El artículo señaló, correctamente, que debido a la forma en la cual los datos habían sido recolectados no era posible determinar a partir del análisis estadístico solamente si existía una relación causal entre el nivel de colesterol y el estado psicologico.

Tipos de Datos

Una **variable** es una característica cuyo valor puede cambiar de un individuo u objeto a otro.

Un conjunto de datos que consiste de observaciones de una única variable constituye un conjunto de datos **univariado**

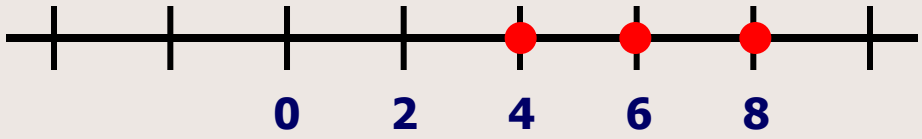


Conjunto de datos **bivariado**: cuando se estudian simultáneamente dos atributos. Ejemplo: peso y altura.

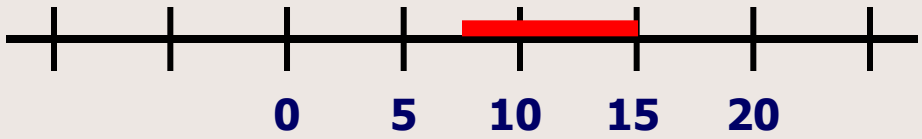
Datos **multivariados**: más de dos variables simultaneamente: peso, altura, frecuencia cardíaca y estado civil.

Variables cuantitativas

discretas: los posibles valores de la variable son puntos separados sobre la recta de números Reales.



continuas: el conjunto de posible valores forma un intervalo sobre la recta de números Reales.



Los datos discretos generalmente aparecen cuando cada observación se determina a través de un proceso de conteo (número de hijos, número de pétalos de una flor, etc.)

El Proceso del Análisis de Datos

- ✓Comprender la naturaleza del problema
- ✓Decidir qué medir y cómo medirlo
- ✓Recolección de los datos
- ✓Resumen de los datos y análisis preliminar
- ✓Análisis de datos formal
- ✓Interpretación de los resultados

Evaluación de un Estudio

Los pasos anteriores pueden ser utilizados como una guía para evaluar estudios publicados:

Qué se intentaba aprender / averiguar? Qué interrogantes motivaron la investigación?

Se recolectó información relevante? Se midió lo que correspondía?

Se recolectaron los datos en forma adecuada?

Se resumieron los datos en forma apropiada?

Se empleó un método de análisis apropiado para el tipo de datos y la forma de recopilar los datos?

Las conclusiones obtenidas, son avaladas por los resultados del análisis de datos?

Ejemplo

Un artículo aparecido en un diario reporta sobre un estudio de la efectividad de una vacuna contra la gripe que se administra por las vías nasales (spray) en lugar de una inyección. El artículo establece lo siguiente:

> Los investigadores administraron el spray a 1070 niños sanos con edades entre 15 meses y 6 años antes del comienzo de la temporada de gripe hace dos inviernos. Uno por ciento (1%) desarrolló gripe (confirmada) comparado con el 18% de 532 niños que recibieron un placebo. Solamente un niño vacunado desarrolló una infección en el oido luego de contrar gripe… Típicamente 30 a 40% de los niños con gripe desarrollan luego una infección en el oido.

Los investigadores concluyeron que la vacuna nasal para la gripe era efectiva en reducir la incidencia de gripe y también en reducir el número de niños con gripe que luego desarrollan infección en el oido.

....

Para evaluar completamente este estudio sería necesario conocer:

Cómo se seleccionaron los niños que participaron en el estudio.

Cómo se determinó si un niño recibía la vacuna o el placebo.

Cómo fueron hechos los diagnósticos subsecuentes de gripe e infección.

Observación y Experimentación

Cuando se recoge información es importante tener presente cuales son las preguntas que se quieren responder en base a los datos recolectados.

En ciertas ocasiones interesa responder preguntas acerca de características de una población o comparar dos o más poblaciones bien definidas.

Para tal fin se selecciona una muestra de cada una de las poblaciones en estudio y se utiliza la información muestral para aumentar el conocimiento que se tiene de la característica en la población(es).

✓Un ecologista está interesado en estimar el grosor promedio de la cáscara de huevos de una determinada especie de ave.

✓Un sociólogo estudia una comunidad rural para determinar si existe una relación entre género y la actitud hacia el aborto.

✓ Un funcionario público desea evaluar si personas de diferentes grupos etáreos difieren con respecto a su apoyo a distintos proyectos de gobierno en una ciudad.

Estudios Observacionales

Un estudio es **observacional** si el investigador observa características de un subconjunto de individuos de una o más poblaciones.

Generalmente, el objetivo de un estudio observacional es obtener conclusiones acerca de la población correspondiente o acerca de las diferencias entre 2 o más poblaciones.

En un estudio observacional es importante obtener una muestra que sea representativa de la población correspondiente, en relación a la o las características estudiadas. (Se debe considerar cuidadosamente la forma en que se selecciona la muestra).

Estudios Experimentales

Un estudio es un **experimento** si el investigador observa como una variable respuesta se comporta cuando se manipulan uno o más factores.

El objetivo de un experimento es determinar el efecto de la manipulación de uno o más factores sobre una variable respuesta.

✓ Un profesor puede investigar que sucedería con las calificaciones en un curso sobre química si el tiempo asignado a laboratorio aumentara de 3 horas a 6 horas por semana.

El tipo de conclusiones que se pueden obtener de un estudio depende de su diseño. Ambos tipos de diseño pueden ser utilizados para comparar grupos, pero en un experimento el investigador controla la asignación a cada grupo.

Observacionales vs. Experimentales

Un experimento bien diseñado puede resultar en datos que provean evidencia suficiente para probar una relación causa-efecto.

Con un estudio observacional es imposible probar una relación causa-efecto ya que no se puede descartar la posibilidad de que el efecto observado se deba a alguna otra variable en lugar de al factor en estudio.

Una variable confundente (o de confusión) es una que está relacionada con la pertenencia a cada grupo y también con la variable respuesta de interés en un estudio.

Ejemplo

Un estudio sobre el hábito de rezar y la presión sanguinea, reporta que 2391 personas con edad mayor o igual a 65 años fueron seguidas durante 6 años.

El estudio reporta que personas que concurrieron a un servicio religioso al menos una vez por semana eran menos probable que tuvieran presión sanguinea alta.

En este ejemplo, se comparan dos grupos pero el investigador no manipula la pertenencia a estos grupos. Como consecuencia no es posible concluir que una relación causa-efecto exista.

Es posible que alguna otra variable, tal como el estilo de vida, se relacione con la concurrencia a la iglesia y con la presión sanguinea. En este caso, estilo de vida es un potencial factor confundente.

Muestreo

Razones para seleccionar una muestra en lugar de obtener información para la población entera:

Recursos: tiempo / dinero

Proceso de medición es destructivo

En un estudio observacional generalmente se trata de generalizar a partir de una muestra a la población correspondiente.

Es importante que la muestra sea **representativa** de la Población.

Generalmente muestras tomadas de determinada manera por cuestiones de conveniencia no resultan representativas.

Sesgos de Muestreo: Sesgo de Seleción

Sesgo es la tendencia de una muestra a difereir de la población correspondiente en una forma sistemática.

Sesgo de selección: se introduce cuando alguna parte de la población es excluida sistemáticamente de la muestra.

✓ Un investigador puede desear generalizar los resultados de un estudio a toda una ciudad pero el método de selección excluye a las personas sin hogar o a las personas sin teléfono.

Si las personas excluidas del estudio difieren en una manera sistemática de las personas incluidas la muestra no será representativa.

También es muy posible que exista sesgo de selección cuando las personas se auto-seleccionan para participar (voluntarios).

Sesgos de Muestreo: Sesgo de Respuesta

Sesgo de respuesta o error de medición ocurre cuando el método de observación tiende a producir valores que difieren sistemáticamente del verdadero valor poblacional en alguna manera.

Errores de medición ocurren cuando un instrumento está mal calibrado o cuando las preguntas en una encuestas están redactadas en una forma que influye la respuesta.

Se estima que los pañales descartables representan menos del 2% de la basura encontrada en los basureros. En contraste, los envases de bebidas representan alrededor del 12% de la basura. Dado lo anterior, en su opinión, considera que sería correcto penalizar o prohibir el uso de pañales descartables?

Sesgos de Muestreo: Sesgo de No Respuesta

Sesgo de no respuesta ocurre cuando las respuestas no son obtenidas para todos los individuos seleccionados para ser incluidos en la muestra.

La no respuesta puede distorsionar los resultados si los individuos que responden difieren de los individuos que no responden.

La no respuesta en estudios o encuestas de opinión varía dramáticamente dependiendo de la forma en que se recoge la información: por correo, telefónicamente o personalmente.

Es importante reconocer que los sesgos se generan por la forma en que la muestra es seleccionada o los datos son medidos. Un aumento del tamaño muestral no ayuda de ninguna manera a reducir el sesgo.

Muestreo Simple al Azar

Una muestra simple al azar de tamaño n es una muestra que se selecciona de una población de forma tal que asegura que cualquier muestra posible del mismo tamaño tiene la misma probabilidad de ser seleccionada.

La definición implica que cada individuo de la población tiene igual chance de ser seleccionado. Sin embargo, el hecho de que cada individuo tenga la misma chance de selección no es suficiente para garantizar una muestra simple al azar.

Un método comunmente empleado para seleccionar una muestra simple al azar es crear una lista llamada **marco muestral**, de objetos o individuos en la población. Se identifica a cada elemento de la lista con un número y se utiliza una tabla de números aleatorios o un generador de números al azar para seleccionar la muestra.

Muestreo con Reemplazo y sin Reemplazo

Cuando se selecciona una muestra aleatoria el muestreo pueder ser:

Con reemplazo: significa que luego que cada elemento es seleccionado en la muestra, dicho elemento es "retornado" a la población y pudiendo ser seleccionado en una paso posterior. (Raramente utilizado en la práctica).

Sin reemplazo: una vez que un elemento es seleccionado en la muestra, no es devuelto a la población.

Aunque las dos formas de muestreo son diferentes, cuando el tamaño de la muestra es pequeño relativo al tamaño de la población, practicamente no existen diferencias prácticas entre los dos métodos.

Muestreo Aleatorio

El objetivo del muestreo aleatorio es producir una muestra representativa de la población estudiada.

Aunque el muestreo aleatorio no garantiza que la muestra sea representativa, es posible utilizar métodos probabilísticos para evaluar el riesgo de obtener una muestra representativa.

Es la habilidad de cuantificar dicho riesgo la que permite generalizar con confianza a partir de una muestra aleatoria a la población correspondiente.

Otros Métodos de Muestreo

En ciertas ocasiones otros métodos de muestreo alternativos pueden ser menos costosos, o más fáciles de implementar que el muestreo simple al azar.

Muestreo estratificado es un método que selecciona independientemente una muestra simple al azar dentro de cada uno de varios subgrupos previamente definidos en una población.

Ejemplo: estimar el costo promedio del seguro por mala práctica en médicos de una determinada comunidad. Dividir en 4 subpoblaciones:

a. cirujanos
b. médicos clínicos
c. obstetras
d. todas las otras especialidades.

Muestreo Estratificado

Comunmente se denomina "estratos" a las subpoblaciones. El muestreo estratificado consiste en seleccionar una muestra simple al azar dentro de cada estrato.

Se puede utilizar muestreo estratificado en lugar de muestreo simple al azar cuando se desea obtener información acerca de características de los individuos en cada estrato asi también como características de la población completa.

Si una población puede ser dividida en estratos, los cuales son homogeneos con respecto a la característica de interés, el muestreo estratificado tiendo a producir estimaciones más precisas que el muestreo simple al azar.

Otros Métodos de Muestreo

Muestreo por conglomerados: las unidades de la población se agrupan en conglomerados. Los conglomerados están formados como si fueran una « mini-población », con cada unidad perteneciente a uno y sólo uno de ellos. Aleatoriamente se seleccionan uno o más conglomerado, así las unidades ingresan a la muestra por conglomerados y no individualmente

Muestreo sistemático: para un muestreo sistemático 1 en k se ordenan las unidades poblacionales de alguna manera y se selecciona aleatoriamente una de las k primeras unidades de dicha lista ordenada.

Esta unidad es la primera unidad a ser incluída en la muestra. Se continúa seleccionando cada k-ésima unidad de dicho listado hasta completar el tamaño de muestra deseado.

Experimentos Comparativos Simples

Investigar el efecto de dos métodos de enseñanza sobre las calificaciones.

Comparar el efecto sobre el desempeño de trabajo de dos diseños de estaciones de trabajo en una fábrica

Comparar un nuevo tratamiento con un tratamiento existente para determinada enfermedad.

Evaluar el efecto del tiempo y la temperatura de cocción sobre el contenido nutricional de determinada clase de pan.

Para responder los problemas anteriores, el investigador lleva a cabo un experimento que permita recolectar la información relevante.

Experimento

Un **experimento** es una intervención planeada llevada a cabo con el fin de observar los efectos de una o más variables explicativas, frecuentemente denominadas **factores**, sobre una variable respuesta.

El propósito fundamental de la intervención es aumentar el conocimiento de la naturaleza de la relación entre las variables explicativas y la respuesta.

Cada combinación de valores para las variables explicativas se denomina "condición experimental" o "**tratamiento**".

El diseño de un experimento es el plan conjunto que se utilizará para realizar el experimento. Un buen diseño minimiza la ambigüedad al momento de interpretar los resultados.

Ejemplo: Efecto de Temperatura en Desempeño

Suponer que interesa determinar cómo la temperatura del salón afecta el desempeño de alumnos de un curso de cálculo. Existen 4 comisiones de cálculo.

Se podría diseñar un experimento de la siguiente manera: asignar a dos comisiones la temperatura 15° y a las otras dos comisiones la temperatura 20°.

Luego comparar las notas promedio para los alumnos en las dos temperaturas diferentes.

Suponer que las notas para alumnos en los salones con 15° de temperatura fueron notablemente más altas que las notas de los alumnos en salones con 20°.

Podemos concluir que el aumento de temperatura produjo menores notas?

...

Comisiones en distintos momentos del día?
Tenían el mismo profesor?
Usaban el mismo libro de texto?
Algunas comisiones dejaban más ejercicios propuestos que otras?
Diferían las comisiones con respecto a la habilidad de los estudiantes?

Cualquiera de estos factores puede ser una explicación de porque los promedios fueron diferentes entre los dos grupos.

No es posible separar los efectos de estos factores de los efectos de la temperatura. Como consecuencia, establecer solamente la temperatura como se describió constituye un experimento pobremente planeado.

Un experimento bien diseñado rquiere mucho más la manipulación de las variables explicativas; el diseño debe también eliminar explicaciones alternativas o los resultados experimentales serán ambiguos.

Factores Externos

El objetivo de la experimentación es diseñar un experimento que permita determianr los efectos de los factores relevantes sobre la variable respuesta elegida.

Para lograrlo, se deben tomar en cuenta otros factores que puedan afectar la variable respuesta denominados **factores externos**.

Un factor externo es uno que no es de interés en el presente estudio pero que se piensa que puede afectar a la variable respuesta.

Un investigador puede **controlar** directamente algunos de los factores externos. Ejemplo: libro de texto utilizado.

Bloques

El efecto de algunos factores externos puede ser filtrado a través de un proceso conocido como **bloqueo**.

El bloqueo crea grupos (bloques) que son similares con respecto a factores de bloqueo y luego cada uno de los tratamientos son empleados dentro de cada bloque.

Ejemplo: instructor. Si existen dos instructores y se considera a los instructores como bloques, cada instructor dictaría una clase con temperatura 15° y otra con 20°.

Si un instructor dictara clases solo en cursos con 15° de temperatura y el otro solo en cursos con 20°, sería imposible distinguir el efecto de la temperatura del efecto del instructor.

Cuando ocurre la situación anterior se dice que los dos factores están "confundidos".

Aleatorización

Es posible controlar factores manteniéndolos constante o través de bloqueo.

Algunos factores no pueden ser controlados por el experimentador como por ejemplo la habilidad de los estudiantes.

Estos factores externos pueden ser manejados mediante el uso de asignación al azar a los grupos experimentales: **aleatorización**.

La aleatorización asegura que el experimento no favorece a una determinada condición experimental e intenta crear grupos experimentales "equivalentes".

Ejemplo: alumnos asignados al azar a cada una de las cuatro cursos.

Aleatorización

No todos los experimentos se refieren a personas evaluadas en diferentes condiciones.

Ejemplo:
comparar 3 diferentes aditivos para el combustible con respecto al rendimiento de un automovil. Se podría realizar el experimento utilizando el mismo automovil, midiendo la distancia recorrida a una velocidad constante luego de haber colocado un litro de combustible con cada uno de los 3 diferentes aditivos.

El procedimiento podría repetirse un determinado número de veces: una secuencia de ensayos.

Debido a que diferentes factores ambientales podrían afectar el rendimiento del automovil, no sería aconsejable realizar todos los ensayos para un tipo de aditivo juntos.

Aleatorización

La asignación aleatoria (de individuos a tratamientos o de tratamientos a ensayos), es una componente crítica de un buen experimento.

La aleatorización puede ser un medio efectivo de equilibrar los efectos de factores externos solamente si el número de individuos u observaciones en cada tratamiento o condición experimental es lo suficientemente grande para que cada grupo experimental refleje de forma confiable la variabilidad presente en la población.

Replicación es la estrategia de diseño que consiste en hacer múltiple observaciones para cada condición experimental.

Experimentos: Resumen

Al planear un experimento o al evaluar un diseño, los siguientes aspectos deben ser considerados:

✓ Replicación: estrategia de diseño que consiste en realizar múltiples observaciones para cada tratamiento experimental.

✓ Control directo: mantener constante el valor de una variable externa de forma que su efecto no se confunde con los factores en el experimento.

✓ Bloqueo: uso de variable externas para crear bloques que son parecidos y luego evaluar todos los tratamientos dentro de cada bloque.

✓Aleatorización: estrategia que permite considerar todos las variables externas que no fueron tomadas en cuenta a través de control directo y bloqueo. La aleatorización permite obtener grupos experimentales "equivalentes".