APUNTES DEL DR. ANDRÉS MENÉNDEZ RAYMAT

EDUC 6390: Estadística aplicada en la educación

Andrés Menéndez Raymat, Ph.D.

Conferencia 1

La estadística en la investigación

Bosquejo

I. Definición

En algunas universidades se habla de Sadística 6390.
Disraeli hablaba de tres tipos de mentiras: lies, damned lies and statistics

A. Importancia de la estadística para los estudiantes

1. Todo ciudadano está en continuo contacto con las estadísticas en todos los medios de comunicación. Debe poder comprender la información que se le ofrece para detectar mentiras y tomar decisiones informadas.
2. Como lector de artículos de investigación debe poder comprender la información cuantitativa que se le ofrece en los artículos que lee
3. Como productor de investigaciones, debe poder utilizar las estadísticas en sus propias investigaciones.

B. Múltiples significados dependiendo del contexto

Meteorología

En este contexto, en la televisión, por estadísticas se refieren a la cantidad de lluvia en el año; la temperatura promedio del mes; la temperatura máxima y mínima del día, etc.

Matemáticas

En este contexto la estadística es el área de aplicación de modelos probabilísticos.

Ciencias sociales

El significado de la estadística en el contexto de la investigación social se enfoca más en los métodos o procedimientos utilizados por los investigadores para comprender e interpretar datos. Es parte integral del proceso de investigación y en la mayoría de las tesis y disertaciones ocupa una posición central.

II. El proceso investigativo

A. Etapas del proceso investigativo

El proceso investigativo tradicional con el que se genera una disertación o tesis consiste de varias etapas o momentos entre los que se distinguen como esenciales
1. el planteamiento del objetivo de la investigación o creación de la pregunta de investigación
2. el planteamiento de las hipótesis de investigación
3. la recolección de datos para someter a prueba la hipótesis de investigación
4. el análisis de los datos recogidos

5. la evaluación de las hipótesis a la luz de estos análisis

B. La función de la estadística dentro del proceso de investigación empírica

Dentro del proceso de investigación se puede llamar estadística a todos y cada uno de los siguientes procesos
1. la recolección de datos a través de cuestionarios y observaciones
2. la presentación de los datos en tablas y gráficas
3. la transcripción de datos por medio de medidas de tendencia central y dispersión
4. el análisis e interpretación de resultados
5. la presentación de conclusiones (donde se interpretan cualitativamente los resultados cuantitativos)
6. el proceso (puntos 1-5) en su totalidad como la justificación para la toma de decisiones.

III. Tipos de estadísticas

A. Estadística descriptiva (Descriptive statistics)

Se origina con la recolección de datos poblacionales para censos. Estos censos se hacen en el imperio romano. El evangelio de Lucas dice : Y aconteció en aquellos días que salió un edicto de parte de César Augusto, mandando que todo el mundo fuera empadronado....
En ella se enfatizan los aspectos de presentar y describir los datos recogidos en la investigación.
En la estadística descriptiva, el investigador debe preocuparse por organizar y presentar los datos de una forma comprensible y sobre todo honesta.
Definición:
Métodos utilizados para organizar, presentar y describir datos de manera adecuada.

B. Estadística inferencial (Inferential statistics)

Se origina en el Renacimiento con el desarrollo de la probabilidad matemática, que a su vez se basa en el estudio de los juegos de azar.
Comienza a desarrollarse plenamente con Karl Pearson (1857-1936) y Ronald Fisher (1890-1962) a principios del siglo XX.
Está íntimamente relacionada con los conceptos de población, muestra, parámetro y estadísticas.

Población (Population)

Es el total de objetos bajo consideración
Es el grupo o conjunto sobre el cual el investigador quiere hacer una inferencia
La mayor parte de las veces es muy grande
Algunas veces es hipotética
Si, por ejemplo, si se quiere demostrar que la semejanza entre personas afecta el nivel de atracción, la población de "personas semejantes" es hipotética pues se hace imposible encontrar una población de personas semejantes en todos los aspectos.

Muestra (sample)

Aunque el investigador se interesa, la mayor parte de las veces en la población, muy pocas veces puede llegar a toda ella. Para hacer cualquier estudio se ve obligado a seleccionar parte de la población.
La muestra es la porción de la población seleccionada para la investigación
La selección se hace porque generalmente el costo, el tiempo y los recursos son limitados para hacer la investigación con toda la población.
Partiendo de los resultados del estudio con la muestra (si ésta es verdaderamente representativa de la población), el investigador puede hacer inferencias sobre la población.

Parámetro (parameter)

Es la medida de una característica numérica de la población. (Media, mediana, varianza, etc.) . Es un elemento descriptivo de la población.

Estadísticas (statistics)

Es una medida que se utiliza para describir una característica numérica de la muestra, no de la población como en el caso del parámetro.
La estadística inferencial sirve para determinar cómo una estadística y un parámetro se relacionan.
Actividad:
(Se organizan grupos de 3 o cuatro estudiantes y se les pide que definan la estadística inferencial utilizando las palabras muestra, población, parámetro, estadística, método. Deben reducir la definición a un mínimo de palabras.)

C. Definiciones posibles de la estadística inferencial

1. Métodos o procedimientos que hacen posible la estimación de una característica de la población basándose exclusivamente en los resultados obtenidos en la muestra.
2. Métodos que hacen posible la estimación de un parámetro basándose

exclusivamente en la estadística correspondiente.
3. Generalizaciones sobre la población basadas exclusivamente en los resultados de la muestra.

Actividades y/o asignaciones:

Tomadas de Holcomb, Z.C. (1997). Real Data. A statistics workbook based on empirical data. Los Angeles, CA: Pyrczak Publishing.
1. What do people look for in prospective partners? A study of personal ads. (Exercise 1. Percentages). pp.1-3.
2. Assaultive behavior of men and women in intimate relationships (Exercise 3. Proportions). pp. 7-10.
Tomadas de Pyrczak, F. (1996). Success at statistics. A worktext with humor. Los Angeles, CA: Pyrczak Publishing.
3.Worksheet 1. Descriptive vs. inferential statistics. (p.3-4)

Lecturas recomendadas:

Hinkle, unidad 1, p.1-20.
Rodríguez-Esquerdo, unidad 1, pp.1-45.
Frankfort-Nachmias & Leon-Guerrero, capítulo 1, pp.1-28
Sirkin, capítulo 1, pp.1-32.

MENU 6390

EDUC 6390: Estadística aplicada en la educación

Andrés Menéndez Raymat, Ph.D.

Conferencia 2

Variables y datos

Bosquejo

I. Variables

En un proceso de observación o en un cuestionario se observan o se hacen preguntas sobre ciertas características de los sujetos o sobre ciertos fenómenos. Estas características que se tratan de medir en la investigación se llaman variables. Cuando en una investigación cuantitativa se pasa de la pregunta o preguntas de investigación a las hipótesis es necesario expresar las hipótesis en términos de variables. Una hipótesis de investigación, por lo general, se expresa como una relación entre dos variables, la variable independiente y la variable dependiente. Éstas se pueden visualizar a veces como la causa y el efecto en la relación. (Aunque pocas veces las relaciones que se establecen en las investigaciones educacionales

corresponden a causas y efectos).

Definición

Una variable es una característica de los sujetos que puede asumir más de un valor. Los valores que asumen las variables en cada uno de los sujetos son los datos.
Ejemplo:
En un estudio en que se trata de determinar el aprovechamiento de los estudiantes en una escuela, una de las variables puede ser la nota obtenida en el curso de estadísticas. Los datos son, en este caso, las notas obtenidas en estadísticas por cada uno de los estudiantes que han tomado el curso.

II. Datos

A. Definición de datos

Los datos pueden definirse como la información recogida, organizada y analizada por los estadísticos.

B. Métodos para obtener los datos en una investigación

1. Datos publicados

Se pueden utilizar datos publicados previamente que el investigador no tiene que recoger. Estamos en la época de la tecnología y la información. Las bibliotecas están equipadas con computadoras y a través de éstas se pueden localizar bancos de datos que otras personas o instituciones han recogido y almacenado. Estas fuentes de datos para las investigaciones pueden ser:
Fuentes primarias. Estas son las personas u organizaciones que recogen los datos directamente.
Fuentes secundarias. Son las personas u organizaciones que han compilado los datos en tablas y gráficas. Por lo general, tanto el gobierno como las universidades son fuentes primarias y secundarias.

2. Datos obtenidos en la experimentación

En la investigación a menudo se utilizan datos obtenidos a través de la experimentación. Esto ocurre principalmente en las investigaciones de medicina y de ciencias naturales. La investigación consiste en el montaje de un experimento en que se controlan todas las variables que pueden influir en los resultados y entonces se maneja la variable independiente y se observan los cambios en la variable dependiente. Cuando esto ocurre se puede hablar de una relación de "causa y efecto". La investigación es un verdadero experimento.
En las ciencias sociales es más difícil puesto que se dificulta imponer controles sobre el medio social.
En el momento de la recolección de datos debe haber control sobre todas las variables que pueden afectar variaciones en el experimento.
Para hablar de un "experimento" y de una relación de causa y efecto es necesario que se den tres condiciones:
a. la variable dependiente es el objetivo de la investigación. Se trata de determinar cómo se modifica esta variable dependiente cuando se ha modificado la variable independiente.
b. la modificación en la variable independiente ocurre antes que la modificación en la variable

dependiente
c. la variable independiente ejerce una influencia directa o indirecta en la variable dependiente.

3. Datos obtenidos a través de cuestionarios

La forma más común de llevar a cabo una investigación en las ciencias sociales es utilizando datos obtenidos a través de cuestionarios.
En estos casos no se ejerce control sobre el comportamiento de las personas. Sólo se hacen preguntas y se observan las dos variables (independiente y dependiente) al mismo tiempo.
En los cuestionarios no se busca una relación de causa y efecto, sino de correlación entre dos variables. Se busca determinar si la magnitud una variable se relaciona con la magnitud de la otra. Por lo general no se habla de variables independientes y dependientes, sino de predictores y criterios. El cambio en el predictor no es la causa del cambio en el criterio aunque un cambio implique el otro.
Ejemplos:
Causa y efecto: vacuna y prevención de la enfermedad. La vacuna es la causa de que la enfermedad no tenga lugar.
Correlación: Se observa en la relación entre preparación académica y salario. Por lo general, mientras mayor es la preparación académica mayor es el salario. Pero esto no siempre ocurre y la causa del mayor salario puede muy bien ser otra diferente de la preparación académica.

Una correlación que hace obvia esta situación es la alta relación que hay entre tamaño de pie y destrezas de lectura. Nadie en su sano juicio puede asegurar que el tamaño del pie es la causa de las destrezas, sin embargo mientras mayor es el tamaño de pie, mayor es la habilidad en la lectura. Hay una variable escondida que es la causa de ambas (el crecimiento).

4. Datos obtenidos de la observación

Se utilizan mucho en antropología y en investigaciones sobre animales.
Este método de recoger datos tiene problemas debido a la subjetividad del observador y al hecho de que la presencia del observador puede modificar la situación.

III. Clasificación de datos y variables

Datos o variables		**Preguntas**	**Respuestas**
Categóricas o cualitativas		Tienes pasaporte	si / no (dicótoma)
Numéricas o cuantitativas	discretas	¿Cuántas camisas tienes?	Número natural
	continuas	¿Cuánto pesas?	Número real

Esta clasificación se puede aplicar tanto a las variables como a los datos.

A. Datos o variables categóricas / cualitativas

Son características que se refieren a categorías como sexo, afiliación política, color de los ojos. Por lo general, estas características no se pueden describir por medio de números.
En los cuestionarios, por lo general, las preguntas sobre estas variables se pueden responder con "si" o "no".
Ejemplos:

¿Posees un carro? ¿Vives en una casa? ¿Tienes los ojos azules?

B. Datos o variables numéricas / cuantitativas

Las preguntas que se hacen sobre estas variables se pueden responder con un número. ¿Cuánto pesas? ¿Cuánto mides? ¿Cuánto dinero ganas? ¿Cuántos hijos tienes?
Las variables numéricas pueden ser :

Discretas

Una variable es discreta si sus valores se pueden contar; si existe una relación biunívoca con el conjunto de los números naturales. Existe una unidad mínima que no puede subdividirse. Ejemplo: cantidad de carros, de hijos, ingreso anual, etc. No se puede tener medio hijo o un cuarto de carro.

Continuas

Los valores de estas variables no se pueden contar puesto que siempre existe un número entre dos de ellos. Generalmente se encuentran en los procesos de medición, como peso, altura, temperatura. Además en la realidad no hay dos sujetos con la misma medida. No existe una unidad indivisible para los datos continuos como ocurre con los datos discretos.
Sin embargo, debido a que los instrumentos de medición generalmente no son muy sofisticados y precisos, en las investigaciones se encuentran sujetos con la misma medida (debido al redondeo)

Por extensión las variables reciben el mismo nombre de los datos. Pueden ser categóricas y numéricas. Si son numéricas pueden clasificarse en discretas y continuas. Cuando una variable numérica o categórica puede tener solamente dos valores se llama **dicótoma**.

IV. Escalas de medición

Variables categóricas

Variable categórica	Categorías	Escala
partido político	PPD; PNP; PIP	nominal
género	mujer, hombre	nominal
colores	negro, rojo,.....	nominal
satisfacción	mucha, mediana, poca	ordinal
nota	A, B, C, D, F	ordinal

Variables numéricas

Variable numérica	Escala
temperatura	intervalar
IQ	intervalar
peso	razón

altura	razón
edad	razón

Uno de los puntos más importantes de una investigación es determinar el tipo de análisis estadístico de los datos que se va a llevar a cabo. En estadísticas el tipo de análisis depende del nivel o escala de medición de las variables de la investigación. Los cursos de estadísticas en las ciencias de la educación y en las ciencias sociales están diseñados para que los estudiantes aprendan diversos métodos de análisis estadístico. Pero antes de aplicar cualquiera de ellos, el estudiante debe haber determinado el nivel de medición de sus variables de investigación. Cada nivel requiere un análisis diferente. La importancia de esta clasificación por niveles reside en el hecho de que mientras más complejo o alto es el nivel de medición, más efectivos son los métodos estadísticos que se pueden utilizar. Para hablar de niveles o escalas de medición es imprescindible primeramente clarificar que es "medir" en las ciencias sociales. Medir no es solamente determinar las dimensiones de un objeto. En las ciencias sociales la idea de medición se aplica en muchas otras ocasiones.

Ejemplos:
Se mide cuando determina:
la religión de una persona
el color de pelo
el ingreso anual
el género
el peso
el tamaño
la puntuación en un examen
la nota

El tipo de medida que se obtiene en la investigación puede caer en una de las siguiente cuatro escalas o niveles de medición: Si los datos son categóricos o cualitativos, entonces dependiendo del grado de precisión posible en la medición, se utilizan las siguientes dos escalas:

A. Escala nominal

La escala nominal se utiliza cuando los datos están clasificados en categorías en las que no hay ninguna idea de ordenamiento. No se puede decir que una categoría es mejor que otra. El propósito en este nivel es solamente clasificar, nombrar los datos. Se refiere a atributos de los sujetos, no a cantidades. En ningún momento se habla de números, aunque a la hora de entrar los datos a la computadora se puede asignar un número para hacer la entrada de datos más simple. A veces se ve asignar el #1 al género femenino y el #2 al masculino. En College Board se utilizan el #7 y el #8 para los dos géneros. Pero esta asignación de valores es arbitraria.
Ejemplos:
colores, religiones, partidos políticos, etc.

B. Escala ordinal

Hay orden en este nivel de medición. Se sugiere un rango en las categoría de forma que una categoría es mejor, más importante o mayor que otra. Sin embargo, no hay un sentido

numérico para este orden. La diferencia entre dos rangos no es una cantidad exacta.
Ejemplo:
En una escala Likert los rangos pueden ser: Acuerdo total, acuerdo parcial, desacuerdo parcial y desacuerdo total.
En este caso es posible que la diferencia entre acuerdo total y acuerdo parcial se pueda interpretar de formas diferentes por personas diferentes. Se hace imposible medir numéricamente la diferencia entre acuerdo total y acuerdo parcial, aunque es obvio que uno es mayor o mejor que otro.
NO existen "unidades de acuerdo" que permitan decir que entre acuerdo parcial y acuerdo total hay " 5 unidades de acuerdo".
En muchas ocasiones se usan números para codificar estas respuestas como:
acuerdo total (5),
acuerdo parcial (4),
indeciso(3),
desacuerdo parcial (2),
desacuerdo total (1).
Estos números sólo representan orden. En ningún momento se implica que la diferencia entre acuerdo total y acuerdo parcial es de una unidad.

C. Escala intervalar

La escala intervalar se utiliza con datos numéricos. Cada sujeto recibe un número.
Ejemplos:
Puntuaciones en la Prueba de Razonamiento Verbal del College Board; IQ; temperatura del agua
Estos datos se pueden sumar y restar. Es posible decir cuánto mejor salió un estudiante en la prueba que otro, o cuanto supera un sujeto a otro en IQ. La diferencia entre dos medidas es significativa
Ejemplo:
79 grados es 2 más que 77 grados de temperatura. La diferencia entre 79 y 77 grados es la misma que entre 55 y 53 grados.
Sin embargo no hay un cero verdadero. El cero en temperatura Fahrenheit es una temperatura seleccionada al azar. El cero en centígrados corresponde a otra temperatura muy diferente. El resultado es que, a pesar que 100 es el doble de 50, en una temperatura de 100 no hace el doble de calor que en una de 50. No se siente el doble de calor.
Ejemplo:
Un niño con un IQ de 150 no tiene el doble de inteligencia de uno con un IQ de 75. Esta es una de las razones por las que cuando se mide un atributo psicológico por medio de un instrumento, el cero no tiene sentido y en muchos de los instrumentos simplemente no existe.

D. Escala de razón

Tiene un cero real. Tanto a - b (a menos b) como a/b (a dividido entre b) tienen significado.
Ejemplo:
peso, altura.
Tiene sentido hablar de que una persona pesa el doble de otra. O que alguien tiene el doble de años que otro.

Esta clasificación de las escalas para medir las variables es sumamente importante, pues dependiendo de la escala se van a seleccionar lo métodos estadísticos que se pueden

emplear.
Los métodos más precisos en términos de predicción son los métodos paramétricos, que son los que se estudian en este curso. Para poder usarlos, las variables tienen que medirse en escalas intervalares o de razón. Cuando las escalas son nominales u ordinales, no queda más remedio que utilizar métodos estadísticos no paramétricos, que no son tan precisos en su medición.
Es siempre posible pasar de una escala a otra menos exigente.
Ejemplo:
Los estudiantes pueden medirse en pulgadas (razón) o simplemente se pueden ordenar de mayor a menor (ordinal).

V. Codificación y entrada de datos a la computadora

Para codificar datos se crea un libro de código (codebook) donde:
1. Se identifican las variables y se asignan las columnas requeridas para cada una de las variables
2. Se asignan códigos para cada valor de las variables no numéricas.
3. Se construye un archivo de datos (data file) en la computadora donde cada fila corresponde a un sujeto diferente.
4. Esta entrada de datos se puede hacer directamente usando un procesador de palabras cualquiera, pero se dificulta el proceso.
5. Generalmente se utiliza una hoja electrónica de datos (spreadsheet) de un programa estadístico como Excel o SPSS en los cuales se identifican mejor las columnas y es más fácil la entrada de datos.
Ejemplo de un libro de c ó digo:

Actividades y/o asignaciones:

Tomadas de Pyrczak, F. (1996). Success at statistics. A worktext with humor. Los Angeles, CA: Pyrczak Publishing.
1.Actividad: Scales of Measurement. Section 2. Pyrzack, pp.5-8
2.Discusión del libro de código y hojas de datos en Word y SPSS
3.Hinkle pp.19-20 ej. 8

Lecturas recomendadas:

Hinkle, unidad 1, p.13-17.
Hinkle, unidad 2, p.21-25
Rodríguez-Esquerdo, unidad 1, pp.1-45.
Frankfort-Nachmias & Leon-Guerrero, capítulo 1, pp.1-28
Sirkin, capítulo 2, pp.33-61.

MENU 6390

EDUC 6390: Estadística aplicada en la educación

Andrés Menéndez Raymat, Ph.D.

Conferencia 3

Organización de datos

Bosquejo

I. Organización de datos categóricos (una sola variable)

A. Forma individual

Cuando los datos recogidos corresponden a una variable categórica, estos se pueden presentar de forma individual, en la columna asignada a la variable se indica la categoría a la que pertenece el sujeto

Ejemplo:

Sujeto	Partido político	Religión
Juana	PPD	católica
Pedro	PIP	católica
María	PNP	protestante

B. Frecuencia

Sin embargo, muy a menudo el investigador está interesado en señalar cuán a menudo ocurre cada valor de la variable. A esta indicación de cuán a menudo ocurre un valor se le llama la frecuencia y su símbolo es una "***f*** "minúscula.

Para indicar la frecuencia se utilizan tablas que sumarizan la información (summary tables) generalmente en términos de frecuencia absoluta y de frecuencia relativa (porciento o proporción) de casos que corresponden a cada categoría.

C. Frecuencia absoluta y relativa

En la tabla sólo es necesario presentar las categorías y la frecuencia absoluta (cantidad de sujetos) o relativa (proporción o porcentaje) de esas categorías.

Ejemplo:

Partido político	frecuencia	porcentaje
PPD	4	44.4
PNP	3	33.3
PIP	2	22.2
Total	9	99.9

D. Datos nominales u ordinales

Nota:

Los datos pueden ser nominales u ordinales. La única diferencia radica en que cuando los datos son ordinales se deben colocar las categorías de mayor a menor o de menor a mayor.

Nivel socioeconómico	frecuencia	proporción
Bajo	10	0.40
Medio bajo	4	0.16
Medio alto	9	0.36
Alto	2	0.08

Total	25	1

II. Organización de datos numéricos (una sola variable)

Una distribución es el nombre que se da a cualquier conjunto organizado de datos. Esta organización se puede hacer por medio de una tabla o de una gráfica. Cuando en estadísticas se habla de cómo los datos están distribuidos en una muestra o en una población nos referimos al conjunto de datos organizados en una tabla o en una gráfica. La idea que se persigue en la estadística descriptiva es dar una estructura a los datos que permita al lector identificar sus aspectos más importantes.

A continuación se observará un proceso de organización creciente que permite distinguir mejor las características más sobresalientes de los datos numéricos.

A. Datos crudos (raw data)

Generalmente cuando se recogen los datos crudos (raw data) de un estudio no tienen una estructura de presentación definida. En esta representación no hay mucho que se pueda decir de los datos.

Puntuaciones en un examen de estadísticas de una clase de 50 estudiantes

76	65	89	86	45
35	66	55	99	95
87	85	83	84	68
72	74	85	68	76
97	52	24	76	77
80	94	90	64	61
83	84	74	76	68
57	47	65	94	97
47	53	52	64	42
32	33	16	47	69

(Transparencia T3.1)

B. Datos ordenados

Después se puede crear una tabla de datos ordenados donde se colocan los

datos en orden (de menor a mayor o viceversa). Este primer orden permite identificar los valores extremos, pero no ofrece mucha más información.

16	24	32	33	35
42	45	47	47	47
52	52	53	55	57
61	64	64	65	65
66	68	68	69	70
72	74	74	76	76
76	76	77	80	83
83	84	84	85	85
86	87	89	90	94
94	95	97	97	99

(Transparencia T3.1)

C. Diagrama de tallo y hoja (stem and leaf display)

Los datos también se pueden organizar en un diagrama de tallo y hoja (stem and leaf display). Para ello los datos se separan en dígitos principales que conforman los tallos y se utiliza el dígito final para las hojas.

1	6
2	4
3	532
4	77275
5	27352
6	149556848
7	064762664
8	9334554076
9	4405977

(Transparencia T3.2)

En el caso de este ejemplo las decenas forman el tallo y las unidades las hojas. Se ha trazado una raza una raya entre tallos y hojas. Se ha establecido un orden (menor a mayor) para los tallos. Las hojas, sin embargo, se pueden incluir en el orden en que se recogieron los datos. Toda tabla debe incluir una explicación del significado del tallo y la hoja.

Ejemplo: 2/ 4 = 24

En un diagrama de tallo y hoja además de determinar fácilmente los valores máximos y mínimos se hace mucho más fácil notar donde existe una mayor concentración de datos.

1. Diagrama de tallo y hoja revisado

1	6
2	4
3	235
4	25777
5	22357
6	144556889
7	024466667
8	0334455679
9	0445779

(Transparencia T3.2)

En este diagrama se han ordenado las hojas de forma ascendente.

2. Diagrama de tallo y hoja modificado

El diagrama de tallo y hoja se puede modificar de múltiples maneras dependiendo de las necesidades de la presentación.

i. Tallos ampliados

Se pueden ampliar los tallos si se desea ver los datos menos agrupados.

1	
1	6
2	4
2	
3	235
4	2
4	5777
5	223
5	57
6	144
6	556889
7	0244

7	66667
8	03344
8	55679
9	044
9	5779

(Transparencia T3.3)

El diagrama de tallo y hoja anterior tiene los tallos separados en dos partes correspondientes a las hojas menores y mayores. En la primera parte del tallo se incluyen las hojas bajas (de 0 a 4) y en la segunda parte del tallo se incluyen las hojas altas (de 5 a 9).

ii. Tallos agrupados

Dos tallos también se pueden agrupar de forma que la separación entre los dos se haga notar por la presencia de una coma. Las hojas correspondientes a cada tallo se indican por medio del uso de negritas para las hojas correspondientes a uno de los tallos.

1, **2**	6**4**
3, **4**	235**25777**
5, **6**	22357**144556889**
7, **8**	02446666**70334455679**
9, **10**	0445779

(Transparencia T3.3a)

iii. Hojas comparadas

Las hojas también se pueden colocar a la derecha y a la izquierda del tallo. Esto se hace sobre todo cuando se quieren comparar dos grupos de datos

Puntuaciones de estudiantes en dos secciones de un curso

Sección A		**Sección B**
0	5	245
	6	5
55431	7	22566
999875542	8	0677
31	9	122233

(Transparencia T3.3b)

En este caso se puede llevar a cabo una comparación de los dos grupos

(Actividad de Tallo y Hoja)

D. Distribución simple de frecuencia absoluta (para variables discretas)

Una distribución simple de frecuencia absoluta es una tabla que indica el número de veces que ha ocurrido cada valor en un conjunto de datos. La representación en una tabla se puede hacer con dos columnas donde una indica el valor de la variable y la otra columna la frecuencia de cada valor. Por lo general estas tablas en su última fila incluyen el total de las frecuencias que se representa con una N mayúscula.

Ejemplo: (T. Table 3.1)

Puntuación	*f*
24	1
25	1
26	0
27	0
28	0
29	1
30	1
31	0
32	2
33	3
34	1
35	2
36	4
37	5
38	4
39	3
40	4
41	5
42	5
43	4
44	4
45	7

46	9
47	7
48	8
49	11
50	7
51	3
52	6
53	7
54	7
55	12
56	14
57	6
58	2
59	3
60	2
61	1
62	3
63	5
64	4
65	2
66	0
67	1
68	2
69	1
Total	N = 180

E. Distribución agrupada de frecuencia absoluta

Sin embargo, a veces, cuando hay muchos valores posibles para la variable es necesario condensar estos valores en clases o intervalos. Esta agrupación se llama una distribución agrupada de frecuencia absoluta porque en ella se indica cuan frecuentemente aparecen datos en cada grupo. La información inicial de la frecuencia de cada valor individual se pierde, pero es más fácil determinar rápidamente las características principales del conjunto de datos.

Aspectos importantes que se tienen que tener en cuenta cuando se crea una distribución agrupada de frecuencias

1. Variables discretas

i. Número de intervalos

El número de intervalos depende del número total de observaciones. No debe haber más de 15 ni menos de 5. Si hay muy pocos se pierde mucha información. Si hay muchos no se ven las características más importantes.

ii. Ancho de los intervalos

Todos los intervalos en una tabla de distribución de frecuencia deben tener el mismo ancho. Pero hay excepciones, especialmente en el último intervalo. En los informes estadísticos del College Board todos los intervalos van hasta 299, 399, etc, excepto por el último que va hasta 800. El programa de Excel se ajusta a esta posibilidad cuando construye las tablas de distribución de frecuencia y los histogramas. En estos casos indica que no se incluya el número final del último intervalo, pues Excel incluye todo lo que resta de la distribución en ese intervalo.

El ancho del intervalo se define de formas diferentes dependiendo del autor.

Weiss (p.52) lo define como la diferencia entre el límite inferior de un intervalo y el límite inferior del próximo intervalo.

Sirkin (p.50) lo define como la diferencia entre el límite superior y el límite inferior del mismo intervalo.

Si el investigador construye una tabla de distribución de frecuencia debe seguir los siguientes pasos:

1. Escoger el número de intervalos que desea tener en la tabla.

2. Determinar el ancho de los intervalos. Para determinar el ancho de cada intervalo, se divide el alcance o amplitud de los datos (diferencia entre el dato mayor y el menor) por el número de intervalos que se desean. Finalmente se redondea el número obtenido

Ejemplo: Se utilizará el ejemplo anterior de 180 datos

El alcance va de 69 a 24 y vamos a considerar que 9 ó 10 intervalos es un número adecuado. Por lo tanto (69-24)/9 = 45/9 = 5 . El ancho será de 5 valores.

iii. Límites de los intervalos

El conjunto de intervalos debe tener las siguientes propiedades:

1. Los intervalos deben incluir todas las observaciones

2. No debe haber solapamiento (overlapping) de intervalos

Para cumplir con estas dos propiedades la mejor forma de proceder es describir el intervalo por medio de su valor mínimo y máximo. Estos deben ser valores fáciles de organizar y recordar.

En el ejemplo en vez de hacer el primer intervalo de 24 a 28 se hace de 20 a 24 pues el límite inferior se identifica y recuerda mejor si es un múltiplo de 5.

El límite superior corresponde al valor mayor que se puede incluir en dicho intervalo. Si suponemos que en cada intervalo se incluyen 5 valores, estos son: 20, 21, 22, 23 y 24.

Ejemplo: (T. Table 3.1a)

puntuación	*f*
20-24	1
25-29	2
30-34	7
35-39	18
40-44	22
45-49	42
50-54	30
55-59	37
60-64	15
65-69	6
Total	N = 180

Nota:

Según Weiss el ancho de estos intervalos es de 25 - 20 = 5

Hinkle (p.29) habla de los límites exactos que se utilizarán posteriormente para construir el histograma. Estos límites exactos se encuentran a mitad de camino entre el límite superior de un intervalo y el límite inferior del próximo intervalo.

(19.5 y 24.5 para el primer intervalo; 24.5 y 29.5 para el segundo, etc.)

2. Variables continuas

Se puede hacer lo mismo con algunas modificaciones cuando la variable es continua. La diferencia radica en que en cada intervalo se incluye el valor extremo mínimo y se excluye el valor mínimo del próximo intervalo.

Aspectos importantes que se deben tener en cuenta cuando se crea una distribución agrupada de frecuencias

i. Número de intervalos

Depende del número total de observaciones. No más de 15 ni menos de 5. Si hay muy pocos se pierde mucha información. Si hay muchos no se ven las características más importantes.

ii. Ancho de los intervalos

Todos deben tener el mismo ancho. Se divide el alcance por el número de intervalos que se desean y se redondea

Ejemplo:

Vamos a considerar 60 datos continuos que se obtuvieron en una investigación sobre el peso en kilogramos de unos perros realengos. El perro más flaco pesó 2.4 kg y el más gordo 12.0 kg. El alcance va de 2.4 a 12.0 y consideraremos que 6 intervalos es un número adecuado. Por lo tanto (12 - 2.4)/6 = 9.6/6 = 1.6 y redondeamos a 2.

iii. Límites de los intervalos

Deben incluirse todas las observaciones. No debe haber solapamiento (overlapping). Cada dato debe pertenecer exclusivamente a un intervalo o clase. Esto se logra describiendo el intervalo por medio de su valor mínimo y máximo. Estos valores se llaman el límite superior y el límite inferior del intervalo o clase. (Weiss p.52)

Transparencia T3.4

Peso de los perros realengos	*f*

2.0 ≤ x < 4.0	13
4.0 ≤ x < 6.0	24
6.0 ≤ x < 8.0	9
8.0 ≤ x < 10.0	8
10.0 ≤ x < 12.0	5
12.0 ≤ x < 14.0	1
Total	60

Nota: Según Weiss el ancho de estos intervalos es de 4.0 - 2.0 = 2

****Sin embargo Hinkle y otros autores prefieren este mismo sistema para las variables discretas y continuas**.

Ejemplo con la variable discreta “puntuación”.

puntuación	*f*
20 ≤ x < 25	1
25 ≤ x < 30	2
30 ≤ x < 35	7
35 ≤ x < 40	18
40 ≤ x < 45	22
45 ≤ x < 50	42
50 ≤ x < 55	30
55 ≤ x < 60	37
60 ≤ x < 65	15
65 ≤ x < 70	6
Total	N = 180

Nota: Hinkle prefiere ir de mayor a menor y distingue entre límites exactos y límites de las puntuaciones. Pero para continuar con lo establecido en el diagrama de tallo y hoja es preferible ir de menor a mayor. Para no complicarse la vida cuando la variable es continua se usa el método de inclusión del valor mínimo y exclusión del valor máximo.

iv. El punto medio del intervalo

(a veces se le llama la marca del intervalo o clase (class mark)

Cuando la variable es discreta el punto medio corresponde al valor que se sitúa

en el mismo medio de los otros valores. Si la variable es discreta, a la hora de crear los intervalos se hace el esfuerzo por tener un número impar de valores en cada intervalo para que el valor del medio sirva de punto medio.

Cuando la variable es continua, o cuando el intervalo se expresa con símbolos de “menor que” ($20 \leq x < 25$), el punto medio es el punto que está a mitad de camino entre los límites de un intervalo. Se halla sumando los límites y dividiendo entre 2. (Weiss, p.57). En este ejemplo el punto medio de $20 \leq x < 25$ es 22.5

F. Distribución de frecuencia relativa (proporción y porcentaje)

La frecuencia relativa se obtiene dividiendo las frecuencias de cada clase por el número total de observaciones. Este resultado se puede expresar como una proporción o como un porcentaje.

Peso de los perros realengos	***f***
$2.0 \leq x < 4.0$	0.22
$4.0 \leq x < 6.0$	0.40
$6.0 \leq x < 8.0$	0.15
$8.0 \leq x < 10.0$	0.13
$10.0 \leq x < 12.0$	0.08
$12.0 \leq x < 14.0$	0.02
Total	1.00

Transparencia T3.4

Por lo general se utiliza más la distribución de frecuencia relativa expresada en porcentajes que expresada en proporciones.

A menudo la frecuencia absoluta y la relativa aparecen en la misma tabla

Peso de los perros realengos	***f***	***f***
$2.0 \leq x < 4.0$	13	0.22
$4.0 \leq x < 6.0$	24	0.40
$6.0 \leq x < 8.0$	9	0.15
$8.0 \leq x < 10.0$	8	0.13
$10.0 \leq x < 12.0$	5	0.08

12.0 ≤ x < 14.0	1	0.02
Total	60	1.00

Transparencia T3.5

La distribución de frecuencia relativa es esencial si se quieren comparar datos de dos distribuciones diferentes.

Ejemplo:

Comparar las frecuencias del estudio de los perros con las frecuencias de otro estudio sobre 45 perros que tienen dueño. Debido al número de perros la comparación no es clara, pues en un ejemplo hay 60 perros y en el otro hay 45 perros.

Peso de los perros	***f realengos***	***f con dueño***
2.0 ≤ x < 4.0	13	1
4.0 ≤ x < 6.0	24	10
6.0 ≤ x < 8.0	9	15
8.0 ≤ x < 10.0	8	10
10.0 ≤ x < 12.0	5	7
12.0 ≤ x < 14.0	1	2
Total	60	45

Sin embargo, si se comparan los porcentajes se puede concluir que :

Peso de los perros	***Porcentaje de realengos***	***Porcentaje con dueño***
2.0 ≤ x < 4.0	22	2
4.0 ≤ x < 6.0	40	22
6.0 ≤ x < 8.0	15	33
8.0 ≤ x < 10.0	13	22
10.0 ≤ x < 12.0	8	16
12.0 ≤ x < 14.0	2	5
Total	100	100

Comparación Los perros con dueño pesan más que los perros realengos

puesto que:

Los porcentajes son menores en los valores bajos para los perros con dueño.

Los porcentajes son mayores en los valores altos para los perros con dueño.

La concentración de perros con dueño está en el intervalo $6.0 \leq x < 8.0$ y la concentración de perros realengos en el intervalo $4.0 \leq x < 6.0$

G. Distribución de frecuencia acumulada

En la distribución de frecuencia acumulada se indica la frecuencia, la proporción o el porcentaje de los casos acumulados hasta cierto intervalo o clase (inclusive).

Construcción:

La columna correspondiente a la frecuencia acumulada se construye cuando se añade a la frecuencia en cada intervalo la frecuencia de todos los intervalos inferiores.

Peso de los perros realengos	***punto medio***	***f absoluta***	***f absoluta acumulada***	***f relativa***	***f relativa acumulada***
$2.0 \leq x < 4.0$	3	13	13	22	22
$4.0 \leq x < 6.0$	5	24	37	40	62
$6.0 \leq x < 8.0$	7	9	46	15	77
$8.0 \leq x < 10.0$	9	8	54	13	90
$10.0 \leq x < 12.0$	11	5	59	8	98
$12.0 \leq x < 14.0$	13	1	60	2	100
Total		60		100	

Transparencia T3.9a

III. Organización de datos (dos o más variables)

Cuando los datos que se obtienen corresponden a dos o más características de los mismos sujetos se pueden crear tablas como las ya estudiadas para cada característica, pero también se puede crear una tabla donde se representa el valor de cada variable para cada sujeto. El siguiente ejemplo con 5 estudiantes

corresponde a un estudio sobre el acceso a las computadoras en la escuela secundaria

sujeto	*lugar de acceso*	*género*
1	escuela	hombre
2	casa	mujer
3	escuela	hombre
4	casa	mujer
5	casa de amigo	mujer

El problema con esta tabla es que se hace sumamente difícil poder apreciar cual es la relación entre las variables.

A. Organización de datos categóricos (dos variables)

Tablas de contingencia

Las tablas en las que se presentan dos variables categóricas se llaman tablas de contingencia (contingency tables) y se utilizan para analizar y comparar las frecuencias de dos variables categóricas. Utilizaremos el ejemplo anterior pero ahora con estudiantes:

	Lugar de acceso			
género	*escuela*	*casa*	*casa de amigo*	total
mujer	45	36	2	93
hombre	32	53	20	95
total	77	89	22	188

Transparencia T5.6

Como las comparaciones se deben hacer siempre en términos de porcentajes o proporciones, las frecuencias absolutas deben convertirse en frecuencias relativas (preferiblemente porcentajes, pues la gente los entiende más).

El problema radica en que hay tres tipos de porcentajes: fila, columna y total. Es básico determinar de antemano cuál es el que se necesita en cada caso y esto depende de qué es lo que se quiere comparar.

1. Frecuencia relativa del total

Para obtener el porcentaje de cada celda con respecto al total se divide el número en cada celda entre 188 (número total)

	Lugar de acceso			
género	*escuela*	*casa*	*casa de amigo*	total
mujer	24	19	1	49
hombre	17	28	11	51
total	41	47	12	100

Gracias a esta tabla podemos concluir que el lugar de acceso preferido es la casa (47%) y que hubo más o menos la misma proporción de hombres y mujeres que contestaron el cuestionario (49% vs 51%).

2. Frecuencia relativa de fila

Sin embargo hay otra información importante que se descubre cuando se obtienen los porcentajes de fila dividiendo cada celda de una fila entre el total de esa misma fila (la primera fila entre 49 y la segunda entre 51)

	Lugar de acceso			
género	*escuela*	*casa*	*casa de amigo*	total
mujer	48	39	2	100
hombre	34	56	21	100

Esta tabla permite llegar a conclusiones mucho más importantes que la anterior pues en ella se observa claramente que las mujeres tienen como lugar de acceso preferido la escuela (48 % vs 39 % ó 2%), mientras que los hombres prefieren la casa (56% vs 34% ó 21%).

3. Frecuencia relativa de columna

Actividad interesante: Halla los porcentajes de columna e indica a qué conclusiones te permiten llegar esos porcentajes.

Actividades y/o asignaciones:

1. Actividad de Tallo y Hoja
2. Hinkle ej. 3 p.50
3. *Worksheet 5. Frequency distributions for grouped data. Success at statistics (p.21-22)

4. *Worksheet 6. Cumulative frequencies, cumulative percentages, and percentile ranks. Success at statistics. (p.25-26)
6. Actividad con tabla de contingencia
*Tomadas de Pyrczak, F. (1992). Success at statistics. A worktext with humor. Los Angeles, CA: Pyrczak

Lecturas recomendadas:

Hinkle, unidad 2, p.25-31
Rodríguez-Esquerdo, unidad 2, pp.47-142.
Frankfort-Nachmias & Leon-Guerrero, capítulo 2, pp.29-70
Sirkin, capítulo 2, pp.33-61.

MENU 6390

EDUC 6390: Estadística aplicada en la educación

Andrés Menéndez Raymat, Ph.D.

Conferencia 4

Representación de datos

Bosquejo

Toda información cuantitativa puede representarse de tres formas: aritmética, geométrica y algebraica, que es como decir con números, gráficas y finalmente con símbolos. En ocasiones, una gráfica puede ayudar a transmitir mucha más información que una tabla de datos y en la mayor parte de las situaciones, el público se ve menos amenazado por una gráfica que por una tabla.

En la presentación de las tablas se comenzó con aquellas apropiadas para datos categóricos y luego numéricos. De igual forma en esta unidad primero se describirán las tablas apropiadas para una variable categórica (nominal u ordinal), después para una variable numérica y finalmente para dos variables, una categórica y la otra numérica o ambas numéricas.

I. Gráficas de datos categóricos (una sola variable)

A. Gráficas circulares (pie charts)

(Transparencia T5.3)

Es una de las gráficas que más se utilizan. Sirve para representar diferencias tanto en frecuencias como en porcientos y proporciones entre las diferentes categorías de una variable nominal u ordinal.

Las categorías aparecen como sectores del interior de un círculo. El tamaño de cada sector depende del porciento que representa y el círculo en su totalidad representa un 100%. A menudo los sectores que se quieren hacer resaltar aparecen sobresaliendo del círculo y en colores que refuerzan el contraste.

Las gráficas circulares no se usan si hay muchas categorías o éstas son tan pequeñas que se hace difícil identificarlas.

Antiguamente se creaban usando el compás y el transportador para determinar la medida de los ángulos. El ángulo se determinaba por medio de la proporción ángulo/360 = por ciento/100

Los programas como Excel han resuelto este problema al construir las gráficas automáticamente.

B. Gráficas de barras (bar graphs)

(Transparencia T5.2)

Estas gráficas proveen una alternativa a la presentación de las gráficas circulares.

Cada categoría se representa por una barra que puede ser horizontal o vertical. Cada día se ven más libros presentando las gráficas de barras horizontalmente

con las categorías descritas en el eje vertical o en la misma barra.

En una gráfica de barras todas las barras deben ser del mismo ancho, donde el largo de la barra indica la frecuencia o porcentaje de cada categoría.

A menudo la gráfica de barras se utiliza para representar dos variables categóricas. En cada categoría de la primera variable se construyen varias barras que representan las categorías de la segunda variable.

Características:

a. A menudo se usan barras horizontales para distinguir las variables categóricas de las numéricas que se representan por medio de barras verticales.

b. Todas las barras tienen el mismo ancho. La diferencia entre ellas radica en el largo.

c. Los espacios entre barras deben ser de más o menos la mitad del ancho de cada barra

d. En el eje horizontal se marcan las frecuencias comenzando con el cero. Este cero para la frecuencia debe indicarse en el eje horizontal.

C. Gráfica de puntos (Dot chart)

La gráfica de puntos es una modificación de la gráfica de barras. Consiste de líneas de puntos que representan las barras y terminan con un punto grande. Son más simples de construir.

II. Gráficas de datos numéricos con una sola variable

A. Histograma

El histograma se utiliza para representar las diferencias en frecuencias absolutas y relativas entre los intervalos o clases de una variable intervalar o de razón.

Es un tipo de gráfica de barras verticales donde el ancho de cada barra

corresponde a los límites de cada clase. Por lo tanto se diferencia de las gráficas de barras de variables categóricas en que las barras son contiguas (se tocan).

Los límites de cada clase aparecen en el eje horizontal y la frecuencia en el vertical. A veces se indica la frecuencia exacta señalándola en la parte superior de la barra.

Cuando se representa un histograma de variables discretas las barras también son contiguas (se tocan) a mitad de camino entre los límites de cada intervalo.

Ejemplo: 20-24, 25-29, 30-34, etc. En estos casos las barras del histograma tienen como límites reales los puntos 24.5, 29.5, 34.5 etc.

En el caso de variables continuas no existe esta situación pues los límites de los intervalos corresponden a los de las barras Ejemplo: $2.0 \leq x < 4.0$; $4.0 \leq x < 6.0$; $6.0 \leq x < 8.0$ intervalos

Los intervalos siempre están ordenados de derecha a izquierda, de mayor a menor, como se espera del eje de x.

Hay histogramas de frecuencia, de proporción o de porciento dependiendo del tipo de distribución que representen.

La forma del histograma se parece mucho a la del diagrama de tallo y hoja cuando los tallos corresponden a los intervalos. El histograma ofrece una muy buena idea visual de la distribución de frecuencias

A veces se utiliza cuando se quieren comparar dos grupos diferentes. Para lograrlo se presentan dos histogramas, uno al lado del otro compartiendo el mismo eje vertical, como lo vimos en los diagramas de tallo y hoja. Esta representación de dos histogramas se hace, a veces, compartiendo el eje horizontal de forma que en un histograma las frecuencias se encuentran hacia arriba y en el otro hacia abajo.

B. Polígono de frecuencia

Como el histograma, también sirve para representar frecuencias absolutas o relativas en los intervalos de una variable intervalar o de razón.

Los límites de los intervalos se indican en el eje horizontal y la frecuencia, proporción o por ciento en el vertical.

La diferencia con respecto al histograma es que el polígono de frecuencias sólo toma en consideración el punto medio como representativo de cada clase.

Construcción:

a. Se colocan los puntos medios de cada clase o intervalo en la parte superior de cada barra del histograma.

b. Se añaden dos puntos medios adicionales correspondientes a un primero y último intervalo inexistentes.

c. Se conectan todos estos puntos medios.

A veces no se añaden dos puntos adicionales, de manera que el polígono de frecuencia se extiende del punto medio del primer intervalo al punto medio del último.

El polígono de frecuencia se utiliza principalmente cuando se comparan dos o más grupos con respecto a la misma variable. Cuando son sólo dos es posible poner un histograma al lado de otro, pero si son más, no hay forma de poder hacerlo. (Transparencia T3.8)

C. Ojiva o polígono de frecuencia acumulada

Corresponde a la tabla de frecuencia acumulada. Puede representar tanto la frecuencia absoluta como la relativa (porciento o proporción)

Cada punto en el eje vertical indica la frecuencia acumulada hasta el límite superior del intervalo.

Construcción:

a. En el límite superior de cada clase se traza el punto que corresponda a la frecuencia, proporción por ciento acumulado hasta ahí.

b. El primer punto que se marca es el límite inferior de la primera clase, que corresponde a 0 %

c. Se conectan todos estos puntos con segmentos.

La ojiva permite la comparación de dos grupos de datos de forma visual y de

manera más efectiva que el polígono de frecuencia. Puesto que con la simple utilización de una regla se puede determinar la frecuencia acumulada que se encuentra por debajo de ciertos valores. (Transparencia T3.11)

Utilizando la gráfica anterior conteste las siguientes preguntas:

¿24% de todos los perros realengos pesan menos de cuánto?

¿Qué por ciento de los perros realengos pesan menos de 9 kg?

¿Cuáles perros están más flacos, los que tienen dueño o los realengos? ¿Qué aspecto de la gráfica te asegura de ello?

III. Gráficas de dos variables numéricas o una numérica y otra categórica)

Estas gráficas se representan en el plano cartesiano. En el caso de que una variable sea categórica y la otra numérica, la categórica se presenta en el eje de x y la numérica en el eje de y.

Es importante observar que en esta situación no se está hablando de frecuencias como ocurría cuando se presentaba una variable categórica o numérica en un eje y su frecuencia en el otro. (Hinkle p.33)

Cuando ambas variables son numéricas la gráfica que se crea se llama un diagrama de dispersión (Transparencia T17.1)

La siguiente gráfica representa las puntuaciones de 30 estudiantes que tomaron una preprueba y una postprueba con puntuaciones desde 200 hasta 800. Cada punto representa un estudiante.

IV. Formas y maneras de crear gráficas engañosas.

Analiza los siguientes ejemplos de gráficas engañosas que se han hecho partiendo de un histograma en el que se representa un aumento muy pequeño de salario de 2001 a 2002.

A. Corte del eje vertical del histograma o polígono de frecuencias para dar la impresión de que el aumento fue mayor

Cuando se construye un histograma o un polígono de frecuencias jamás se debe cortar el eje vertical de las frecuencias. Si esto se hace la gráfica es engañosa. Sin embargo el eje horizontal si se puede cortar y hasta el cero se puede excluir, siempre que todas las clases aparezcan en la gráfica.

B. Estiramiento de los ejes del histograma para dar la impresión de que el aumento fue mayor

_

Si se estira el eje horizontal o se encoge el eje vertical, se puede modificar el impacto visual de la gráfica dando a entender que el cambio ha sido mayor.

C. Ampliación del histograma para dar la impresión de que el aumento fue mayor

A menudo se utilizan barras o figuras para representar frecuencias. El problema con las barras y otras figuras es que los cambios se representan aumentando el volumen total de la figura, no solamente el alto. Esto crea la impresión de que el cambio ha sido mayor.

Actividades y/o asignaciones:

Tomadas de Holcomb, Z.C. (1992). Interpreting statistics. A guide and workbook based on excerpts from journal articles. Los Angeles, CA: Pyrczak Publishing.
1. Construcción del histograma, polígono y ojiva utilizando Excel (en el próximo taller)
2. (Histograma) Young children who drown in hot tubs. Exercise 6. Interpreting statistics. pp.21-23
3. (Polígono) Hand manipulation skills. Interpreting statistics. Exercise 7. pp.25-27

Lecturas recomendadas:

Hinkle capt. 2 pp.31-41
Rodríguez-Esquerdo, unidad 2, pp.47-142
Frankfort-Nachmias, unidad 3, pp.71- 108.

MENU 6390

EDUC 6390: Estadística aplicada en la educación

Andrés Menéndez Raymat, Ph.D.

Conferencia 5

Medidas de localización

Bosquejo

En muchos textos de estadísticas se habla de “medidas de tendencia central” en vez de “medidas de localización”. Pero en esta sección se incluyen las porcentilas que no son medidas de tendencia central. Las porcentilas presentan

un lugar, una localización en la distribución. Lo mismo puede decirse de la moda, media y mediana, las otras medidas que se presentan en la conferencia. Todas ellas representan una localización en la distribución. Por lo tanto el nombre que se ha adoptado para la conferencia es **medidas de localización**.

En muchas ocasiones el conjunto de datos bajo estudio es grande, no sólo en cuanto a la cantidad de sujetos, sino también en términos de la cantidad de variables que se quieren estudiar. En tales casos, no es eficiente utilizar tablas y gráficas para llevar a cabo la comparación entre las variables. Es preferible utilizar ciertas medidas mucho más simples que permiten la comparación. Las medidas de localización son las primeras que permiten hacer eso. Las medidas de dispersión, que se tratan en el próximo capítulo, permiten precisar estas comparaciones.

Las tres medidas de localización más importantes son la moda, la mediana y la media. Se les llama medidas de tendencia central porque son números o categorías que describen lo que es típico o promedio en la distribución.

I. Moda

La moda es la medida de tendencia central más fácil de determinar. Corresponde a la categoría o valor de la variable con la frecuencia mayor (la que aparece más a menudo)

La moda se utiliza principalmente con variables nominales y es la única medida de tendencia central que se puede usar con variables nominales.

A veces no hay moda porque ningún valor se repite. Otras veces hay dos o más modas puesto que varios valores se repiten la misma cantidad de veces.

Ejemplos:

En la distribución (2, 3, 4, 4, 5 8, 12) la MODA es **4**

La distribución (2, 3, 3, 4, 5, 5, 12) es **bimodal** y las modas son **3** y **5**

En la distribución (2, 3, 6, 7, 8, 10, 12) NO HAY MODA

Al igual que la mediana, la moda no se afecta por los valores extremos de la distribución.

II. Mediana

A. Características

La mediana se utiliza principalmente con variables ordinales y junto con la moda son las únicas dos medidas de tendencia central que se puede usar con variables ordinales.

La mediana se define como la puntuación que queda en el medio exacto de la distribución. En términos visuales corresponde a la puntuación en el mero medio, después que todas las puntuaciones han sido colocadas en orden (ascendente o descendente).

El método para determinar la mediana en el caso de variables numéricas depende de si el número de observaciones es par o impar. Si hay un número impar de observaciones, la mediana corresponde al valor que se encuentra en el medio. Pero si el número de observaciones es par, entonces se toman los dos valores que se hallan en el medio de la distribución, se suman y se divide entre dos. Cuando el número de observaciones es impar, la mediana corresponde a un valor de la distribución. Cuando el número de observaciones es par, la mediana no es necesariamente un valor de la distribución

Los empates se cuentan como puntuaciones diferentes.

Ejemplos:

En la distribución (2, 7,**9**,12,15), la mediana = **9**

En la distribución (2,7,9,12,15,20), la mediana = (9+12)/2 = **10.5**

En la distribución (2,7,9,9,15,20), la mediana = (9+9)/2 = **9**

Sin embargo, cuando la variable es ordinal, no es apropiado promediar los dos valores medios. Simplemente se dice que la mediana se encuentra entre esos dos valores

Ejemplos:

1.En un cuestionario que utiliza la escala Likert, las respuestas a una pregunta fueron “nunca, nunca, de vez en cuando, a menudo, muy frecuentemente”.

mediana = **de vez en cuando**

2.En un cuestionario que utiliza la escala Likert, las respuestas a una pregunta fueron “nunca, nunca, de vez en cuando, a menudo”.

mediana = se encuentra entre **"nunca" y "de vez en cuando"**

Una de las características de la mediana es que no se afecta por los valores extremos de la distribución.

Ejemplo:

En la distribución (2, 7, **9**, 12, 15), la mediana = **9**

En la distribución (2, 7, **9**, 12, 245), la mediana = **9**

B. La mediana en frecuencias agrupadas (variable categórica con escala ordinal)

Cuando las observaciones han sido tabuladas en una tabla de distribución de frecuencias, la mediana corresponde a la categoría en la que se encuentra la frecuencia acumulativa del 50% de las observaciones.

Ejemplo:

En la siguiente tabla de frecuencias se observa que la mediana corresponde a la categoría **"algo gordo"** a pesar de que dicha categoría no es la categoría que aparece en el medio de la lista. Esto se debe a que el 50% de la distribución acumulada se encuentra en la categoría "algo gordo"

Categorías	***frecuencias***	***frec. relativa***	***frec. acumulada***
Muy gordo	35	6.6	6.6
Gordo	80	15.0	21.6
Algo gordo	183	34.3	55.9
Peso adecuado	124	23.2	79.1
Algo flaco	69	12.9	92
Flaco	37	6.9	98.9
Muy flaco	6	1.1	100
Total	***534***	***100***	

C. La mediana en frecuencias agrupadas (variable intervalar o de razón)

Cuando la variable es intervalar o de razón y las frecuencias se encuentran agrupadas en una tabla, el proceso de determinación de la mediana es más complicado y se utiliza una fórmula. Hoy día los datos, por lo general, se guardan individualmente en el programa de computadora y en muy raras ocasiones se necesita de esta fórmula.

$$Mediana = L + \left[\frac{n(0.5) - cf}{f}\right]w$$

donde :

L = límite (exacto) inferior del intervalo que contiene la mediana

n = número total de casos

cf = la frecuencia absoluta acumulada en el intervalo anterior al que contiene la mediana

f = frecuencia del intervalo que contiene la mediana

w = ancho del intervalo que contiene la mediana

Ejemplo:

Intervalos	***frecuencias***	***frec. relativa***	***frec. acumulada***
10-19	35	6.6	6.6
20-29	80	15.0	21.6
30-39	183	34.3	55.9
40-49	124	23.2	79.1
50-59	69	12.9	92
60-69	37	6.9	98.9
70-79	6	1.1	100
Total	***534***	***100***	

La mediana en la distribución precedente se obtiene utilizando la fórmula de la siguiente forma:

$$Mediana = 29.5 + \left[\frac{534(0.5) - 105}{183}\right]10 = 38.4$$

Nota:

Debe observar que L [el límite (exacto) inferior del intervalo que contiene la mediana) es 29.5 y que w (el ancho del intervalo] se obtuvo de la resta de los límites reales 19.5 – 9.5 = 10

III. Porcentilas y rango porcentil

A. Definición y propiedades

La mediana es un caso especial de las medidas de localización llamadas porcentilas.

La porcentila es una puntuación o dato en el cual o por debajo del cual se encuentra un porciento específico de la distribución. La porcentila **n** es el dato por debajo del cual (e incluyéndose) se encuentra el n porciento de la población.

Las porcentilas se utilizan a menudo cuando se informan las puntuaciones en las pruebas estandarizadas. Le permiten al examinado determinar qué porciento de la población de examinados se encuentra por debajo de él.

La porcentila es un dato

Ejemplo:

En la siguiente tabla

Puntuación	*frec.*	*frec. acum.*
24	1	1
25	1	2
26	0	2
27	0	2
28	0	2
29	1	3
30	1	4
31	0	4
32	2	6
33	3	9

34	1	10
35	**2**	**12**
36	4	16
37	5	21
38	4	25
39	3	28
40	4	32
41	**5**	**37**
42	5	42
43	4	46
44	4	50
45	7	57
46	9	66
47	7	73
48	8	81
49	**11**	**92**
50	7	99
51	3	102
52	6	108
53	7	115
54	7	122
55	12	134
56	14	148
57	6	154
58	2	156
59	3	159
60	2	161
61	1	162
62	3	165
63	**5**	**170**
64	4	174
65	2	176
66	0	176
67	1	177
68	2	179
69	1	180
Total	***N = 180***	

¿Qué valor o puntuación es la porcentila 20? ¿Qué valor o puntuación tiene el 20% de los datos por debajo (incluyéndose él mismo)?

$$\frac{x}{180} = \frac{20}{100}$$

x = 3600/100 = 36

Este 36 corresponde al número de datos (frecuencia acumulada) comenzando por el valor más pequeño. El valor 41 tiene las frecuencias 33, 34, 35, 36 y 37. Por lo tanto

$P_{20} = 41$

porque hay 36 puntuaciones por debajo (incluyendo el 41) de un total de 180 puntuaciones.

Ejemplo:

$P_{50} = ?$

¿Qué valor o puntuación es la porcentila 50? ¿Qué valor o puntuación tiene el 50% de los datos por debajo (incluyéndose él mismo)?

$$\frac{x}{180} = \frac{50}{100}$$

x = 9000/100 = 90

Este 90 corresponde al número de datos (frecuencia acumulada) comenzando por el valor más pequeño. El valor **49** tiene las frecuencias 82,83,84,85,86,87,88, 89,90, 91,92. Por lo tanto

$P_{50} = 49$

porque hay 90 puntuaciones por debajo (incluyendo el 49) de un total de 180 puntuaciones.

El **rango porcentil** es la **posición** que ocupa un dato. El rango porcentil de un valor dado se determina hallando el por ciento de datos con valores iguales o inferiores al dato dado.

Ejemplo:

El rango porcentil de 35 es 6.7, puesto que hay **12** datos con valor igual o menor de 35 y 12 es el 6.7 por ciento de 180.

$$\frac{12}{180} = \frac{x}{100}$$

x = 1200/180 = 6.7

De igual manera el rango porcentil de 63 es 94.4 puesto que hay 170 datos en por debajo de 63.

$$\frac{170}{180} = \frac{x}{100}$$

x = 1700/180 = 94.4

170 es el 94.4% de 180

Para determinar porcentilas y el rango porcentil se utiliza Excel. (Excel utiliza una definición semejante, pero no exactamente igual a esta definición intuitiva)

La ojiva nos permite pasar de la porcentila al rango porcentil y viceversa de forma visual.

Ejemplo:
Determina en la ojiva anterior
1. ¿Cuál es el rango porcentil de los estudiantes que obtuvieron 36 puntos en la prueba?
2. ¿Cuál es el rango porcentil de los estudiantes que obtuvieron 63 puntos en la prueba?
3. ¿Cuál es la porcentila 20 ?
4. ¿Cuál es la porcentila 60 ?

B. Propiedades de los rangos porcentiles

1. Sirven para comparar valores en una distribución en términos de posición. Un estudiante con un rango porcentil de 70 está en mejor posición que uno con un rango porcentil de 50.

2. Esta comparación es categórica pues una diferencia en rango porcentil no representa una diferencia semejante en puntuaciones crudas.

3. Por esta razón los rangos porcentiles no se pueden sumar o restar.

4. En una distribución normal la diferencia en rangos porcentiles en el centro de la distribución representa poca diferencia en valores crudos, pero la misma diferencia en los extremos de la distribución significa mucha diferencia.

Ejemplo:

En la gráfica que aparece en la p. 60 de Hinkle observe que la diferencia entre P_{40} y P_{20} es mucho mayor que la diferencia entre P_{40} y P_{60} a pesar de que la diferencia es siempre de 20 puntos porcentiles.

$P_{40} - P_{20} = 45 - 36 = 11$

$P_{60} - P_{40} = 51 - 45 = 9$

C. Las porcentilas y los rangos porcentiles en una tabla de distribución de frecuencias agrupadas (variable intervalar o de razón)

Cuando la variable es intervalar o de razón y las frecuencias se encuentran agrupadas en una tabla el proceso de determinación de las porcentilas y los rangos porcentiles es más complicado y por lo general se utilizan fórmulas. Hoy día los datos por lo general se guardan individualemente en el programa de computadora y en muy raras ocasiones se necesita de estas fórmulas. (Estas fórmulas no se trabajarán en la clase. Se incluyen aquí como referencia exclusivamente)

Fórmula para hallar la porcentila en una distribución de frecuencias agrupadas

$$porcentila = L + \left[\frac{np - cf}{f}\right]w$$

donde :

L = límite (exacto) inferior del intervalo que contiene la porcentila deseada

n = número total de casos

p = proporción correspondiente a la porcentila deseada

cf = la frecuencia absoluta acumulada en el intervalo anterior al que contiene la mediana

f = frecuencia del intervalo que contiene la porcentila

w = ancho del intervalo que contiene la porcentila

Ejemplo:

Halla la porcentila 20 en la siguiente distribución:

Intervalos	***frecuencias***	***frec. relativa***	***frec. acumulada***
10-19	35	6.6	6.6
20-29	80	15.0	21.6
30-39	183	34.3	55.9
40-49	124	23.2	79.1
50-59	69	12.9	92
60-69	37	6.9	98.9
70-79	6	1.1	100
Total	***534***	***100***	

$$P_{20} = 19.5 + \left[\frac{534(0.2) - 35}{80}\right]10 = 28.48$$

Nota:

Debe observar que L [el límite (exacto) inferior del intervalo que contiene la porcentila 20] es 19.5 y que w (el ancho del intervalo) se obtuvo de la resta de los límites reales 19.5 – 9.5 = 10

Esta fórmula de la porcentila es muy semejante a la que se utiliza para hallar la mediana en una tabla de distribución de frecuencias agrupadas.

Fórmula para hallar el rango porcentil en una distribución de frecuencias agrupadas

$$RP_X = \left[\frac{cf + \frac{X - L}{w} \otimes f}{n} \right] (100)$$

donde :

X = puntuación para la cual se busca el rango porcentil

L = límite (exacto) inferior del intervalo que contiene la puntuación dada

n = número total de casos

p = proporción correspondiente a la porcentila deseada

cf = la frecuencia absoluta acumulada en el intervalo anterior al que contiene la puntuación dada

f = frecuencia del intervalo que contiene la puntuación

w = ancho del intervalo que contiene la porcentila

Ejemplo:

Halla el rango porcentil de 38 en la siguiente distribución:

Intervalos	***frecuencias***	***frec. relativa***	***frec. acumulada***
10-19	35	6.6	6.6
20-29	80	15.0	21.6
30-39	183	34.3	55.9
40-49	124	23.2	79.1
50-59	69	12.9	92
60-69	37	6.9	98.9
70-79	6	1.1	100
Total	***534***	***100***	

$$RP_{38} = \left[\frac{115 + \frac{38 - 29.5}{10} \otimes 183}{534} \right] (100) =$$

$RP_{38} = 50.66$

Nota:

Debe observar que L [el límite (exacto) inferior del intervalo que contiene la puntuación 20] es 29.5 y que w (el ancho del intervalo) se obtuvo de la resta de los límites reales 19.5 – 9.5 = 10

IV. Cuartilas

Las cuartilas son otra medida de uso común en estadística.

"Q_1" o primera cuartila corresponde a la porcentila 25.

"Q_2" o segunda cuartila o mediana corresponde a la mediana o porcentila 50.

"Q_3" o tercera cuartila corresponde a la porcentila 75.

Para hallar las cuartilas se determina primero la posición utilizando la fórmula posicional de las porcentilas que se ha presentado anteriormente en la conferencia.

(Q_1)Primera cuartila corresponde a P_{25} . Corresponde a la puntuación (n + l) / 4

Segunda cuartila o mediana corresponde a P_{50} . Corresponde a la puntuación (n + l) / 2

(Q_3) Tercera cuartila corresponde a P_{75} . Corresponde a la puntuación 3 (n + l) / 4

Sólo que

a. si el resultado de la fórmula posicional es un **entero**, se usa la puntuación

correspondiente a esa posición

b. si el resultado de la fórmula posicional termina en **.5**, se usa la puntuación promedio como se hace con la mediana

c. si el resultado de la fórmula posicional termina en otro decimal, se redondea el resultado y se usa la puntuación correspondiente.

(*En otras definiciones, para ser más exacto se interpola. Pero como Excel hace eso por nosotros, no vale la pena complicarse la vida*)

Ejemplo:

Halla las tres cuartilas en la muestra siguiente: n = 6

2.1 ; 3.4 ; 4.2 ; 5.6 ; 7.8 ; 9.0

Q_1: (6+1)/4 = 1.75 y por lo tanto se utiliza la **segunda** puntuación

Q_2: (6+1)/2 = 3.5 y por lo tanto se utiliza la **puntuación entre la tercera y la cuarta**

Q_3: 3(6+1)/4= 5.25 y por lo tanto se utiliza la puntuación en la **quinta** posición

Por lo tanto

Q_1 = 3.4

mediana = (4.2 + 5.6)/2

Q_3 = 7.8

Para hallar las cuartilas también se determina la posición utilizando una forma intuitiva cuando son pocos valores.

Se escoge la mediana como el valor en el medio exacto de la distribución. Como la media separa la distribución en dos grupos del mismo tamaño, se seleccionan como primera cuartila y tercera cuartila las medianas de dichos grupos

Ejemplo:

Halla la mediana y las cuartilas de la siguiente muestra:

2 ; 2 ; 3 ; 4 ; **5** ; 6 ; 8 ; 8 ; 8

La mediana es 5, pues se encuentra en el medio exacto

Q_1 es 2.5 pues se encuentra en el medio del primer grupo (entre 2 y 3)

Q_3 es 8 pues se encuentra en el medio del segundo grupo (entre 8 y 8)

Ejemplo:

Halla la mediana y las cuartilas de la siguiente muestra:

2 ; 2 ; 3 ; 4 ; 8 ; 8 ; 9 ; 9

La mediana es 6, pues se encuentra en el medio entre 4 y 8

Q_1 es 2.5 pues se encuentra en el medio del primer grupo (entre 2 y 3)

Q_3 es 8.5 pues se encuentra en el medio del segundo grupo (entre 8 y 9)

V. La gráfica de caja y bigote (box plot)

La gráfica de caja y bigote provee visualmente una cantidad considerable de información sobre la distribución. Con las medidas de localización que se han presentado hasta ahora es posible construirla, pero el total de la información que puede ofrecer no quedará claro hasta que se hayan visto las medidas de dispersión y las formas de las distribuciones.

La gráfica de caja y bigote se construye utilizando el valor mínimo, Q_1, la mediana, Q_3 y el valor máximo.

Está formado por:

1. una caja que se extiende desde la primera hasta la tercera cuartila

2. un segmento sólido vertical que marca la mediana dentro de la caja

3. dos segmentos horizontales (bigotes) que se extienden a derecha e izquierda desde la primera y la tercera cuartila hacia los valores mínimo y máximo de la distribución

Estas gráficas también se pueden representar verticalmente. El programa SPSS permite cambiar la orientación de estas gráficas dependiendo de nuestra preferencia.

(En la clase se podría discutir un ejemplo utilizando la Transparencia T4.6 que se refiere a la siguiente muestra de la matrícula en 6 colegios universitarios de Pennsylvania: 4.9 ; 6.3 ; 7.7 ; 8.9 ; 10.3 ; 11.7.

Las medidas que se necesitan para construir la gráfica de caja y bigote son las siguientes:

Q_1 = 6.3 mediana= 8.3 Q_3 = 10.3 L = 4.9 H = 11.7

VI. Media

A. Definición y propiedades

La media aritmética es la medida de tendencia central más conocida. La mayor parte de la gente la llama el promedio. Se puede utilizar solamente con variables intervalares o de razón. Esto se debe a que en su cómputo es necesario usar suma y división. Estas operaciones sólo tienen sentido con valores numéricos.

En una muestra el símbolo de la estadística de la media es

pero en una población el parámetro se indica por medio de la letra griega μ .

La media corresponde a la suma de todas las observaciones dividida por el número de observaciones

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

donde X_i representa los valores y n la cantidad de valores.

Ejemplo:

Las siguientes son tres muestras de la matrícula en 6 colegios universitarios de tres estados de EEUU

Pennsylvania:

4.9 6.3 7.7 8.9 10.3 11.7

Texas:

4.9 6.4 6.4 8.5 11.6 12.0

Carolina del Norte:

7.6 7.9 8.3 8.3 8.7 9.0

La media es 8.3 en cada caso

Propiedades de la media

1. Como la media requiere de la suma y la división para su cómputo, sólo puede determinarse para variables intervalares o de razón.

2. La media es el centro de gravedad de la distribución.

Visualmente se puede interpretar como el fiel de una balanza en la que se han puesto todas las puntuaciones.

Esta propiedad, matemáticamente puede verse como el hecho de que **la suma de todas las desviaciones con respecto a la media es cero**

$$\sum\left(X_i - \overline{X}\right) = 0$$

Ejemplo:

En la muestra de matrícula de Pennsylvania los valores son:

4.9 6.3 7.7 8.9 10.3 11.7

Por lo tanto, la suma de las desviaciones con respecto a la media es:

(4.9-8.3) + (6.3-8.3) + (7.7-8.3) + (8.9-8.3) + (10.3-8.3) + (11.7-8.3) = 0

3. Por esta misma razón de ser el centro de gravedad, la media puede verse muy afectada por los valores extremos de la distribución

Si en el ejemplo de la muestra de universidades del estado de Carolina del Norte uno de los valores fuera 30.0 en vez de 9.0

7.6 7.9 8.3 8.3 8.7 30.0

La media pasaba a ser 11.8 y dejaba de ser representativa del grupo como tal. El ejemplo de Hinkle (p.67) señala que con el salario del presidente de una compañía se obtiene una idea errónea de los salarios en la compañía.

En estos casos en que la media no es un buen descriptor de la muestra es conveniente recurrir a la mediana.

(Actividad de media y mediana)

4. La suma de los cuadrados de las desviaciones con respecto a la media es menor que con respecto a cualquier otro valor de la distribución

$$\sum\left(X_i - \overline{X}\right)^2$$

Ejemplo:

En Pennsylvania (ejemplo anterior) donde los valores son:

4.9 6.3 7.7 8.9 10.3 11.7

Si a cada puntuación de la muestra se le resta la media, el resultado es siempre menor que si se resta a cada puntuación cualquier otro valor de la distribución:

$(4.9-8.3)^2 + (6.3-8.3)^2 + (7.7-8.3)^2 + (8.9-8.3)^2 + (10.3-8.3)^2 + (11.7-8.3)^2 = 31.84$

$(4.9-8.9)^2 + (6.3-8.9)^2 + (7.7-8.9)^2 + (8.9-8.9)^2 + (10.3-8.9)^2 + (11.7-8.9)^2 = 34$
etc.

B. La media de dos grupos

Cuando se combinan dos grupos el proceso de combinar sus medias es un poco más complejo de como puede aparecer a primera vista.

Puesto que puede haber más sujetos en un grupo que en el otro se debe dar peso a la media de cada grupo dependiendo de la cantidad de valores que tiene. Por lo tanto se multiplica la media de cada grupo por la cantidad de valores del grupo, se suma y luego se divide el total obtenido entre el número de total de puntuaciones en el grupo.

$$\overline{X} = \frac{n_1\overline{X_1} + n_2\overline{X_2}}{n_1 + n_2}$$

Ejemplo:

106 mujeres y 74 hombres tienen medias de 45.26 y 54.89 en el examen final. Por lo tanto la media del grupo es:

$$\overline{X} = \frac{(106)(45.26) + (74)(54.89)}{106 + 74}$$

En ciertos momentos la media no es un buen representante de la población y es conveniente acudir a la mediana.

Actividades y/o asignaciones:

1. Actividad: Media y mediana
2. Exercise 10: Mean and Median. Students' part-time work, GPA, and days absent from school. (Real Data, pp.33-35)
3. Exercise 8: (Cumulative percentage and percentile ranks). Spelling component test. (Interpreting basic statistics, pp.29-31)
4. Hinkle pp.84-88 ej. 6a, 6b, 7a, 10a, 12
5. Hinkle p. 50 hacer la gráfica de caja y bigote en los ejercicios #1,2,3 y señalando los valores atípicos, si los hubiese.

Lecturas recomendadas:

Hinkle capt. 3 pp.52-72
Rodríguez-Esquerdo, cap. 2, pp.143-234
Frankfort-Nachmias & Leon-Guerrero, capítulo 4, pp.109-148
Sirkin, capítulo 4, pp.81-122.

MENU 6390

EDUC 6390: Estadística aplicada en la educación
Andrés Menéndez Raymat, Ph.D.
Conferencia 6
Medidas de dispersión
Bosquejo

Las medidas de tendencia central permiten describir una distribución por medio de sus valores típicos. Sin embargo estas medidas son sólo parte de la información que se puede obtener de la distribución. A menudo, al conformarse la descripción a una medida de tendencia central se cae en la sobresimplificación y el estereotipo. Hacen falta otras medidas que reflejen la variedad y la multiplicidad. Estas medidas que hablan de las diferencias y la diversidad son las medidas de dispersión.

Ejemplo: Dos grupos de estudiantes toman un mismo examen y ambos grupos obtienen una media o promedio de C. Pero en uno de los grupos las 15 personas que tomaron la prueba obtuvieron una nota de C. En el otro grupo las notas de las 15 personas fueron desde A hasta F. Con la presentación de media exclusivamente no se describe adecuadamente ninguno de los dos grupos que, obviamente, son muy diferentes.

De nuevo la escala de medición de la variable va a ser la clave en la selección que se haga en el estudio sobre la medida de dispersión adecuada para la situación.

I. Datos categóricos

El Índice de Variación Cualitativa

(Index of Qualitative Variation or **IQV**)

El IVC se utiliza para medir la variación de variables nominales. En EEUU se utiliza mucho para medir la diversidad racial y cultural en escuelas y organizaciones.

El IVC varía desde 0 hasta 1. El cero se obtiene cuando todos los casos de la distribución están en una sola categoría. El 1 se obtiene cuando los casos están distribuidos equitativamente en todas las categorías.

Definición

El IVC es la razón entre el número total de diferencias reales en la distribución y el número máximo de posibles diferencias en la misma distribución.

Cómputo del IVC

En un grupo de cuatro personas hay tres cristianos (Pedro, Jesús y María), un judío (Neftalí) y un musulmán (Alí).

Las diferencias en religión se obtienen formando todos los pares posibles de religiones diferentes: Pedro y Neftalí, Pedro y Alí, Neftalí y Alí, Jesús y Neftalí, Jesús y Alí, María y Neftalí, María y Alí. Hay **siete** pares posibles

1. El número total de diferencias se obtiene sumando todos estos pares. Esto se puede simplificar matemáticamente multiplicando la frecuencia de cada categoría por la frecuencia en cada una de las otras categorías y sumando cada producto: Como hay 3 cristianos, 1 judío y 1 musulmán, los productos son

$$(3x1) + (3x1) + (1x1) = 7$$

La fórmula se expresa como

2. El máximo de posibles diferencias se interpreta como el hecho de que en cada categoría haya el mismo número de personas.

La fórmula que se emplea para determinarlo es

donde

K = número de categorías en la distribución

N = número total de casos en la distribución

El número total de diferencias en el ejemplo anterior es

3. El IVC se obtiene dividiendo el número total de diferencias entre el máximo posible de diferencias. Se utiliza la siguiente fórmula:

El IVC se puede representar como un porciento al transformar la proporción en porciento cuando se multiplica por 100.

En el ejemplo anterior

Actividad:

Calcula el IVC de una organización estudiantil universitaria que cuenta con 10 estudiantes PPD, 8 PNP y 15 PIP

Número total de diferencias (10x8) + (10x15) + (8x15) = 350

Máximo posible de diferencias $[(3x2)/2]\ [33/3]^2 = 3 \times 121 = 363$

IVC = 350/363 = 0.964 = 96.4%

II. Datos numéricos

A. Amplitud o alcance **(range)**

La amplitud o alcance (no debe usarse el anglicismo rango que significa otra cosa) se utiliza con variables intervalares o de razón. Es la medida de dispersión más fácil de obtener. Se obtiene hallando la diferencia entre la observación mayor y la menor (el valor máximo menos el valor mínimo). Hinkle le añade 1, pero nosotros no lo haremos.

El alcance es una medida muy influenciada por los valores extremos y por lo tanto puede dar una impresión falsa sobre los valores reales de la distribución.

Ejemplo: En una distribución los valores que se obtienen son 2.1 3.4 4.2 5.6 7.8 9.0 ¿Cuál es la amplitud de la muestra?

El alcance o amplitud es (9.0 - 2.1) = 6.9

Ejemplo: Si en la distribución anterior los valores fueran 2.1 3.4 4.2 5.6 7.8 52.1

El alcance o amplitud es (52.1 - 2.1) = 50

Realmente las dos distribuciones se diferencian solamente por un dato con un valor extremo que en la segunda distribución da una impresión falsa de los otros valores.

B. Amplitud intercuartila

1. Definición

Para evitar la descripción errónea de los datos cuando en la distribución hay valores atípicos se ha diseñado otra medida de dispersión llamada la amplitud intercuartila.

Se ha definido como la diferencia entre Q_3 y Q_1

Ejemplo:

En la distribución 2.1 3.4 4.2 5.6 7.8 9.0

La amplitud intercuartila es (7.8 - 3.4) = 4.4

2. Valores atípicos (outliers)

La definición de valor atípico depende mucho de la distribución y del investigador. Para determinarlo es preciso situar los valores en el contexto de la gráfica de caja y bigote.

Muchos autores definen un valor como atípico cuando éste se encuentra a una distancia mayor de 1. 5 de la amplitud intercuartila de Q_1 (hacia la izquierda) o

de Q_3 (hacia la derecha). En términos visuales esto quiere decir que todo valor que se encuentre en los bigotes a una distancia mayor de caja y media de la cuartila más cercana es un valor atípico.

Algunos autores prefieren tomar como distancia para determinar los valores atípicos 2 veces y hasta 2.5 veces la amplitud intercuartila. En esta clase se utilizará l.5 veces la amplitud intercuartila como la distancia para determinar los valores atípicos.

Ejemplo:

Determina si hay algún valor atípico en la siguiente distribución

0.1 3.4 4.2 5.6 7.8 19.0

La primera y tercera cuartila son:

$Q_1 = 3.4$ $Q_3 = 7.8$

Por lo tanto la amplitud intercuartila es (7.8 - 3.4) = 4.4

Una y media veces la amplitud intercuartila es (4.4) x (1.5) = 6.6

Por lo tanto cualquier valor mayor de 7.8 + 6.6 = 14.4 es un valor atípico.

Por lo tanto 19.0 es un valor atípico.

A veces en los diagramas de caja y bigote el bigote llega solamente hasta **el último valor dentro de los límites razonables** (reasonable lower and upper bounds)(RLB y RUB). Los valores atípicos aparecen como puntos fuera del bigote.

Nota: Debe tenerse cuidado de no llevar los bigotes hasta los límites razonables. Los límites razonables, por lo general, no son valores reales de la distribución, sino límites matemáticos obtenidos para determinar qué es un valor atípico.

C. Desviación media

Es la suma de los valores absolutos de la diferencia entre cada valor y la media; dividido todo por la cantidad de observaciones. Esta medida generalmente no

se usa, pues resulta difícil trabajar matemáticamente con valores absolutos.

Mientras mayor es la dispersión de los valores, mayor es la desviación media.

D. Varianza

Esta medida refleja cuánto, en promedio, cada puntuación de la distribución se desvía de la media.

En una muestra el símbolo que se usa es $\mathbf{s^2}$ y en una población es σ^2

Es un promedio de los cuadrados de las diferencias entre cada valor y la media.

Se puede escribir como

donde la media es

n = tamaño de la muestra

X_i = i^{th} valor de la variable

Debe notarse que el denominador no es n para la muestra sino **n-1.** Para una población, sin embargo, el denominador si es **n**

Esta diferencia entre la varianza de la muestra y de la población se debe a que el promedio de las varianzas de todas las muestras de un tamaño dado es igual a la varianza de la población solamente si en las muestras se usa n-1. La importancia de esto se verá más tarde cuando se describa lo que es un "unbiased estimate"

Se pueden usar dos fórmula básicas que presenta Hinkle, pero ya se inventaron las computadoras y las calculadoras.

Sin embargo es importante recordar que **la varianza es un promedio de desviaciones de cada valor con respecto a la media**.

Todo estudiante es responsable de saber hallar la varianza y la desviación estándar con su calculadora.

E. Desviación estándar

En una muestra el símbolo es **s** y en una población es σ

La desviación estándar es la raíz cuadrada de la varianza.

1. Cómputo de la varianza y la desviación estándar

a. Se obtiene la diferencia entre cada valor y la media.

b. Se cuadra cada diferencia

c. Se suman los cuadrados

d. Se divide la suma por n-1

Ejemplo: 4.9 6.3 7.7 8.9 10.3 11.7

$s^2 = 6.368$

$s = 2.523$

Nótese que ni la varianza ni la desviación estándar pueden ser negativas. La varianza es una suma de cuadrados y la desviación estándar es un radical.

s = 0 sólo cuando todas las puntuaciones tienen el mismo valor que no es otro que el valor de la media.

2. ¿Qué indican la varianza y la desviación estándar?

Ambas muestran cuán separadas están las puntuaciones de la media. Mientras más grandes son estas medidas, más dispersión hay.

La desviación estándar se prefiere a la varianza pues usa la misma unidad de las observaciones. No tiene sentido hablar de unidades cuadradas.

La varianza cuadra la diferencia entre cada valor y la media, pues si no lo hiciera la suma de las diferencias sería cero.

Ejemplo: 4.9 6.3 7.7 8.9 10.3 11.7

(4.9 - 8.3)+ (6.3 - 8.3) + (7.7 - 8.3) + (8.9 - 8.3) + (10.3 - 8.3) + (11.7 - 8.3) = 0

En el proceso de cuadrar se preservan estas diferencias

Actividades y/o asignaciones:

1. Hinkle pp.84-88 ej. 7b, 9, 10b, 13
2. Exercise 13: Mean and standard deviation (Population estimate from a sample). Effects of Progressive and sex-object images of women in advertisements. (Real Data, pp.43-45)
3. Exercise 13: (Means and standard deviations for two groups). Mentoring among men and women. (Interpreting basic statistics, pp.49-51)

* Es importante hablar de la normal y hacer el ejercicio 7 antes de asignar este último ejercicio.

Lecturas recomendadas:

Hinkle capt. 3 pp.73-79
Rodríguez-Esquerdo, cap. 2, pp.235-286
Frankfort-Nachmias, unidad 5, pp.149- 199.

MENU 6390

EDUC 6390: Estadística aplicada en la educación

Andrés Menéndez Raymat, Ph.D.

Conferencia 7

Forma de la distribución

Bosquejo

I. Distribuciones simétricas y distribuciones sesgadas

II. La relación entre la gráfica de la distribución y las medidas de tendencia central y dispersión

A. Distribuciones unimodales cuando la media, la moda y la mediana coinciden

B. Distribuciones unimodales cuando la media, la moda y la mediana no coinciden

III. Relación entre las gráficas de dos distribuciones con medidas de tendencia central y dispersión iguales o diferentes

A. Dos distribuciones con desviación estándar igual, pero medias diferentes

B. Dos distribuciones con medias iguales pero desviación estándar diferentes

IV. La relación entre la gráfica de la distribución y la gráfica de caja y bigote.

Las distribuciones pueden describirse según la forma que asume su gráfica. Cuando se construye el polígono de frecuencias la gráfica tiene una forma que puede ser **simétrica** o **asimétrica** (skewed).

I. Distribuciones simétricas y distribuciones

sesgadas

Se dice que la distribución es simétrica si se puede dividir en dos mitades que parecen ser la imagen una de la otra. En estos casos las frecuencias en los extremos de la distribución son idénticas. La gráfica puede tener diferentes formas. Una de estas formas es la de campana.

Otra forma es la rectangular

Si la distribución tiene algunos valores extremos muy bajos, entonces en la gráfica se nota una cola larga y fina hacia la izquierda de la distribución y se dice que la distribución está sesgada negativamente o que tiene un sesgo a la izquierda.

Si la distribución tiene algunos valores extremos altos, entonces en la gráfica se nota una cola larga y fina hacia la derecha de la distribución y se dice que la distribución está sesgada positivamente o que tiene un sesgo a la derecha.

II. La relación entre la gráfica de la distribución y las medidas de tendencia central y dispersión

A. Distribuciones unimodales cuando la media, la moda y la mediana coinciden

En distribuciones unimodales cuando la media, la moda y la mediana coinciden la distribución es simétrica.

Ejemplo

(Hinkle, p. 67 fig.3.3)**.** La media, mediana y moda coinciden en la distribución

B. Distribuciones unimodales cuando la media, la moda y la mediana no coinciden

En distribuciones unimodales cuando la media, la moda y la mediana no coinciden la distribución es sesgada.

Si la media es mayor que la mediana (la media a la derecha de la mediana) entonces la distribución está sesgada a la derecha (positivamente)

Ejemplo

(Hinkle, p. 67 fig.3.3)

Si la media es menor que la mediana (la media a la izquierda de la mediana) entonces la distribución está sesgada a la izquierda (negativamente)

Ejemplo

(Hinkle, p. 67 fig.3.3)

Nota
En estos casos la media siempre está más cerca del sesgo que la mediana.

III. Relación entre las gráficas de dos distribuciones con medidas de tendencia central y dispersión iguales o diferentes

A. Dos distribuciones con desviación estándar igual, pero medias diferentes

Si dos distribuciones tienen la misma desviación estándar, pero medias diferentes; entonces van a tener la misma forma. La diferencia consiste en que se encuentran desplazadas a lo largo del eje de x. (Hinkel p.232, fig. 10.2)

B. Dos distribuciones con medias iguales pero desviación estándar diferentes

Si dos distribuciones tienen la misma media, pero sus desviaciones estándar son diferentes; entonces se diferencian en que la que tiene la desviación estándar más pequeña tiene los valores más concentrados alrededor de la media y por lo tanto es más "alta". (Hinkel p.43, fig. 2.13)

Ejercicio

¿Cuál de las dos distribuciones (Hinkle p.43, fig. 2.13) tiene la desviación estándar mayor?

IV. La relación entre la gráfica de la distribución y la

gráfica de caja y bigote.

a. Si ambas partes de la caja son iguales (la mediana en el medio de la caja) y los dos bigotes también son iguales, aunque algo más largos que las partes de la caja entonces la distribución tiene tipo de campana (bell shaped distribution)

b. Si los bigotes son diferentes y la mediana no se encuentra en el medio de caja entonces la distribución está sesgada. Negativamente, si el bigote y la parte de la caja largos se encuentran a la izquierda. Positivamente, si el bigote y la parte de la caja largos se encuentran a la derecha.

d. Si los bigotes y las partes de la caja son todos del mismo largo, entonces la distribución es rectangular o uniforme. Tiene la misma frecuencia en cada uno de sus valores.

e. Si los bigotes son cortos y la caja muy larga la distribución tiene forma de U, con mucha concentración de valores en los extremos.

Actividades y/o asignaciones

1. Hinkle p. 86 # 11

Lecturas recomendadas

Hinkle capt. 2 pp. 41-43;Capt. 3 pp. 66-68
Frankfort-Nachmias, unidad 4, pp.130- 136; unidad 5, pp.170-172.

MENU 6390

EDUC 6390: Estadística aplicada en la educación

Andrés Menéndez Raymat, Ph.D.

Conferencia 8

Puntuaciones z y la distribución normal

Bosquejo

I. Puntuaciones estandarizadas o puntuaciones z

En muchas ocasiones se quieren comparar puntuaciones que pertenecen a dos distribuciones diferentes y no se hace posible porque las medias y las desviaciones estándar de las dos distribuciones son diferentes.

Ejemplo:

En qué asignatura salió mejor Pepe si sabemos lo siguiente:

Asignatura	Puntuación	Promedio	s
Psicología	68	65	6
Mat	77	77	9
Historia	83	89	8

Una manera aproximada de resolver el problema es ver si la puntuación está por arriba o por debajo del promedio. Pero esto es sólo aproximado. Si las puntuaciones están todas bajo el promedio o sobre el promedio no se hace fácil identificar cual es mejor a simple vista.

La dificultad radica principalmente en que la media y la desviación estándar son diferentes para cada prueba.

Sin embargo la comparación se hace posible en términos de las puntuaciones z o puntuaciones estandarizadas. Las puntuaciones estandarizadas se obtienen restando la media a cada observación y dividiendo entre la desviación estándar

Donde μ_X es la media de la distribución y σ_X es su desviación estándar.

En el ejemplo tenemos:

Asignatura	Puntuación	Promedio	s	z
Psicología	68	65	6	.5
Mat	77	77	9	0
Historia	83	89	8	-.75

Por lo tanto salió mejor en Psicología y peor en Historia. (Nótese que el orden

puede cambiar con respecto a las puntuaciones crudas)

B. Propiedades

a. La forma de la distribución de las puntuaciones z es idéntica a la de la distribución original.

b. La media es siempre igual a cero

c. La desviación estándar es siempre igual a 1

Por lo tanto al calcular las puntuaciones z transformamos la distribución original en una con $\mu = 0$ y $\sigma = 1$, pero con forma idéntica a la distribución original.

Estas propiedades de las puntuaciones z se van a utilizar para trabajar con distribuciones normales

Actividad con asignación de puntos extra: Halla la media y la desviación estándar de todos los valores z en la siguiente distribución (Hinkle pg. 80 Table 3.6). Haz ambos histogramas en Excel y compara la forma de la distribución de ambos. Determina las propiedades de la nueva distribución de valores z.

Sujetos	Puntuaciones crudas	Puntuaciones z
A	10	1.26
B	9	0.94
C	3	-0.94
D	10	1.26
E	9	0.94
F	2	-1.26
G	2	-1.26
H	10	1.26
I	5	-0.31
J	5	-0.31
K	1	-1.57
L	6	0.00
M	8	0.63
N	6	0.00
O	6	0.00
P	1	-1.57
Q	3	-0.94

R	6	0.00
S	10	1.26
T	8	0.63
N = 20		
Media	6	
Desviación estándar	3.18	

C. Transformaciones

A pesar de que permiten comparar puntuaciones en diferentes exámenes, las puntuaciones z tienen el inconveniente de poder ser negativas y expresarse con decimales (generalmente dos decimales). En el público esto no es muy aceptable.

Por lo tanto varias empresas transforman estos valores z en otros más aceptables que no los cambian en su verdadera significación.

Para eso usan la siguiente transformación:

$$x' = (s)(z) +$$

Donde

s = desviación estándar

= promedio

College Board usa 100 y 500. Muchas otras compañías usan 10 y 50.

Lectura:

Hinkle capt. 3 pp.79-84

Asignación:

Hinkle pp.87-88 #16,17, 20 (sin describir el proceso),21,22
Interpreting Basic Statistics Ex 15 The Stanford Binet Fourth Edition.

II. Modelos matemáticos

Cuando el fenómeno que se quiere trabajar se ha observado bien se puede

crear un MODELO MATEMÁTICO que lo describa. Esta es la función más importante de las matemáticas, facilitar el estudio de la naturaleza con modelos que la interpretan. Por ejemplo, la forma de decir que un grupo de cosas miden el doble de otro grupo se expresa por medio de la función $y = 2x$

Lo mismo sucede con las distribuciones. Hay algunas para las que sus modelos matemáticos son de mucha utilización y provecho en las estadísticas. Existen modelos tanto para las para variables discretas como para las variables continuas. De entre todos veremos el ejemplo por excelencia, el modelo de la función normal para una variable continua.

III. Distribución normal

A. El modelo matemático

Cuando la variable es continua el modelo que más se usa es la distribución normal.

Cuando la distribución es normal siempre es posible pasar de la porcentila al rango porcentil por medio de la transformación al valor z y el uso de la tabla de la distribución normal. Para hacerlo hay que visualizar siempre el área debajo de la curva como el porciento de puntuaciones de la distribución en un intervalo dado.

1. Importancia de la distribución normal (Hinkle, p.94)

Es el modelo matemático por excelencia en muchas situaciones de la vida real. Es la base de la inferencia estadística

2. Propiedades de la distribución normal

a. Es simétrica, y tiene forma de campana

Al ser simétrica la media, la moda y la mediana coinciden.

b. Las medidas de tendencia central son iguales

c. Q_1 y Q_3 están situados a 2/3 de una desviación estándar.

El 68 % del área de la curva (probabilidad) se encuentra a una desviación estándar de la media.

d. La variable tiene un alcance infinito.

3. El modelo matemático

La función matemática que se usa como modelo es (Transparencia):
Transparencia de la fórmula (Hinkle p.91)

donde:

e = 2.71

$\pi = 3.14$

μ_X = media de la población

σ_X = desviación estándar de la población

x = un valor de la variable continua

Como e y π son constantes, la forma de la curva normal depende solamente de los dos parámetros de la distribución normal, la media μ_X y la desviación estándar σ_X.

Las curvas normales varían dependiendo de estos dos parámetros

Distribuciones A y B en la primera gráfica; C y D en la segunda gráfica; E y F en la tercera.

A y B tienen medias diferentes lo que las coloca en diferentes posiciones con respecto al eje de x.

C y D tienen la misma media (que aparece en el mismo punto del eje de x) pero como las desviaciones estándar son diferentes, la forma cambia. La que tiene la desviación estándar menor es más estrecha y alargada.

E y F tienen medias diferentes y desviaciones estándar diferentes por lo que tienen formas diferentes y se colocan en lugares diferentes del eje de x.

Puesto que hay un número infinito de combinaciones para los dos parámetros,

hay un número infinito de curvas normales diferentes. Para facilitar el trabajo con la distribución normal, se llevan a cabo todos los cálculos con una distribución normal llamada la distribución normal estandarizada (standard normal curve) que tiene las siguientes propiedades:

B. La curva normal estandarizada

Propiedades:

i. La media de la distribución normal estandarizada es siempre cero

$\mu_Z = 0$

ii. La desviación estándar de la distribución normal estandarizada es siempre igual a uno.

$\sigma_Z = 1$

iii. El área bajo la curva que aparece en las tablas corresponde al porcentaje o la proporción de puntuaciones en el intervalo dado

Se observa que a una desviación estándar de la media en ambas direcciones, debajo de la curva se encuentra el 68% de las puntuaciones. El 95% se halla a dos desviaciones estándar y el 99.7% a tres desviaciones estándar. Esta distribución simétrica de puntuaciones va a servir de base para resolver los ejercicios sobre la normal que se presentan a continuación.

1. Ejercicios para obtener el área bajo la curva partiendo de valores z dados

Es imprescindible trazar la gráfica de la normal y sombrear el área que se desea obtener antes de utilizar la tabla.

En todos los ejemplos siguientes z es una variable con una distribución normal estandarizada.

a) Halla el porcentaje de puntuaciones donde z es menos de 0.85

Para resolver este problema, debes determinar en la tabla de la p.633 cuánto mide el área bajo la curva entre z = 0 y z = 0.85

En la columna bajo (Area between y z) encontrarás que el área es 0.3023. Por

lo tanto debes añadir 0.5 a esa área para obtener el resultado, como ocurre en la gráfica que aparece a continuación.

Práctica: Halla el porcentaje de puntuaciones donde z es menos de 1.15
Respuesta: 0.8749

b) Halla el porcentaje de puntuaciones donde z es mayor de 0.85

En este caso puedes utilizar la respuesta anterior y obtener el área bajo la curva que buscas restando el área que obtuviste en el ejercicio anterior del total (1.00) que se encuentra bajo la curva.

Por lo tanto el área correspondiente a las puntuaciones z mayores de 0.85 es

1 - 0.8023 = 0.1977

Práctica: Halla el porcentaje de puntuaciones donde z es mayor de 1.15.
Respuesta: 0.1251

También es posible hallar el porcentaje de puntuaciones mayores de una puntuación dada haciendo otras combinaciones de la información obtenida en la tabla.

c) Halla el porcentaje de puntuaciones donde z es mayor de 0.77.

En este caso es conveniente trazar la gráfica y observar que el resultado se puede obtener directamente de la tabla de Hinkle (p. 633) si se toma como respuesta el área más allá de z (Area beyond z)

Por lo tanto el área correspondiente a las puntuaciones z mayores de 0.77 es 0.2206

Como la gráfica de la distribución es simétrica es posible obtener los resultados de la misma tabla de Hinkle (p.633) simplemente reflejando la situación en la parte positiva e interpretando la puntuación z como si fuese positiva.

d) Halla el porcentaje de puntuaciones donde z es menor de -0.25.

Para resolver este problema, traza la gráfica correspondiente.

Cuando reflejas el área en la parte positiva de la gráfica es posible determinar en la tabla de la p.633 cuánto mide el área bajo la curva cuando z es mayor de 0.25.

En la columna bajo (Area beyond z) encontrarás que el área es 0.4013. Por lo tanto debes interpretar este resultado partiendo de la gráfica que has hecho para obtener el resultado, como ocurre en la gráfica que aparece anterior. Por lo tanto el área correspondiente a las puntuaciones z menores de -0.25 es 0.4013

Práctica: Halla el porcentaje de puntuaciones donde z es menor de –2.20. Respuesta: 0.0139

e) Halla el porcentaje de puntuaciones z menores de z = -1.5 ó mayores de z = 2.0

Para resolver este problema, traza la gráfica correspondiente.

En la tabla de Hinkle p.634 observarás que el área correspondiente a z mayor de 2.0 (Area beyond z) es 0.0228 y el área correspondiente a z mayor de 1.5 (Area beyond z) es 0.0668. El resultado se obtiene sumando ambas áreas.

0.0228 + 0.0668 = 0.0896

f) Halla el porcentaje de puntuaciones z entre z = – 0.5 y z = 1.5.

Para resolver este problema, traza la gráfica correspondiente.

En la tabla de Hinkle p.634 observarás que el área correspondiente a z menor de 1.5 (Area between) es 0.4332 y el área correspondiente a z menor de z = 0.5 (Area between) es 0.1915. El resultado se obtiene sumando ambas áreas.

0.1915 + 0.4332 = 0.6247

Práctica: Halla el porcentaje de puntuaciones z menores de z = – 0.5 ó mayores de z = 1.5. (En matemáticas cuando se usa la conjunción "ó" se suman las áreas, pues puede ser tanto una, como la otra, como ambas)

g) Halla el porcentaje de puntuaciones z entre z = 1 y z = 2

Para resolver este problema, traza la gráfica correspondiente

En la tabla de Hinkle p.633-34 observarás que el área correspondiente a z menor de 2 (Area between) es 0.4772 y el área correspondiente a z menor de z = 1 (Area between) es 0.3413. El resultado se obtiene restando del área mayor, el área menor.

0.4772 - 0.3413= 0.1359

Práctica: Halla el porcentaje de puntuaciones z menores de z = 1 ó mayores de z = 2. (En matemáticas cuando se usa la conjunción "ó" se suman las áreas, pues puede ser tanto una, como la otra, como ambas)

2. Ejercicios para obtener los valores z partiendo del área bajo la curva

En todos los ejemplos siguiente z es una variable con una distribución normal estandarizada.

a) Halla el valor de z para el cual el cual el porcentaje de puntuaciones menores de z es 80.23%

Para resolver este problema, traza la gráfica correspondiente

Ahora debes determinar en la tabla de la p.633 qué valor de z corresponde a un área (Area between) de 0.3023. En este caso el valor de z es 0.85.

Cuando el valor del área no es igual al que aparece en la tabla es necesario aproximarlo.

Práctica: Halla el valor de z para el cual el cual el porcentaje de puntuaciones menores de z es 60%. Respuesta: z = 0.25

b) El 35% de los posibles valores de z son menores de _____

Para resolver este problema, traza la gráfica correspondiente

Ahora debes determinar en la tabla de la p.633 qué valor de z corresponde a un área (Area beyond) de 0.3500 aproximadamente. En este caso el valor de z es 0.39. Pero observa que el valor de z es negativo ya que se encuentra, en el

eje de z antes del cero que corresponde a la media. Por lo tanto la respuesta ha de ser **–0.39**

Práctica: Halla el valor de z para el cual el cual el porcentaje de puntuaciones **mayores** de z es 65%. Respuesta: z = -0.39

c) ¿Para qué z el 80% de los posibles valores de z son menores?

Para resolver este problema, traza la gráfica correspondiente

Ahora debes determinar en la tabla de la p.633 qué valor de z corresponde a un área (Area between) de 0.30 aproximadamente. En este caso el valor de z es 0.84. En este caso el valor de z es positivo ya que se encuentra en el eje de z después del cero que corresponde a la media. Por lo tanto la respuesta ha de ser **0.84**

Práctica: Halla el valor de z para el cual el cual el porcentaje de puntuaciones **mayores** de z es 80%. Respuesta: z = -0.84

Práctica: Halla el valor de z para el cual el cual el porcentaje de puntuaciones **mayores** de z es 20%. Respuesta: z = 0.84

d) Entre qué dos valores de z se encuentra el 96% de los posibles valores de z (simétricamente distribuidos alrededor de la media).

Para resolver este problema, traza la gráfica correspondiente

Ahora debes determinar en la tabla de la p.633 qué valor de z corresponde a un área (Area between) de 0.48 (la mitad de 0.96) aproximadamente. En este caso el valor de z es 2.05. En este caso hay dos valores de z , uno positivo y el otro negativo, pero ambos corresponden al número obtenido de la tabla. Por lo tanto la respuesta es **–2.05 y +2.05**

C. Porcentilas y rangos porcentiles.

La pregunta de cuál es el rango porcentil de una puntuación dada se puede transformar en preguntar cuál es el porcentaje de puntuaciones bajo una puntuación dada o cuál es el área correspondiente a la parte de la curva que se encuentra a la izquierda de una puntuación z dada.

Ejemplos

a. Halla el rango porcentil de z = 1.15

ie. Halla el área bajo la curva donde z es menos de 1.15 Respuesta: 0.8749. Por lo tanto se puede decir que el rango porcentil de z = 1.15 es 87.49

b. Halla el rango porcentil de z = -2.20

ie. Halla el área bajo la curva donde z es menos de -2.20 = 0.0139. Por lo tanto se puede decir que el rango porcentil de z = -2.20 es 1.39

La pregunta de cuál es la porcentila se puede transformar en preguntar cuál es la puntuación z que corresponde al área bajo la curva a la izquierda de una puntuación dada.

Ejemplos

a. Halla la porcentila 10 (P_{10}) se puede transformar en la pregunta ¿10% de los posibles valores de z son menores de cuánto?

Respuesta: -1.28

b. Halla la porcentila 85 (P_{85}) se puede transformar en la pregunta: ¿85% de los posibles valores de z son menores de cuánto?

Respuesta: 1.04

IV. Uso de las puntuaciones estandarizadas o puntuaciones z en las distribuciones normales

A. Fórmula

En muchas ocasiones se quieren comparar puntuaciones que pertenecen a dos distribuciones normales con curvas diferentes, pero de primera instancia no se puede debido a que sus dos parámetros, la media y la desviación estándar son diferentes para cada una de las dos distribuciones.

Sin embargo la comparación se hace posible cuando se convierten las puntuaciones crudas de las distribuciones en puntuaciones z o puntuaciones estandarizadas. Estas puntuaciones estandarizadas se logran restando la media a cada observación y dividiendo entre la desviación estándar.

donde μ_X es la media de la distribución y σ_X su desviación estándar.

Por lo tanto al calcular las puntuaciones z se transforma la distribución original en una con $\mu = 0$ y $\sigma = 1$, que no es otra que la distribución normal estandarizada. Esta transformación a puntuaciones z se utiliza muy frecuentemente para trabajar con distribuciones normales

B. Ejemplos

1. Ejercicios para obtener el área bajo la curva partiendo de la puntuación cruda

Todos los ejemplos que siguen a continuación corresponden a una distribución de puntuaciones distribuida normalmente con $\mu_X = 50$ y $\sigma_X = 7$

Ejemplo 1

¿Qué porcentaje de puntuaciones se encuentra entre 50 y 57?

a. Primero hay que asegurarse que la variable está distribuida normalmente (Lo dice el problema) y trazar la gráfica que va a servir de modelo para el problema.

b. Después hay que convertir la escala dada a la escala de la distribución normal estándar usando la fórmula

$z_1 = (50-50)/7 = 0$

$z_2 = (57-50)/7 = 1$

c. Por medio de la tabla se observa que el área entre 0 y 1 es 0.3413

d. Como porcentaje y área bajo la curva significan lo mismo, es posible decir que el porcentaje o la proporción de puntuaciones entre 50 y 57 es 0.3413 o 34.13%

Ejemplo 2

¿Qué porcentaje de puntuaciones hay entre 43 y 57?

$z_1 = (43-50)/7 = -1$

$z_2 = (57-50)/7 = 1$

En la tabla se observa que el área que hay entre 0 y 1 es 0.3413. El área entre -1 y 0 es también 0.3413, puesto que el área es siempre positiva. Por lo tanto la probabilidad es 0.3413 + 0.3413 = 0.6826

Nota: Observa que el 68% del área de la curva está a una desviación estándar de distancia de la media.

Ejemplo 3

Halla el rango porcentil de 45 $P(x \leq 45)$

$z_1 = (45-50)/7 = -0.71$

En la tabla el área correspondiente es 0.2389 (beyond z)

Ejemplo 4

Halla el porcentaje de puntuaciones menores de 50 ó mayores de 57. $P(x \leq 50$ ó $x \geq 57)$

a. Halla $P(x \leq 50) = 0.5$ (área bajo la curva)

b. Halla $P(x \geq 57) = 0.1587$ (beyond z)

c. $P(x \leq 50$ ó $x \geq 57) = P(x \leq 50) + P(x \geq 57)$

Por lo tanto:

$P(x \leq 50$ ó $x \geq 57) = 0.5 + 0.1587 = 0.6587$

Ejemplo 5

Halla $P(40 \leq x \leq 47)$

$z_1 = (40-50)/7 = -1.43$

$z_2 = (47-50)/7 = -0.43$

Halla $P(40 \leq x \leq 50) = P(-1.43 \leq z_1 \leq 0) = 0.0764$

Halla $P(47 \leq x \leq 50) = P(-0.43 \leq z_2 \leq 0) = 0.3336$

Por lo tanto:

$P(40 \leq x \leq 47) = 0.3336 - 0.0764 = 0.2572$

donde μ_X es la media de la distribución y σ_X su desviación estándar.

2. Ejercicios para obtener la puntuación cruda partiendo del área bajo la curva

Ejemplo 1

¿Qué puntuación tiene el 50% de las puntuaciones por debajo de ella? (ie. Halla P_{50})

La media, 50

Ejemplo 2

¿Qué puntuación tiene el 10% de las puntuaciones por debajo de ella?

La proporción es 0.1 El número que más se aproxima en la tabla en términos de área es 0.1003 (beyond z)

$z_1 = -1.28$

Por lo tanto $-1.28 = (x_1 - 50)/7$

$x_1 = 41.04$

ie: $P_{10} = 41.04$

Lectura:

Hinkle capt. 4 pp. 89-104 (ed. 3 pp.85-101)

Asignaciones:

Hinkle pp.103-104 ej. 1-8 excepto #5

MENU 6390

EDUC 6390: Estadística aplicada en la educación

Andrés Menéndez Raymat, Ph.D.

Conferencia 9

Correlación

Bosquejo

VI. Coeficiente Punto biserial (r_{pb})

VII. Coeficiente Phi (Φ)

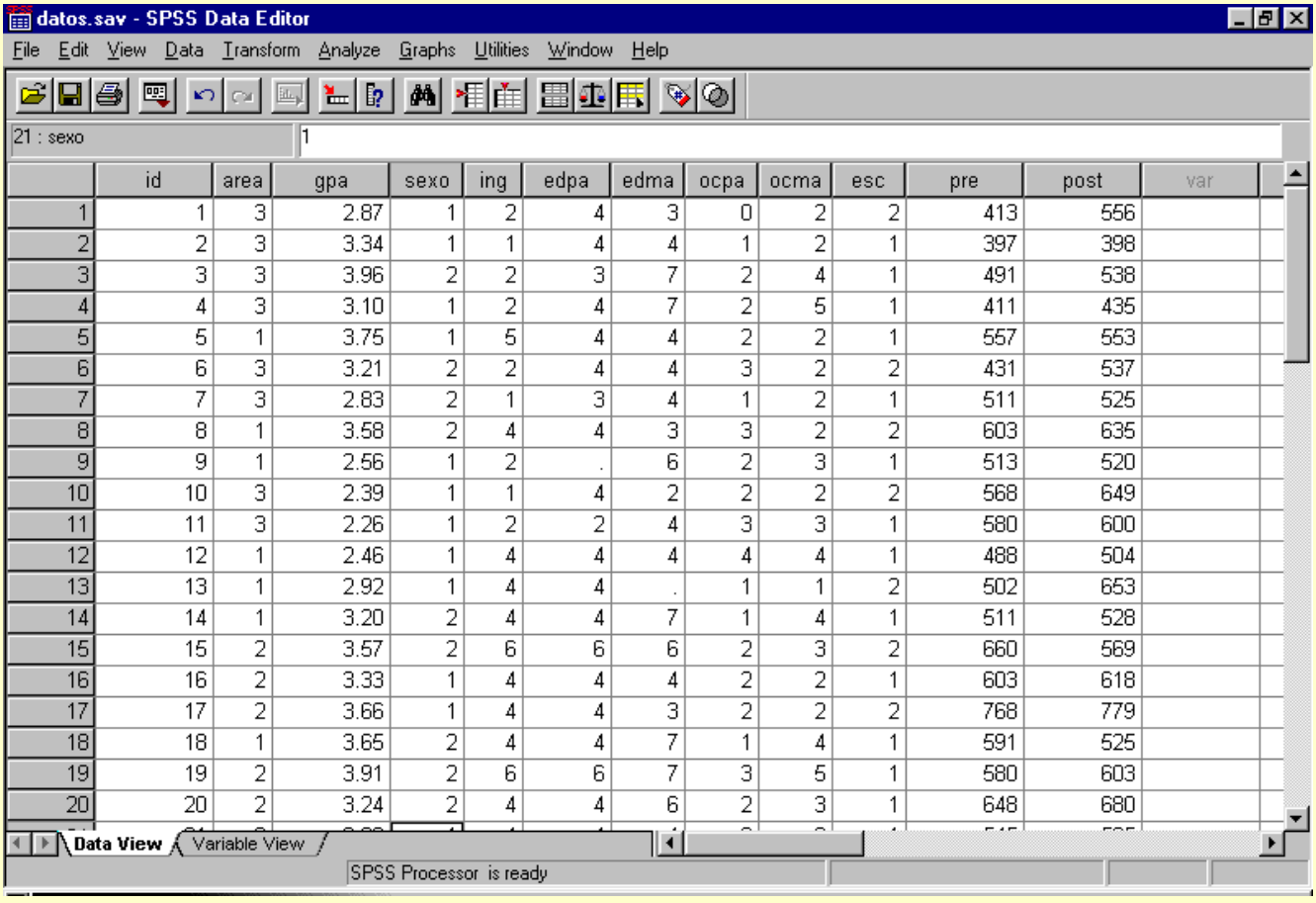
I. Introducción

A. Organización de datos para dos variables

Al trabajar con la organización de datos encontramos que cuando los datos que se obtienen corresponden a dos o más características de los mismos sujetos se puede crear una tabla donde se presentan los valores de las dos variables.

Ejemplo:

Las notas (valores numéricos) de varios exámenes para cada individuo aparecen en columnas diferentes. El problema con esta tabla es que se hace sumamente difícil poder apreciar cual es la relación entre las variables.



	id	area	gpa	sexo	ing	edpa	edma	ocpa	ocma	esc	pre	post	var
1	1	3	2.87	1	2	4	3	0	2	2	413	556	
2	2	3	3.34	1	1	4	4	1	2	1	397	398	
3	3	3	3.96	2	2	3	7	2	4	1	491	538	
4	4	3	3.10	1	2	4	7	2	5	1	411	435	
5	5	1	3.75	1	5	4	4	2	2	1	557	553	
6	6	3	3.21	2	2	4	4	3	2	2	431	537	
7	7	3	2.83	2	1	3	4	1	2	1	511	525	
8	8	1	3.58	2	4	4	3	3	2	2	603	635	
9	9	1	2.56	1	2	.	6	2	3	1	513	520	
10	10	3	2.39	1	1	4	2	2	2	2	568	649	
11	11	3	2.26	1	2	2	4	3	3	1	580	600	
12	12	1	2.46	1	4	4	4	4	4	1	488	504	
13	13	1	2.92	1	4	4	.	1	1	2	502	653	
14	14	1	3.20	2	4	4	7	1	4	1	511	528	
15	15	2	3.57	2	6	6	6	2	3	2	660	569	
16	16	2	3.33	1	4	4	4	2	2	1	603	618	
17	17	2	3.66	1	4	4	3	2	2	2	768	779	
18	18	1	3.65	2	4	4	7	1	4	1	591	525	
19	19	2	3.91	2	6	6	7	3	5	1	580	603	
20	20	2	3.24	2	4	4	6	2	3	1	648	680	

B. Gráficas de datos para dos variables

Las gráficas de datos para dos variables utilizan el plano cartesiano.

Una variable puede ser categórica y la otra numérica. En este caso la categórica por lo general ocupa el eje de x.

Si ambas variables son numéricas se puede crear un diagrama de dispersión

C. Correlación

Hasta ahora se ha descrito cada variable independientemente utilizando tablas, gráficas y medidas de tendencia central y dispersión.

Ahora veremos que es posible describir, no solamente las variables por separado, sino la relación que existe entre ellas.

Ejemplo: Hinkle p.106 (103) presenta un grupo de estudiantes que obtuvieron una nota en el College Board y otra en un examen final. (Tabla 5.1, fig. 5.1)

Se quiere saber si los que sacaron notas altas en el CB también sacaron notas altas en el examen, etc. Para eso, una de las primeras cosas que se hace es que se grafican los puntos en el plano cartesiano donde cada punto corresponde a un estudiante

Los estudios de correlación tratan de medir el grado de asociación que existe entre dos variables. Estos estudios sobre la relación entre variables son muy comunes en las ciencias sociales.

Sin embargo, como hay diferentes escalas para medir las variables veremos que la medida o coeficiente de correlación que se utilice va a depender directamente de las escalas de medición de las variables.

D. Correlación y causalidad

La correlación no implica causalidad

Ejemplo:

Existe una correlación alta entre la talla del zapato y las destrezas de lectura pero es obvio que la talla del zapato no es la causa de las destrezas de lectura. Existe una variable oculta (el crecimiento de los niños) que resulta ser una de las causas.

A menudo una tercera variable o una combinación de variables que no vemos puede ser la causa de la correlación. Por lo tanto siempre es importante asegurarse de que al hablar sólo se menciona asociación y relación, jamás causa y efecto o dependencia.

II. Escalas de medición

Hay cuatro formas o escalas básicas para medir datos (nominal, ordinal, intervalar y de razón). Si las variables son categóricas entonces dependiendo del grado de precisión posible en la medición se utilizan las siguientes dos escalas:

1. Escala nominal

Se utiliza cuando los datos están clasificados en categorías en las que no hay ninguna idea de ordenamiento. No se puede decir que una categoría es mejor que otra.

Ejemplo:

colores, religiones, partidos políticos, etc.

Cuando se trabaja con correlaciones hay un tipo de escala nominal sumamente importante. Consta de sólo dos niveles. Las variables clasificadas en estas dos categorías se llaman dicótomas.

Ejemplo:

Sexo, fuma o no fuma, rico o pobre, etc.

Generalmente cuando se codifica, para identificar la presencia del atributo se usa 1 y su ausencia 0.

2. Escala ordinal

Hay orden en este nivel de medición. Se obtiene mayor información sobre la variable. Implica que una categoría es mejor que otra.

Ejemplo:

Escala Likert: Acuerdo total, acuerdo parcial, desacuerdo, etc.

En estos casos no se puede medir la diferencia entre uno y otro, aunque es obvio que uno es mayor o mejor que otro.

En muchas ocasiones se usan números para codificar estas respuestas como acuerdo total (5), acuerdo parcial (4). Estos números sólo representan orden. En ningún momento se implica que la diferencia entre acuerdo total y acuerdo parcial es de una unidad.

También se usa mucho la clasificación en dos categorías aunque la variable en sí sea cuantitativa, continua y tenga una distribución normal.

Esto se debe, en la mayoría de los casos, a que no se recogieron los datos íntegros Ejemplo: Niños con IQ sobre 100 y bajo 100. Hay puntuación para cada niño, pero sólo se recogió si estaba sobre el promedio o no.

Cuando las variables son cuantitativas entonces es posible usar una de las siguientes dos escalas:

3. Escala intervalar

En esta escala la diferencia entre dos medidas es significativa.

Ejemplo:

79 grados es 2 más que 77 grados de temperatura. La diferencia entre 79 y 77 grados es la misma que entre 55 y 53 grados.

Sin embargo no hay un cero verdadero. El cero en temperatura Fahrenheit es una temperatura seleccionada al azar. El cero en centígrados corresponde a otra temperatura muy diferente.

El resultado es que, a pesar que 100 es el doble de 50, en una temperatura de 100^0 no hace el doble de calor que en una de 50^0.

4. Escala de razón

Tiene un cero real.

Ejemplo:

peso, altura.

Tiene sentido hablar de que una persona pesa el doble de otra.

Nota:

A veces los investigadores convierten una variable cuantitativa a rangos o dicotomías, pero esto no lo hacen sin tener razones muy poderosas para ello, pues en realidad estarían perdiendo información muy valiosa. Generalmente cuando se hace es porque se trabaja con datos ya recogidos en términos de rangos o dicotomías.

Lectura

Hinkle Capt. 20 pp. 548-551

III. Coeficientes de correlación para variables cuantitativas v cualitativas.

En la siguiente tabla aparecen las combinaciones posibles de dos variables y los coeficientes de correlación que se pueden utilizar en cada caso.

			Variable X	
		Nominal	Ordinal	Interv/Raz

	Nominal	a. Phi (Φ) b. Coeficiente C c. Coeficiente V d. λ y λ_Y	Rango-biserial	Punto-biserial
Var Y	Ordinal	Rango-biserial	a.Tetrachoric b. Spearman ρ	Biserial
	Interv/ Razón	Punto-biserial	Biserial	Pearson r

IV. Pearson r

A. Diagramas de dispersión

(scattergram)

Este tipo de diagrama presenta una imagen de la relación entre dos variables numéricas.

En la gráfica de la transparencia se observa un patrón que indica una correlación positiva, puesto que los puntos suben a medida que nos movemos hacia la derecha.

La correlación negativa ocurre cuando los puntos bajan a medida que nos movemos a la derecha. (Ver pendiente en la recta)

La correlación sería perfecta (1 ó -1) si los puntos todos formaran una recta. Cuando no hay una tendencia hacia arriba o hacia abajo, la correlación está cerca de cero.

El coeficiente de correlación puede tomar valores entre +1 y -1, donde el signo indica dirección de la relación.

Cuando se observan valores como +0.9 ó -0.9 se dice que la relación es fuerte. Valores como ± 0.1 indican una correlación débil.

La magnitud de la correlación (si es fuerte o débil) se mide utilizando el valor absoluto. La dirección se determina con el signo.

B. Cómputo de Pearson r

La idea principal es multiplicar el valor de la variable x por el de la variable y para cada individuo y hallar el promedio.

Pero hay varias fórmulas para computar el coeficiente de Pearson.

1. Puntuaciones estandarizadas

Se utiliza la idea de multiplicar las puntuaciones de las dos variables para cada individuo y hallar el promedio. Sólo que se utilizan los valores estandarizados puesto que es la forma de que los valores para las dos variables sean comparables.

Si se hace a mano es sumamente tedioso y largo pues hay que convertir cada valor de la variable al valor estandarizado (Hinkle p.110)

$$r_{xy} = \frac{\sum z_x z_y}{n-1}$$

2. Fórmula con la desviación de la media

En esta fórmula por medio de unas manipulaciones matemáticas se transforman los valores de z en **x y**, donde estas variables representan la desviación de cada puntuación con respecto al promedio (Hinkle, p.112)

x = X - X

y = Y - Y

$$r_{xy} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

3. Fórmula con las puntuaciones crudas

Esta fórmula requiere menos cómputos, pues usa las puntuaciones crudas sin necesidad de convertirlas en desviaciones o puntuaciones estandarizadas. (Hinkle, p.113)

$$r_{xy} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

4. Fórmula con la covarianza

Esta es la fórmula que urtiliza el programa Excel.

La covarianza es otra forma de expresar la relación entre dos variables.

No se utiliza a menudo, pues no está entre los valores de +1 y -1 como la correlación. Sin embargo es útil para esta fórmula. (Hinkle, p.114)

$$S_{xy} = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{n-1} = \frac{\sum xy}{n-1}$$

Por lo tanto

$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

Esto es la covarianza dividida por las desviaciones estándar de X y de Y.

C. Condiciones para poder utilizar el coeficiente de Pearson r

1. Las variables que se correlacionan tienen que ser pareadas para el mismo sujeto. No se puede tomar la variable X de un sujeto y la variable Y de otro.

2. Las variables tienen que ser medidas utilizando la escala intervalar o de razón.

3. La distribución de ambas variables tiene que ser normal.

D. Factores que afectan al coeficiente de Pearson r

1. Linealidad

Si los puntos tienden a caer cerca de una línea recta, se puede decir que hay una relación lineal.

Pero puede haber otro tipo de relación no lineal, como la curvilineal donde los valores de X y Y aumentan al principio, pero luego cuando X aumenta Y disminuye.

(Transp. Fig. 5.3, Hinkle p.115)

Ejemplos:

Relación entre ansiedad y ejecutoria. Poca ansiedad o mucha ansiedad produce ejecutoria pobre.

Relación entre la edad y la dependencia de otros es curvilineal también pues tanto los jóvenes como los ancianos dependen mucho..

El problema es que si se computa una relación curvilineal como lineal arroja que no hay relación pues las oposiciones se cancelan.

2. Homogeneidad del grupo

Si el grupo es muy homogéneo, quiere decir que la dispersión es poca.

Mientras menos dispersión hay, más pequeño es el coeficiente de correlación.

En la transparencia anterior si se reduce el alcance de las puntuaciones de aptitud (se eliminan las puntuaciones altas y las bajas) se observa que hay menos tendencia hacia una recta. Por lo tanto la correlación se ha reducido considerablemente.

A veces al no tomar el grupo en su totalidad la correlación parece ser muy pequeña cuando en realidad no lo es.

Ejemplo: Si se correlacionan los resultados de las pruebas de Razonamiento Matemático del CB de los estudiantes con su índice académico de primer año. La correlación parece pequeña puesto que se han eliminado todos los estudiantes que no fueron admitidos a la universidad.

3. Tamaño del grupo

Sin embargo el tamaño del grupo no afecta el valor de la correlación, lo que puede afectar es su precisión. Con pocos datos no hay seguridad de que siempre pase lo mismo.

E. Interpretación del coeficiente de Pearson r

1. En términos de la escala

La escala de r es ordinal. Por lo tanto en los casos en que es r = 0.40; r= 0.60; r= 0.80 **No** podemos decir que para r = 0.80 hay el doble de correlación que para r = 0.40.

No se puede decir que existe la misma diferencia entre r = 0.40 y r = 0.60 que entre r = 0.60 y r = 0.80

Lo más que se puede decir es que la relación lineal entre las variables es mayor o menor.

Para determinar si la correlación es alta o baja, se puede pensar en términos de la siguiente tabla, pero hay que tener en cuenta de qué se está hablando, pues la interpretación depende siempre de la situación. En términos de admisiones a la universidad es difícil hallar una relación mayor de 0.50 entre el promedio del primer año de universidad y el promedio de graduación de escuela superior, por lo tanto en ese caso una correlación de 0.50 es alta.

Tamaño de la correlación	Interpretación
0.90-1.00	Muy alta
0.70-0.90	Alta
0.50-0.70	Moderada
0.30-0.50	Baja
0.00-0.30	Muy poca

2. En términos de la varianza

Otro significado y uso de la correlación tiene que ver con el porciento de la variación en una variable se relaciona con la variación en la otra variable.

Ejemplos:

¿Cuánto de la educación se relaciona con la escuela?

¿Cuánto de las notas del primer año de universidad está asociada con el índice de graduación?

¿Cuánto de la inteligencia se relaciona con la herencia?

Definición

Coeficiente de determinación

El cuadrado del coeficiente de correlación se llama el coeficiente de determinación = r^2

El coeficiente de determinación representa el porciento de la varianza en una variable que está asociado con la otra variable.

$$r^2 = \frac{S_a^2}{S_y^2}$$

donde $(S_a)^2$= varianza en Y asociada con X

$(S_Y)^2$ = varianza total de Y

Ejemplo:

Hay 75 estudiantes y el coeficiente de correlación entre sus puntuaciones de aptitud y su índice académico del primer semestre es r = 0.69.

Por lo tanto $(0.69)^2$ = 0.48 de la varianza en el índice académico del primer semestre se relaciona con la variación de la puntuación en aptitud. El 52% restante de la variación está asociada con otros factores que no son la aptitud.

Esta es una de las razones para decir que un coeficiente de correlación es alto o bajo. Por ejemplo un coeficiente de correlación de 0.90 implica que un 81% de la varianza de la segunda variable está asociada con la varianza de la primera. Se puede decir que la correlación es alta.

Sin embargo un coeficiente de correlación de 0.30 implica sólo un 9% de la varianza de la segunda variable. Se puede decir que la correlación es baja.

Lectura

Hinkle capt. 5 pp.105-126
Hinkle capt. 20 pp.548-552

V. Spearman Rho

Se utiliza cuando las dos variables son ordinales. No se encuentra en Excel, así que hay que hacerlo a mano o con SPSS.

El rango más alto es 1.

Cuando hay empates en las puntuaciones de los sujetos en una variable, se da el promedio de los posibles rangos a cada uno de los empatados.

Ejemplo:

Si hay empates para los rangos 10,11 y 12 (10+11+12)/3 = 11 rango para los tres.

Fórmula:

$$\rho = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

donde

n = número de rangos pareados

d = diferencia entre los rangos pareados

Ejemplo:

$$\rho = 1 - \frac{6\sum 36.50}{15(225 - 1)}$$

ρ = 0.93

Lectura

Hinkle capt. 5 pp.124-125

VI. Coeficiente Punto biserial (r_{pb})

Es un caso especial de la correlación de Pearson cuando una variable se mide en la escala intervalar o de razón y la otra es nominal y dicótoma.

Se utiliza principalmente en los exámenes estandarizados para determinar si un ejercicio debe estar o no en el examen. Si se determina que los buenos estudiantes lo hacen mal y los malos lo hacen bien, el ejercicio no discrimina y debe eliminarse.

Así que en todos los exámenes estandarizados se busca la correlación punto biserial para relacionar el ejercicio (bien o mal contestado) con la puntuación de cada estudiante.

Fórmula:

$$r_{pb} = \frac{\overline{X_1} - \overline{X_0}}{\sigma_Y}\sqrt{pq}$$

Donde

x_1 = media de los estudiantes que sacaron 1 en el ítem

x_0 = media de los estudiantes que sacaron 0 en el ítem

σ_Y = desviación estándar de todas las puntuaciones en la prueba

p = proporción de individuos que sacaron 1 en el ítem

q = proporción de individuos que sacaron 0 en el ítem

Ver Hinkle Capt.20 p.552-554 (Transp. Tabla 20.2, p.553 (19.2)

Persona	Ejercicio (X)	Examen (Y)
A	1	10
B	1	12
C	1	16
D	1	10
E	1	11
F	0	7
G	0	6
H	0	11
I	0	8
J	0	5

x_1 = media de los estudiantes que sacaron 1 en el ítem

= (10+12+16+10+11)/5 = 11.80

x_0 = media de los estudiantes que sacaron 0 en el ítem

= (7+6+11+8+5)/5 = 7.4

σ_Y = desviación estándar de todas las puntuaciones en la prueba

= 3.07 (con calculadora)

p = proporción de individuos que sacaron 1 en el ítem

= 0.5

q = proporción de individuos que sacaron 0 en el ítem

= 0.5

$$r_{pb} = \frac{11.8 - 7.4}{3.07}\sqrt{(0.5)(0.5)}$$

r_{pb} = 0.716

VII. Coeficiente Phi (Φ)

Es un caso especial de la correlación de Pearson cuando ambas variables son nominales y dicótomas.

La fórmula se puede reducir a:

$$\phi = \frac{BC - AD}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

cuando se organiza una tabla de contingencia de la siguiente manera:

		Variable X		
		0	1	Total
Variable Y	1	A	B	A+B
	0	C	D	C+D
	Total	A+C	B+D	N

Ejemplo: Hinkle p.555 (514) (Tablas 20.3, 20.5)

Determinar la relación entre género y partido político

Persona	Género	Partido
A	1	1
B	1	1
C	1	0
D	1	1
E	1	1
F	0	0
G	0	1

H	0	1
I	0	0
J	0	0

1 = MUJER

1 = REPUBLICANO

0 = HOMBRE

0 = DEMOCRATA

Tabla de contingencia

		GENERO		
		0 (masc)	**1(fem)**	Total
PARTIDO	**1(rep)**	A (2)	B (4)	A+B (6)
	0 (dem)	C (3)	D (1)	C+D (4)
	Total	A+C (5)	B+D (5)	N (10)

$$\phi = \frac{(4)(3) - (2)(1)}{\sqrt{(2+4)(3+1)(2+3)(4+1)}}$$

$\phi = 0.408$

Conclusión: Hay baja relación positiva entre partido político y sexo. Las mujeres se asocian con el partido republicano y los hombres con el demócrata. Esto sucede porque 0 = demócrata y 0 = hombre ; 1 = republicano y 1 = mujer.

Si la correlación hubiese sido negativa, entonces podíamos decir que las mujeres se asociaban con los demócratas y los hombres con los republicanos.

Lectura

Hinkle capt. 20 pp.552-556

MENU 6390

EDUC 6390: Estadística aplicada en la educación
Andrés Menéndez Raymat, Ph.D.
Conferencia 10
Métodos descriptivos en la regresión

Bosquejo

I. Introducción

Anteriormente se había podido determinar el nivel de correlación entre dos variables mediante el cálculo del coeficiente de correlación. En el caso de dos variables numéricas medidas en la escala intervalar o de razón se utilizó el coeficiente de Pearson. Se había visto en el diagrama de dispersión como a veces, cuando la correlación era alta, se podía visualizar una recta que pasaba entre los puntos. Si la correlación era +1 ó -1 todos los puntos coincidían en esta recta.

A partir de esta conferencia a esa recta se le llamará la línea de regresión de y en x. Va a servir para predecir los valores de una variable en términos de los valores de la otra variable. Finalmente se podrá determinar cuánto error existe cuando se logra predecir una variable en términos de la otra.

II. Ecuaciones lineales con una variable independiente

Para poder entender qué es una línea de regresión es necesario volver un poco al álgebra y al plano cartesiano para repasar qué es una ecuación lineal.

$y = b_0 + b_1x$ es una ecuación lineal

b_0 y b_1 son constantes, x es la variable independiente y es la variable dependiente, pues sus valores van a depender de los de x.

La gráfica de esta ecuación es una línea que se obtiene después de determinar dos puntos, graficarlos y trazar la recta que pasa por ellos.

Esto se puede hacer, pues en geometría se aprende que dos puntos determinan una recta.

Ejemplo:

Traza en el plano cartesiano las siguientes gráficas lineales

y = -3 + 4x

y = 2 - 3x

y = 5x

Ejemplo:

Una maestra cobra $15 por hora por sus servicios, pero exige que se le paguen $50 por firmar el contrato. Crea el modelo matemático que represente mejor la cantidad de dinero que debe cobrar al finalizar su trabajo.

y = 50 + 15x

donde x es el número de horas que trabaja.

La gráfica se obtiene cuando se determinan dos puntos incluidos en la recta, se grafican y se traza una recta que los incluya.

La **pendiente** representa la inclinación de la recta y su signo corresponde al de la correlación. Si la pendiente es positiva, la correlación es positiva y los valores de y aumentan a medida que aumentan los valores de x (la recta sube hacia la derecha). Si la pendiente es negativa, la correlación es negativa los valores de y disminuyen a medida que aumentan los valores de x (la recta baja hacia la derecha). Si la pendiente es cero, la correlación es cero y la recta es horizontal. Si la recta es vertical entonces la pendiente no existe.

La pendiente se representa con el coeficiente de x.

Físicamente representa el cambio en **y** sobre el cambio en **x** cuando se comparan dos puntos en la recta.

Ejemplo:

y = 5x - 2

Dos posible puntos en la recta son (0,-2) y (1,3). Por lo tanto la pendiente es (-2-3) /(0-l) = 5

El **intercepto de y** representa el valor de y cuando x es cero. Gráficamente es el punto en el eje de y en que la recta corta ese eje.

Ejemplo:

En la recta y = 2 – 3x 3l intercepto de y es (0,-2), puesto que la recta corta al eje de y en ese punto.

III. La línea de regresión

A. El criterio de los cuadrados menores

Las situaciones que se dan en la vida real no son siempre tan simples como la de la maestra y sus horas de trabajo.

Muchas veces se encuentran valores para dos variables presentados en una tabla con la que se puede construir un diagrama de dispersión.

Ejemplo:

En el siguiente caso un profesor quiere predecir el índice académico que obtendría un nuevo estudiante durante el primer semestre en su clase de 14 estudiantes.

estudiante	Indice académico	Razonamiento Verbal
1	2.3	200
2	2.4	210
3	3.6	210
4	3	245
5	3	365
6	2.3	450
7	3	470
8	3.8	500
9	3.2	555
10	3.1	575
11	3	575
12	3.4	680
13	3.6	790
14	3.9	800
Media	3.11	

Si sólo contara con los índices académicos de los 14 estudiantes que ya están en la clase, lo único que pudiera hacer es utilizar la media de los índices de esos estudiantes para predecir cuál sería el índice del nuevo estudiante. Por lo tanto el maestro sólo podría decir que espera que el nuevo estudiante hubiera obtenido 3.11 de índice. Pero en esta predicción el error puede ser mayúsculo puesto que no se tiene en cuenta ninguna de las características del nuevo estudiante.

En una gráfica, esta situación se puede representar de la siguiente forma.

El error de predicción se puede visualizar como la suma de las distancias desde cada uno de los puntos hasta la recta

$y = 3.11$

Sin embargo, esta predicción se puede mejorar si pensamos que hay otra recta que pasa entre los puntos de forma que la suma de las distancias entre los puntos y la recta se minimice. Visualmente se puede pensar que la recta representada en la gráfica anterior puede rotar hasta situarse de la siguiente manera

Esto se logra cuando se añade un predictor a la ecuación (en este caso, la puntuación que obtuvieron los estudiantes en una prueba de razonamiento verbal). Gracias a este predictor se obtiene la línea de regresión que minimiza la suma de las distancias de cada punto hasta la recta.

La nueva predicción del índice del nuevo estudiante se obtendría utilizando la ecuación de la línea de regresión que en este caso es

$y = 0.0015x + 2.41$

y sustituyendo la puntuación que el nuevo estudiante obtuvo en la prueba de razonamiento verbal en el lugar de la variable x.

Si el estudiante hubiese obtenido 300 en la prueba de razonamiento verbal es posible predecir que obtendría un índice de

$y = 0.0015(300) + 2.41 = 2.86$

Esta nueva predicción utilizando la línea de regresión no ofrece certitud, pero sí reduce considerablemente el error que se hubiese cometido si no se tiene en cuenta ninguna de las características del estudiante.

La característica principal de la línea de regresión es que minimiza la distancia. Matemáticamente hablando esta característica se expresa diciendo que **la línea de regresión cumple con el criterio de los cuadrados menores**.

En la siguiente situación, se consideran solamente cuatro puntos una recta cualquiera que pasa entre ellos y la línea de regresión. El objetivo es demostrar que la suma de las distancias de los puntos a la recta es menor en el caso de la línea de regresión.

Ejemplo:

estudiante	Razonamiento verbal	índice
1	200	2.3
2	365	3
3	360	4
4	790	3.6

En este ejemplo hay solamente 4 puntos. El diagrama de dispersión aparece a continuación. Como hay varias rectas que se pueden ajustar para que pasen entre los puntos se escogerán dos. Una de ellas será la línea de regresión y la otra será una recta cualquiera que pase entre los puntos.

La primera recta (A) es la línea de regresión y su ecuación es y = 0.0016x + 2.54

La segunda es la recta (B) representada por la ecuación y = 0.002x + 3

En cada una de ellas se calculará el error e que se va a visualizar como la distancia que hay entre el punto trazado en el diagrama de dispersión y el que correspondería en la recta a ese valor de x. Este valor que la recta predice para y, se denota como y*. Por lo tanto e = y - y*

Cuando x = 200 en la recta (A) y = 0.0016(200) + 2.54 = 2.86

pero el punto real de la tabla es (200,2.3).

Por lo tanto,

e = y - y* = 2.3 – 2.86 = -0.56

Cuando x = 365 en la recta (A) y = 0.0016(365) + 2.54 = 3.124

pero el punto real de la tabla es (365, 3).

Por lo tanto,

e = y - y* = 3 – 3.124 = -0.124

Si hallamos todos los errores e para ambas líneas de regresión, vamos a encontrar que cuando los cuadramos y sumamos, la suma de todos los e^2 es menor para la línea de regresión que para la recta B.

Línea de regresión

Estudiante	x	y	y pred	e	e^2
1	200	2.3	2.86	-0.56	0.31360
2	365	3	3.124	-0.124	0.01537
3	360	4	3.116	0.884	0.78150
4	790	3.6	3.804	-0.204	0.02170
				Suma	1.1328

Recta B

Estudiante	x	y	y pred	e	e^2
1	200	2.3	3.4	-1.1	1.2100
2	365	3	3.73	-0.73	0.5329

3	360	4	3.72	0.28	0.0784
4	790	3.6	4.58	-0.98	0.9604
				Suma	2.7817

Si hallamos todos los errores e para ambas líneas de regresión, vamos a encontrar que cuando los cuadramos y sumamos, la suma de todos los e^2 es menor para la línea de regresión que para cualquier otra recta que hayamos escogido.

La línea de regresión es, por lo tanto, la recta de mejor ajuste según el criterio de los cuadrados menores. Sin embargo, para poder determinar esta recta es necesario utilizar algunas fórmulas que eventualmente se aplicarán por medio del programa Excel.

B. La ecuación de la línea de regresión

Si la línea de regresión está representada por la ecuación

y = bx + a

entonces las fórmulas para obtener b y a son las siguientes

$$b = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = (r)\frac{S_Y}{S_X}$$

$$a = \overline{Y} - b\overline{X}$$

donde

r = correlación

S_Y = desviación estándar de Y

S_X = desviación estándar de X

= media de X

= media de Y

b se llama el coeficiente de regresión

a se llama la constante de regresión.

Ejemplo:

Estudiantes	Raz.Mat.	Educ 6390
1	565	54
2	543	85
3	690	78
4	710	90
5	235	34
6	345	47
7	675	79
8	578	67
9	800	95
10	520	54
11	430	56
12	235	32
13	346	45
14	445	50
15	490	60

En este ejemplo tenemos 15 estudiantes y lo que cada uno de ellos obtuvo en una prueba de razonamiento matemático y en el curso EDUC 6390 . Queremos determinar la línea de regresión que nos permitirá determinar la puntuación en EDUC 6390 de otro estudiante del cual sólo sabemos su puntuación en razonamiento matemático. Si utilizamos Excel para trazar la línea de regresión y obtener la ecuación de regresión, éstan son las siguientes.

y = 0.107x + 7.234

En este caso los puntos correspondientes a las puntuaciones observadas se acercan pero no están en una perfecta recta. La correlación (0.928) es positiva y alta. Parece que se puede ajustar una recta entre los puntos que modele la relación entre las puntuaciones de la prueba de razonamiento matemático y las puntuaciones de EDUC 6390.

También utilizando las fórmulas se puede obtener el coeficiente de regresión

a = 0.107

y la constante de regresión

b = 7.234

Por lo tanto la línea de regresión es

y = 0.107x + 7.234

A partir de este momento se pueden predecir valores de y dados valores de x (puntuaciones en EDUC 6390 partiendo de las puntuaciones en la prueba de razonamiento matemático)

Ejemplo:

¿Qué puntuación obtendría en EDUC 6390 un estudiante que hubiese obtenido 700 en razonamiento matemático ?

y = 0.107(700) + 7.234

y = 74.9 + 7.234

y = 82.134

C. Extrapolación

Es razonable decir que la línea de regresión permite hacer predicciones sobre valores de y dentro del dominio de x, pero las predicciones que se hagan fuera de este intervalo no son necesariamente correctas.

Ejemplo:

En el mismo caso anterior, ¿Cuánto obtendría en EDUC 6390 un estudiante que hubiese obtenido una puntuación de 900 en la prueba de razonamiento matemático?

Si utilizamos la misma línea de regresión nos encontramos con que

y = 0.107(900) + 7.234

y = 103.534

Pero en la clase de EDUC 6390 las puntuaciones no van más allá de 100. Este error se debe a que se ha escogido una puntuación de la prueba de razonamiento matemático superior a la mayor que aparece en el ejemplo que correspondió al estudiante # 9 quien obtuvo 800 en dicha prueba.

Por lo tanto se puede utilizar la línea de regresión básicamente dentro de los límites impuestos por los datos que se han recolectado. En el ejemplo anterior las puntuaciones de razonamiento matemático que se pueden utilizar van desde 235 hasta 800.

IV. Predicción de valores z

Con un poco de manipulación de las fórmulas es posible llegar a una nueva fórmula que relaciona valores estándar de la variable predictora (X) con valores estándar de la variable predicha (Y). En esta nueva fórmula la correlación desempeña un papel importante.

La fórmula es $z_y = r\, z_x$

V. Error estándar del estimado

A pesar de que se utilizó el método de los cuadrados menores para determinar la línea de regresión, la ecuación no es un predictor perfecto. Al igual que no todos los puntos de los datos coincidieron en la recta, los valores actuales de Y no van a coincidir con los valores predichos para Y. Hace falta desarrollar una estadística que represente esta dispersión o variabilidad de los puntos con respecto a la línea de regresión.

El error estándar del estimado (standard error of the estimate) es algo semejante a la desviación estándar de una variable con respecto a la media. Sólo que ahora es la dispersión con respecto a la línea de regresión.

Se define como

$$S_{YX} = \sqrt{\frac{\sum \left(Y_i - \overline{Y}_i\right)^2}{n-2}}$$

donde

Y_i = valor actual de Y para X_i

$\overline{Y_1}$ = valor predicho de Y para X_i

Por lo tanto para poder determinarlo es necesario hallar primero los valores predichos de Y para compararlos con los valores actuales. Se pueden usar otras fórmulas, pero lo mejor es dejar todo este proceso a la computadora.

El error estándar del estimado se mide en la misma unidad de medida de la variable dependiente Y.

Sin embargo, cuando se trata de muestras grandes la fórmula se puede intercambiar con una que utiliza la correlación

$$S_{XY} = S_Y \sqrt{1 - r^2}$$

donde

S_y = desviación estándar de Y

r = correlación entre X & Y

Sin embargo, si la muestra es pequeña, esta misma fórmula puede dar un estimado menor que el adecuado para el error estándar.

En estos casos en que la muestra es pequeña, si la correlación es alta, el error estándar que se obtiene es más pequeño que el error real. Si la correlación es baja, entonces el error estándar que se obtiene es más grande que el error real.

Actividades:

Hinkle pp.147-151 ej. 2,3,7. No utilizar ninguna fórmula, sólo Excel

Lecturas:

Hinkle capt.6 pp. 132-151

MENU 6390

EDUC 6390: Estadística aplicada en la educación
Andrés Menéndez Raymat, Ph.D.
Conferencia 11
La estadística inferencial y las distribuciones de probabilidad

Bosquejo

d. La curva normal estandarizada

I. Tipos de estadísticas

A. Descriptiva

Se origina con la recolección de datos poblacionales para censos. Se usa en Roma y se habla de ella en los evangelios. Su énfasis recae en los aspectos de presentar y describir datos.

Definición:

Consiste de los métodos utilizados para recolectar, presentar y describir datos de manera adecuada.

B. Inferencial

Se origina en el Renacimiento con el desarrollo de la probabilidad matemática, que a su vez se basa en el estudio de los juegos de azar. Tiene que ver con poblaciones, muestras, parámetros y estadísticas.

Población

La población es el total de objetos bajo consideración. Es el grupo sobre el cual se quiere hacer una inferencia. La mayor parte de las veces es muy grande. Algunas veces es hipotética. Si, por ejemplo, se quiere probar que la semejanza entre personas afecta el nivel de atracción, se hace imposible encontrar una población de personas semejantes en todos los aspectos.

Muestra

Una muestra es la porción de la población seleccionada para un experimento o investigación. Esta selección se hace porque generalmente el costo, el tiempo y los recursos son limitados para llevar a cabo el experimento con toda la población. Partiendo del estudio de la muestra, el investigador puede hacer inferencias sobre la población.

Parámetro

El parámetro es una medida de una característica numérica de la población. (Media, mediana, varianza, etc.). Es un elemento descriptivo de la población.

Estadística

Es una medida que se utiliza para describir una característica numérica de la muestra, no de la población como en el caso del parámetro. La estadística inferencial sirve para determinar como una estadística y un parámetro se relacionan.

Definiciones posibles de la estadística inferencial:

1. Consiste de los métodos y procedimientos que hacen posible la estimación de una característica de la población basándose exclusivamente en los resultados obtenidos en la muestra.

2. Es el conjunto de métodos que hacen posible la estimación de un parámetro basándose exclusivamente en la estadística correspondiente.

3. Son las generalizaciones sobre la población basadas exclusivamente en los resultados de la muestra.

Pero antes de entrar de lleno en la estadística inferencial es preciso clarificar un par de conceptos importantes sobre probabilidad.

II. Distribución de probabilidad de variables discretas

A. Definición

Una **distribución de Probabilidad** es una lista o tabla que incluye todos los posibles eventos o valores de una variable y su probabilidad.

Ejemplo 1:

Si se lleva a cabo un experimento que consiste en lanzar un dado una sola vez y los eventos son los valores obtenidos. La distribución de probabilidad del experimento debe incluir todos los posibles valores que se pueden obtener y su probabilidad

Valor	Probabilidad
1	1/6
2	1/6

3	1/6
4	1/6
5	1/6
6	1/6

De esta tabla se pueden obtener otras probabilidades mediante la suma de probabilidades

¿Cuál es la probabilidad de obtener 2 ó 3?

P(2 ó 3)= P(2) + P(3) = 1/6 + 1/6 = 2/6 = 1/3

¿Cuál es la probabilidad de obtener 3 ó menos?

P(1 ó 2 ó 3) = P(1)+P(2)+P(3) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2

¿Cuál es la probabilidad de obtener un número par?

P(2 ó 4 ó 6) = P(2)+P(4)+P(6) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2

Ejemplo 2:

Si se lleva a cabo un experimento que consiste en lanzar dos dados una sola vez y los eventos son la suma de los valores obtenidos. La distribución de probabilidad del experimento debe incluir todos los posibles valores que se pueden obtener y su probabilidad

Valores	combinaciones posibles	Prob
2	(1,1)	1/36
3	(1,2),(2,1)	2/36
4	(1,3),(3,1),(2,2)	3/36
5	(1,4),(4,1),(3,2),(2,3)	4/36
6	(1,5),(5,1),(2,4),(4,2),(3,3)	5/36
7	(1,6),(6,1),(2,5),(5,2),(3,4),(4,3)	6/36
8	(2,6),(6,2),(3,5),(5,3),(4,4)	5/36
9	(3,6),(6,3),(4,5),(5,4)	4/36
10	(4,6),(6,4),(5,5)	3/36
11	(5,6),(6,5)	2/36
12	(6,6)	1/36

De esta tabla se pueden obtener otras probabilidades mediante la suma de probabilidades

¿Cuál es la probabilidad de obtener una suma de 4 o menos?

P(2 ó 3 ó 4) = P(2)+P(3)+P(4) = 1/36 + 2/36 + 3/36 = 6/36 = 1/6

¿Cuál es la probabilidad de obtener por lo menos 11?

P(11 ó 12) = P(11)+P(12) = 2/36 + 1/36 = 3/36 = 1/12

B. Características

Las características más importantes de una distribución de probabilidad son la media y la desviación estándar.

1. Valor esperado

El valor esperado de una distribución de probabilidad no es otra cosa que la media aritmética. El símbolo que se utiliza es la letra griega μ_X con el suscrito x. Se obtiene multiplicando cada valor de la variable por su probabilidad y sumando todo eso.

Nota: Las siguientes son diez observaciones de un experimento:

2.1 2.1 2.1 2.1 3.2 3.2 4 4 4 4.3

La media es la suma dividida entre 10

$$\frac{2.1 + 2.1 + 2.1 + 2.1 + 3.2 + 3.2 + 4 + 4 + 4 + 4.3}{10}$$

Pero esto es lo mismo que:

$$\frac{(4)(2.1) + (2)(3.2) + (3)(4) + (l)(4.3)}{10}$$

(4/10)(2.1) + (2/10)(3.2) + (3/10)(4) + (1/10)(4.3) = 3.11

La distribución de probabilidad en este caso hubiera sido:

Valor	Probabilidad
2.1	4/10
3.2	2/10
4	3/10
4.3	1/10

Definición de valor esperado:

$$E(X) = \mu_X = \sum_{i=1}^{N} X_i P(X_i)$$

donde

X = variable

X_i = i^{th} valor de X

$P(X_i)$ = probabilidad de X_i

i = 1,2,3,...N

Nota:

La media no es necesariamente un valor de la distribución.

Ejemplo:

El valor esperado de lanzar el dado una vez es el promedio que se obtendría si se lanzara el dado muchas veces.

(1)(1/6) + (2)(1/6) + (3)(1/6) + (4)(1/6)+ (5)(1/6) +(6)(1/6) =

21/6 = 3.5

2. Varianza y desviación estándar

La varianza de una distribución de probabilidad es el promedio de las diferencias cuadradas que hay entre cada valor y la media.

Nota: Las siguientes son diez observaciones de un experimento:

2.1 2.1 2.1 3.2 3.2 4 4 4 4.3

La media que se obtuvo fue 3.11 y la varianza debe ser:

$[(2.1-3.11)^2 +(2.13.11)^2 + (2.1-3.11)^2 + (2.1-3.11)^2 + (3.2-3.11)^2 +(3.2-3.11)^2 + (4-3.11)^2 + (4-3.11)^2 + (4-3.11)^2 +(4.\ 3-3.11)^2]/\ 10$

= 9.6721

que no es otra cosa que la fórmula de la varianza $(\sigma_X)^2$

$$\sigma_X^2 = \sum_{i-1}^{N} \left(X_i - \mu_X\right)^2 P\left(X_i\right)$$

donde

X = variable

X_i = i^{th} valor de X

$P(X_i)$ = probabilidad de X_i

i = 1,2,3,...N

La desviación estándar es la raíz cuadrada positiva de la varianza

$$\sigma_X = \sqrt{\sigma_X^2}$$

III. Modelos matemáticos

Se ha señalado que una distribución de probabilidad de una variable discreta no es otra cosa que una tabla donde aparece el valor de la variable y su probabilidad. Estas distribuciones de probabilidad generalmente surgen de observaciones o de fenómenos cuyas leyes se conocen bien como en el caso de los dados. Cuando el fenómeno que se quiere trabajar se ha observado cuidadosamente se puede crear un MODELO MATEMÁTICO que lo describe. Esta es la función más importante de las matemáticas, facilitar el estudio de la

naturaleza con modelos que la interpretan. Por ejemplo, la forma de decir que las cosas en un grupo miden el doble de las de otro grupo se expresa como la función

$$y = 2x$$

Lo mismo sucede con las distribuciones de probabilidad. Hay algunas para las que hay modelos matemáticos que evitan el trabajo de calcular todo lo que se ha estado haciendo hasta ahora. Estos modelos o funciones se llaman Funciones de distribuciones de probabilidad. Hay modelos para variables discretas y para variables continuas. De entre todos se van a estudiar dos ejemplos, la función binomial para una variable discreta y la función normal para una variable continua.

A. La distribución binomial

1. El modelo

La función binomial se puede utilizar como modelo solamente cuando las observaciones son independientes unas de otras y cada observación se puede clasificar como un éxito o un fracaso.

Ejemplos:

Sacar bolas blancas o negras de una urna; sacar cara o cruz al lanzar una moneda; escoger una respuesta en una pregunta de selección múltiple cuando uno no sabe nada.

Ejemplo:

En un experimento se llama éxito al hecho de obtener un 5 cuando se lanzan dos dados. ¿Cuál es la probabilidad de obtener un cinco?¿Cuál es la probabilidad de obtener dos cincos? Al crear la distribución de probabilidad de lanzar dos dados se obtiene la siguiente tabla:

Valores	combinaciones posibles	Prob
2	(1,1)	1/36
3	(1,2),(2,1)	2/36
4	(1,3),(3,1),(2,2)	3/36
5	(1,4),(4,I),(3,2),(2,3)	4/36
6	(1,5),(5,1),(2,4),(4,2),(3,3)	5/36

7	(1,6),(6,1),(2,5),(5,2),(3,4),(4,3)	6/36
8	(2,6),(6,2),(3,5),(5,3),(4,4)	5/36
9	(3,6),(6,3),(4,5),(5,4)	4/36
10	(4,6),(6,4),(5,5)	3/36
11	(5,6),(6,5)	2/36
12	(6,6)	1/36

Los únicos eventos de éxito son:

(1,5), (5,1), (2,5), (5,2), (3,5), (5,3), (4,5), (5,4), (5,5), (5,6), (6,5).

Por lo tanto la probabilidad de **2 éxitos** es 1/36 = 0.028; de **un éxito** es 10/36 = 0.278; y de **ningún éxito** es 25/36 = 0.694. Este mismo resultado se puede obtener sin necesidad de crear la tabla, pero utilizando el modelo matemático de la **distribución binomial**

$$P(X = x / n,p) = \frac{n!}{x!(n-x)!} P^x (1-P)^{n-x}$$

donde

P (X = x/n,p) es la probabilidad de que X = x

Cuando se conocen p y n

n = tamaño de la muestra

p = probabilidad de éxito

1 - p = probabilidad de fracaso

x = número de éxitos en la muestra.

En el ejemplo anterior en vez de hacer el trabajo intuitivamente, se puede utilizar la fórmula de la distribución binomial

n = 2 (dos dados); p = 1/6 = 0.17;

1-p = 5/6 = 0.83

x = 0; x = 1; x = 2

P (no éxito) = P(0) = (2!)/(2!0!)[$(0.17)^0(0.83)^2$] = 0.694

P(1 éxito) = P(1) = (2!)/(1!1!)[$(0.17)^1(0.83)^1$] = 0.278

P (2 éxitos) = P(2) = (2!)/(0!2!)[$(0.17)^2(0.83)^0$] = 0.028

2. Características

Media

$E(X) = \mu_X = np$

Desviación estándar

$$\sigma = \sqrt{np(1-p)}$$

B. Distribución normal

1. El modelo matemático

La distribución binomial y otras como la Poisson son modelos matemáticos que se utilizan cuando la variable es discreta y satisface los requisitos del modelo. Si la variable es continua entonces se usa principalmente la distribución normal. La gran diferencia entre ambas es que con variables discretas siempre es posible hallar la probabilidad de un valor dado puesto que los valores de la variable son discretos. Sin embargo cuando la variable es continua sólo se puede hallar la probabilidad de un intervalo dado. Esta probabilidad se visualiza siempre como el área debajo de la curva que representa la distribución.

a. Importancia de la distribución normal

i. Es el modelo matemático por excelencia en muchas situaciones de la vida real

ii. Sirve para aproximar la binomial y otras distribuciones discretas

iii. Es la base de la inferencia estadística

b. Propiedades de la distribución normal

i. Es simétrica y tiene forma de campana

ii. Las medidas de tendencia central son iguales

iii. Q_I y Q_3 están situados a 2/3 de una desviación estándar. El 68 % del área de la curva (probabilidad) se encuentra a una desviación estándar de la media.

iv. La variable tiene un alcance infinito.

c. El modelo matemático

La función matemática que se usa como modelo es:

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma_X} e^{-(1/2)[(X-\mu_X)/\sigma_X]^2}$$

donde:

e = 2.71

π = 3.14

μ_X = media de la población

σ_X = desviación estándar de la población

x = un valor de la variable continua

Como e y π son constantes, la forma de la curva normal depende solamente de los dos parámetros de la distribución normal, la media μ_X y la desviación estándar σ_X. Las curvas normales varían dependiendo de estos dos parámetros. En matemáticas, el área debajo de la curva se halla por medio del integral de la función. Para evitar el proceso de hallar el integral, en estadísticas se utilizan tablas que ya traen el área de diferentes secciones de la curva.

Puesto que hay un número infinito de combinaciones para los dos parámetros, hay un número infinito de curvas normales diferentes.

Este problema se ha resuelto prácticamente cuando se transforman todas estas posibles curvas normales en una llamada la curva normal estandarizada. (**standard normal curve**).

d. La curva normal estandarizada

Propiedades

i. $\mu_Z = 0$

ii. $\sigma_Z = 1$

iii. El área bajo la curva que aparece en las tablas corresponde a la probabilidad (Hinkle p.618)

Nota:

Obsérvese que el área bajo la curva corresponde ahora a la probabilidad, de igual manera que en conferencias anteriores correspondía al porcentaje o la proporción de puntuaciones en el intervalo dado. Por lo tanto la búsqueda de la probabilidad es exactamente igual a la búsqueda del porcentaje o proporción de puntuaciones en un intervalo dado.

iv. Cualquier variable normal puede ser transformada en la normal estandarizada por medio de la siguiente fórmula:

$$Z = \frac{X - \mu_X}{\sigma_X}$$

donde μ_X es la media de la distribución y σ_X su desviación estándar.

Ejemplos:

En una fábrica el tiempo que le toma a un trabajador ensamblar una pieza está distribuido normalmente con $\mu_X = 50$ seg. y $\sigma_X = 7$ seg.

Ejemplo 1:

¿Cuál es la probabilidad de que un obrero pase entre 50 y 57 segundos

ensamblando la pieza?

a. Primero hay que asegurarse que la variable está distribuida normalmente (Lo dice el problema) y trazar la gráfica que va a servir de modelo para el problema.

b. Después hay que convertir la escala dada a la escala de la distribución normal estándar usando la fórmula

$z_1 = (50-50)/7 = 0$

$z_2 = (57-50)/7 = 1$

c. Por medio de la tabla se observa que el área entre 0 y 1 es 0.3413

d. Como probabilidad, porcentaje y área bajo la curva significan lo mismo, es posible decir que:

ie: La probabilidad de que un obrero seleccionado al azar ensamble la pieza en ese tiempo es 0.3413

ie: El porcentaje de obreros que pueden ensamblar esa pieza en ese lapso de tiempo es 34.13 %

ie: De cada 100 obreros cerca de 34 pueden ensamblar la pieza en ese lapso de tiempo

Ejemplo 2:

Halla $P(x \leq 45)$

$z_1 = (45-50)/7 = -0.71$

En la tabla el área correspondiente es 0.2389 (beyond z)

$P(z \leq 45) = 0.2389$ (beyond z)

Dada la probabilidad o el porcentaje de obreros, se puede hallar el tiempo

Ejemplo 3:

¿Cuánto tiempo debe pasar antes que 50% de los obreros puedan ensamblar una pieza?

La media, 50 segundos.

Ejemplo 4:

¿Cuánto tiempo pasará antes que 10% de los obreros pueda ensamblar una pieza?

Probabilidad es 0.1 y el número que más se aproxima en la tabla en términos de área es 0.1003 (beyond z)

$z_1 = -1.28$

Por lo tanto $-1.28 = (x_1 - 50)/7$

$x_1 = 41.04$

ie: $P_{10} = 41.04$

Actividades:

Hinkle pp. 186 ej. 11,12

Lectura:

Hinkle capt. 7 pp. 152-170

MENU 6390

EDUC 6390: Estadística aplicada en la educación

Andrés Menéndez Raymat, Ph.D.

Conferencia 12

Muestras y la distribución muestral

Bosquejo

_

E. Teorema Central del Límite

III. Clarificación de conceptos por medio de ejemplos.

I. Muestras

A. Representatividad

Hay dos formas básicas de obtener información sobre una población: A través de un censo, donde se recoge información sobre toda la población. A través de una muestra, donde se recoge información sobre una muestra y se sacan conclusiones sobre la población.

Para hacer estas inferencias de la muestra a la población es imprescindible que la muestra sea representativa de la población. La muestra debe tener las mismas características de la población que se quiere estudiar.

B. Muestras probabilísticas y no probabilísticas

Hay dos tipos de muestras que se utilizan en los estudios: muestras probabilísticas y no probabilísticas. Las no-probabilísticas se basan principalmente en el juicio de expertos y se estudian en cursos especiales de muestreo. De las probabilísticas hay 4 tipos y se basan en la selección aleatoria.

Para llevar a cabo una selección aleatoria se pueden utilizar varios instrumentos que garanticen que la selección sea al azar. Puede ser: tirar una moneda al aire; sacar nombres de una urna; utilizar una tabla de números aleatorios; utilizar un programa computadorizado para hacer la selección.

La característica más importante de las muestras probabilísticas es que ellas garantizan que se puedan usar las técnicas de la estadística inferencial. Pero el hecho de que una muestra sea probabilística no garantiza que sea representativa de la población, el azar puede jugar malas pasadas. Sin embargo, sí se garantiza que la falta de representatividad, de haberla, no es intencional.

C. Tipos de muestras

1. Muestra aleatoria simple

La muestra aleatoria simple es el tipo más sencillo de muestra aleatoria, pero es la base de todas las demás. Su función principal es que todos los miembros de la población de la que se toma la muestra tengan la misma probabilidad de ser seleccionados para formar parte de la muestra. Hay dos tipos de muestras aleatorias simples, con reposición y sin reposición

a. Con reposición

Cuando se selecciona un miembro éste se devuelve a la población antes de seleccionar al próximo, de forma que la misma persona puede ser seleccionada en varias ocasiones. Ejemplo: Tomar una carta de un paquete de 52 y devolverla antes de seleccionar la próxima. Así cada carta en una muestra de 5 tiene 1/52 probabilidad de ser seleccionada.

b. Sin reposición

Cuando se selecciona un miembro de la población no se devuelve de manera que una persona u objeto puede ser seleccionado una sola vez. Sucede entonces que la probabilidad de selección es más pequeña a medida que se va construyendo la muestra.

Ejemplo:

Si no se devuelve la carta, la segunda tiene 1/51 probabilidad de ser seleccionada. La tercera 1/50 etc. En las ciencias sociales se utiliza principalmente este método de selección aleatoria, pues no se puede utilizar el mismo individuo más de una vez en las muestras. En este caso se logra que todas las posibles muestras del mismo tamaño tengan la misma probabilidad de ser seleccionadas. La diferencia en términos de probabilidad entre la muestra con y sin reposición es insignificante cuando la población es grande.

Ejemplo: En una población de 500 personas no hay mucha diferencia entre 1/500 = 0.002 y 1/499 = 0.002004008

Sin embargo, muy a menudo se encuentran investigaciones sociales y educacionales en que no se usan muestras aleatorias a pesar de que parezcan serlo. Un ejemplo interesante y muy frecuente es el de los estudios en que se envían cuestionarios a una muestra aleatoria de sujetos. Sin embargo, el estudio no resulta ser aleatorio, pues sólo los voluntarios responden al cuestionario.

Inconvenientes

Las muestras aleatorias no son adecuadas cuando hay dentro de la población bajo estudio se encuentran subpoblaciones de las que se quiere tener información. Cuando esto ocurre la información que se obtiene concierne a la población, pero no dice gran cosa sobre las subpoblaciones que la componen.

Las muestras aleatorias son imprácticas cuando los miembros de la población están alejados geográficamente unos de otros. En estos casos se hace casi imposible para los investigadores acceder a todos los sujetos de la muestra.

2. Muestra aleatoria sistemática

La muestra aleatoria sistemática se utiliza generalmente cuando existe una lista de la población preparada previamente. Esta lista puede ser la guía telefónica, una lista de miembros de un club, etc.

En estos casos el procedimiento que se lleva a cabo es el siguiente.

1. Se determina la fracción de muestreo. Se divide el tamaño de la muestra entre el tamaño de la población. Si la población es de 1,500 sujetos y la muestra es de 300 sujetos la fracción de muestreo es (300/1500 = 1/50)

2. Se selecciona aleatoriamente un sujeto que esté en la primera fracción del grupo de la lista . (Supongamos que el sujeto seleccionado entre los primeros 50 haya sido el número 23).

3. Se seleccionan los demás sujetos como múltiplos de la fracción. (Se seleccionan los sujetos 73 [23 + 50]; 123 [23 + 100]; 173 [23 +150] etc.

4. Es importante asegurarse de que la lista no esté hecha siguiendo algún patrón cíclico que eventualmente haga que la selección no sea representativa.

3. Muestra de conglomerado (cluster)

En este tipo de muestra, la unidad bajo investigación no es el individuo sino el grupo. Se utiliza principalmente cuando los miembros de la población están muy separados geográficamente. Una vez se selecciona un grupo, se utilizan todos los individuos que forman ese grupo.

Ejemplo:

Una muestra de conglomerado es aquella en que se seleccionan escuelas

aleatoriamente en un distrito y entonces todos los maestros de las escuelas seleccionadas participan.

Inconvenientes

El problema con la muestra de conglomerado es que generalmente los grupos son más homogéneos en su interior que la población y entonces cada grupo no representa adecuadamente la población. Este es el caso de la selección de escuelas para hacer un estudio sobre los maestros. Dentro de cada escuela el grupo de maestros es más homogéneo en términos socioeconómicos. Sin embargo, este problema se puede reducir si se seleccionan muchos conglomerados.

4. Muestra aleatoria estratificada

La muestra aleatoria estratificada se utiliza cuando en la población hay subgrupos de los cuales se quiere obtener información.

En estos casos el procedimiento que se lleva a cabo es el siguiente.

1. Se determina el porcentaje que representa cada subgrupo dentro de la población.

2. Se seleccionan aleatoriamente, de cada subgrupo, tantos sujetos como sean necesarios para que en la muestra total haya el mismo porcentaje de cada subgrupo que hay en la población

3. Esto se llama una alocación proporcional

II. La distribución muestral de la media (sampling distribution)

A. Concepto

Para obtener información sobre una población por lo general sólo tenemos la información que nos provee una muestra. Pero esta muestra, por muy aleatoria que sea sólo provee información sobre parte de la población.

El error que se genera al estimar un parámetro partiendo de una estadística es lo que se llama el **error de muestreo**. La única forma científica de poder hacer una inferencia de la muestra a la población requiere que se construya la

distribución muestral de la media. Esto se logra por medio del siguiente proceso:

1. Se seleccionan en una población todas las muestras posibles (de un tamaño dado)

2. Se halla la media de cada muestra.

3. Se construye una nueva distribución con todas las medias obtenidas.

4. Esta nueva distribución de medias se llama la distribución muestral (sampling distribution)

B. Ejemplo

Dada una población de sólo 3 personas A tiene $20; B tiene $15 y C tiene $10. Construye la distribución muestral de la media de dos personas (la distribución formada por las medias de todas las muestras de dos personas):

media de A y B = 17.5

media de A y C = 15

media de B y C = 12.5

La distribución muestral de la media es {17.5, 15, 12.5}

Notación:

es la media de una de las muestras

μ_X es la media de la población

es la media de la distribución muestral de la media (la media de todas las medias)

C. Propiedades de la media

1. Insesgada **(unbiased)**

Se dice que la media es insesgada porque la media de todas las medias de todas las muestras de un tamaño dado es igual a la media de la población

Ejemplo:

Dada una población de sólo 3 personas donde A tiene $20; B tiene $15 y C tiene $10. La media de la población es $15. La media de la distribución muestral de la media de dos personas es (17.5+15+12.5)/3 = 15

2. Eficiente

La media es la medida de tendencia central que menos cambia de muestra en muestra. Hay más diferencias entre las modas o entre las medianas de las diversas muestras.

3. Consistente

A medida que el tamaño de la muestra aumenta la media de las muestras se acerca más a la media de la población.

D. Error estándar de la media

1. Definición

La distribución muestral de la media tiene también una desviación estándar que representa la variabilidad de las medias de todas las muestras de un tamaño dado. Esta desviación estándar se llama **error estándar de la media** y se representa con el siguiente símbolo.

2. Notación

El símbolo "s" representa la desviación estándar de una muestra.

El símbolo " σ_x " representa la desviación estándar de la población.

El símbolo representa la desviación estándar de la distribución muestral de la media (La desviación estándar de todas las medias)

3. Relación entre " σ_x " y

Hay mucha más variabilidad en la población que en la distribución muestral. Esto se debe a que al calcular $\sigma_{\bar{x}}$ es necesario incluir el valor mínimo y el máximo de la población. Sin embargo, estos valores máximos y mínimos no se tienen que incluir en el cálculo de la distribución muestral ni en el del error estándar de la media.

Ejemplo:

Dada una población de sólo 3 personas donde A tiene $20; B tiene $15 y C tiene $10. La distribución muestral de la media de dos personas (las medias de todas las muestras de dos personas) es {17.5, 15, 12.5}

media de A y B = 17.5

media de A y C = 15

media de B y C = 12.5

$\sigma_{\bar{x}} = 4.082$

Hay dos fórmulas para relacionar $\sigma_{\bar{x}}$ y

Cuando el muestreo es aleatorio **sin reposición**:

Cuando el muestreo es aleatorio con reposición o la población es infinita entonces:

Esta última fórmula se usa también cuando la muestra es pequeña comparada con la población. Como éste es generalmente el caso en ciencias sociales, por lo general se usa esta fórmula.

E. Teorema Central del Límite

Si la población subyacente es normal con media $\mu_{\bar{x}}$, desviación estándar $\sigma_{\bar{X}}$ y el muestreo es aleatorio con reposición entonces la distribución muestral de la media para cualquier tamaño de muestra es normal y

además

Cuando la población subyacente no es normal se puede aplicar el **TEOREMA**

CENTRAL DEL LÍMITE que dice lo siguiente:

Si el tamaño de la muestra es suficientemente grande, entonces la distribución muestral de la media se puede aproximar por medio de la distribución normal.

Este teorema se cumple independientemente de la forma de la distribución de la población subyacente.

1. Para cualquier distribución una muestra de 30 ó más es suficientemente grande para podérsele aplicar el teorema.

2. Si la población subyacente es normal, entonces la distribución muestral de la media es normal para cualquier tamaño de muestra.

III. Clarificación de conceptos por medio de ejemplos.

Todos los ejemplos que siguen a continuación están basados en el siguiente caso:

En una fábrica de cereales la cantidad de cereal que se pone dentro de una caja está normalmente distribuida y tiene una media de 368 gramos y una desviación estándar de 15 gramos. Se hacen 10,000 cajas de cereal diariamente. Si se quiere ejercer un control de calidad se selecciona una muestra de 25 cajas cada cierto tiempo y se pesa cada caja para ver si la máquina empacadora funciona bien

El propósito del experimento del control de calidad es tomar muestras de 25 cajas de la población y determinar si la media de cada una de estas muestras no se encuentra muy lejos de la media hipotética de la población (368 grs). Obviamente, las muestras van a tener medias diferentes entre sí y diferentes de la media de la población, pero si se encuentran dentro de unos límites razonables con respecto a la media de la población, es posible achacar la diferencia entre la media de la muestra y la media de la población a la selección aleatoria. Si la media de la muestra se encuentra dentro de estos límites no hay porqué poner en duda que la media de la población sea 368 grs.

Sin embargo, si la media de la muestra es muy diferente de la media hipotética de la población (368 grs) entonces es posible que la diferencia no se deba sólo a la selección aleatoria, sino que la máquina ha dejado de funcionar adecuadamente y la media de la población ha cambiado. Cuando esto ocurre con varias muestras es necesario detener la producción y arreglar la máquina.

Por lo general el problema va a ser hallar el intervalo dentro del cual el 90% o el 95% de las medias de las muestras de 25 cajas deben caer para estar seguros de que la máquina está funcionando bien.

De acuerdo al Teorema Central del Límite, con N = 10,000

Como la distribución de la población es normal es posible hacer preguntas sobre cada caja individualmente o sobre cada muestra de 25 cajas. En el primer caso se utilizaría la distribución de la población subyacente y en el otro la distribución muestral.

Ejemplo 1: ¿Cuál es la probabilidad que **la media de una caja** se encuentre entre 365 y 368 gramos?

En este caso se habla de una sola caja, por lo tanto se utiliza la desviación estándar de la población para obtener el valor de z que se va a utilizar en las tablas para la distribución normal.

$$z = \frac{[365-368]}{15} = -3/15 = -0.2$$

$$P(365 < x < 368) = P(-0.2 < z < 0) = 0.0793$$

Por lo tanto se puede decir que la probabilidad de que la media de una caja se encuentre entre 365 y 368 gramos es 0.0793.

También se puede decir que el 7.93 % de las cajas tiene una media entre 365 y 368 gramos.

Sin embargo algo muy diferente sucede si examinamos cada muestra de 25 cajas.

Ejemplo 2: ¿Cuál es la probabilidad que la media de una muestra de 25 cajas se encuentre entre 365 y 368 gramos?

En este caso se habla de una muestra de 25 cajas, por lo tanto se utiliza la desviación estándar de la distribución muestral ($15/\sqrt{25} = 3$) para obtener el valor de z que se va a utilizar en las tablas para la distribución normal.

$$z = \frac{[365-368]}{3} = -3/3 = -1$$

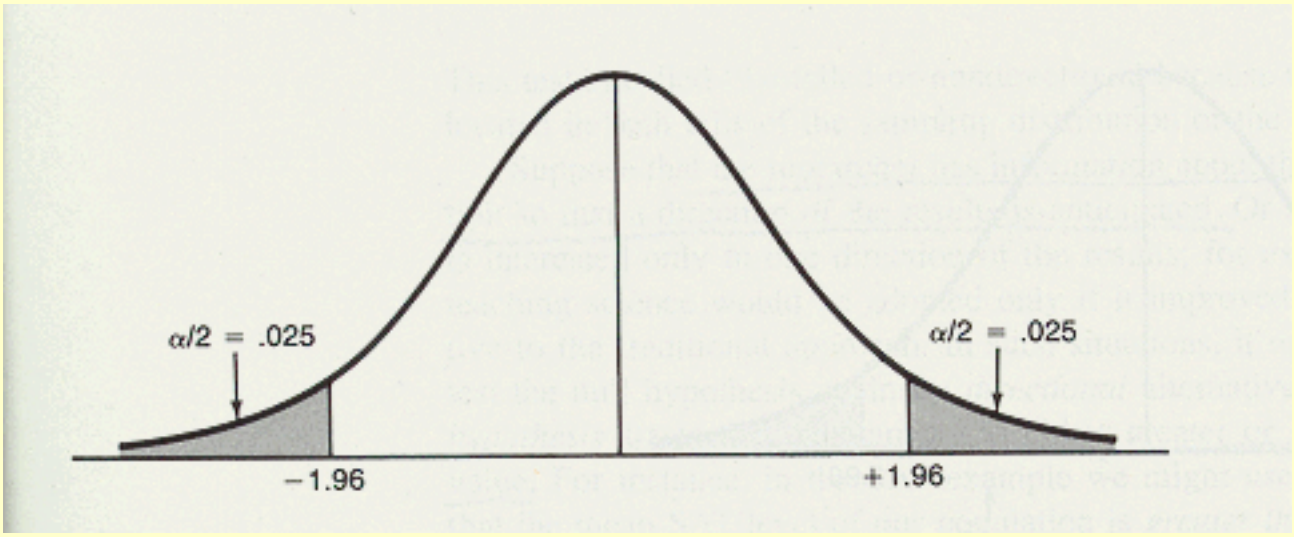
P(365 < x < 368) = P(-1 < z < 0) = 0.3413

Por lo tanto se puede decir que la probabilidad de que la media de una muestra de 25 cajas se encuentre entre 365 y 368 gramos es 0.3413.

También se puede decir que el 34.13 % de las cajas tiene una media entre 365 y 368 gramos.

El problema más importante va a ser hallar el intervalo dentro del cual el 90% o el 95% de las medias de las muestras de 25 cajas deben caer para estar seguros de que la máquina está funcionando bien.

Ejemplo 3: Halla el intervalo dentro del cual el 95% de las medias de la distribución muestral debe estar si la media de la población es 368. Visualmente se busca el área bajo la curva que contiene el 95% del área total. En otras palabras se desea obtener un área igual al 47.5% de cada lado de la media y determinar la z que corresponde a dicha área.



El área de 0.4750 bajo la curva corresponde en la tabla a z = 1.96. Por lo tanto

z = 1.96 y z = -1.96

1.96 = (x -368)/3

5.88 = x - 368

x = 373.88

-1.96 = (x -368)/3

-5.88 = x - 368

x = 362.12

Por lo tanto es posible decir que el 95% de las medias de las muestras de 25 cajas deben estar entre 362.12 y 373.88 gramos.

Lectura:

Hinkle capítulo 7 pp.171-184

Actividades:

Hinkle pp. 184 ej. 15,16,17,18

MENU 6390

EDUC 6390: Estadística aplicada en la educación

Andrés Menéndez Raymat, Ph.D.

Conferencia 13

La prueba paramétrica de hipótesis

Bosquejo

I. Introducción

El propósito principal de esta unidad es desarrollar una metodología para inferir si la estadística obtenida en una muestra corresponde al parámetro que se propone en la hipótesis.

II. Metodología

En toda la unidad se utilizará el ejemplo que se había presentado en la unidad anterior sobre la fábrica:

En una fábrica de cereales la cantidad de cereal que se pone dentro de una caja está normalmente distribuida y tiene una media de 368 gramos y una desviación estándar de 15 gramos. Se hacen 10,000 cajas de cereal diariamente. Si se quiere ejercer un control de calidad se selecciona una muestra de 25 cajas cada cierto tiempo y se pesa cada caja para ver si la máquina empacadora funciona bien

La investigación se verá como un proceso de control de calidad donde los resultados de la evaluación pueden ser dos:

1. Continuar la producción si la evidencia indica que el promedio es 368 gr.

2. Detener la producción y arreglar la máquina si la evidencia indica que el promedio no es 368 gr.

A. La hipótesis nula y la hipótesis alterna

La prueba de hipótesis siempre comienza planteando que un parámetro dado de la población es cierto. Este planteamiento se llama la hipótesis nula y se usa el símbolo H_0 para referirse a ella. En el caso del ejemplo la hipótesis nula se escribe:

$H_0: \mu_x = 368$

Esta hipótesis se considera cierta hasta que se haya encontrado evidencia indicando que es falsa. Es importante notar que la lógica de la estadística es que: **Nunca se prueba nada**. Solamente se acumula evidencia contra la hipótesis nula para rechazarla o no rechazarla. El vocabulario estadístico será:

Hay suficiente evidencia para rechazar H_0

No hay suficiente evidencia para rechazar H_0

La hipótesis alterna se simboliza con H_1 y representa el opuesto de la hipótesis nula. Es el opuesto absoluto de H_0. Así que cuando se rechaza una hipótesis se sustenta la otra. H_1 generalmente representa lo que el investigador quiere demostrar. En el caso del ejemplo la hipótesis alterna se escribe:

$H_1: \mu_x \neq 368$

En el ejemplo de la línea de producción si la media de la muestra está muy por arriba o muy por debajo de la media propuesta en H_0 entonces se rechaza H_0 y se sustenta H_1. Pero tiene que quedar claro que el hecho de que se rechace H_0 no es prueba de que H_1 sea cierta o de que H_0 sea falsa. Solamente se ha acumulado evidencia a favor de una en contra de la otra. Siempre va a haber un margen de error con respecto a la decisión que se tome. El lenguaje es bien importante. Siempre se dirá:

Se rechaza H_0 a favor de H_1

No se rechaza H_0 pues no hay suficiente evidencia para justificar su rechazo

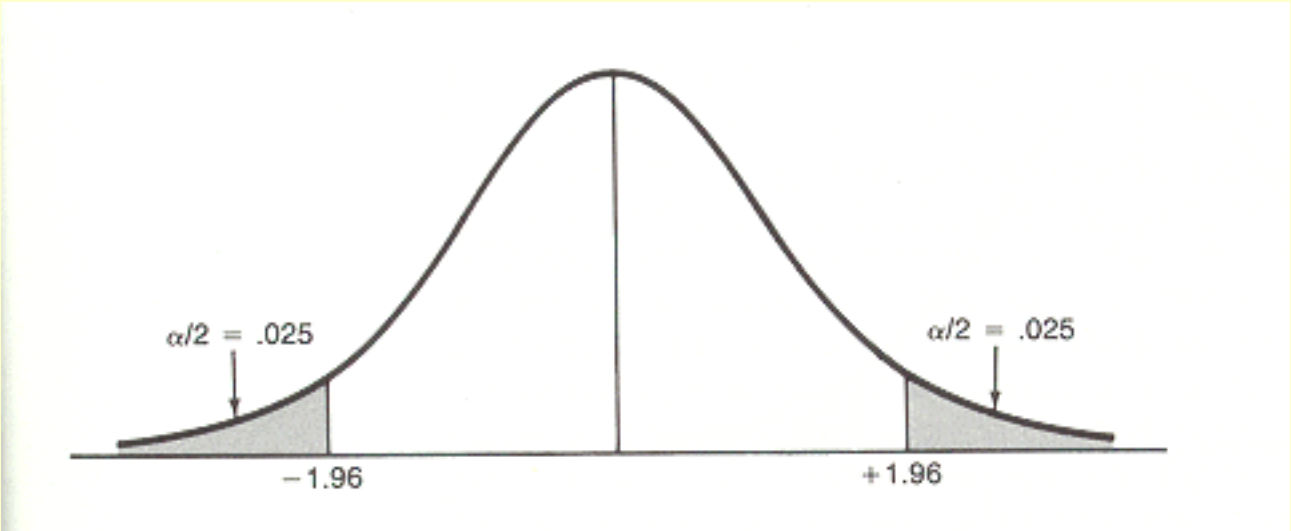
Nunca se dice que se ha probado nada. Para probar algo en matemáticas es imprescindible asegurarse que siempre va a ser cierto. En estadísticas, sin embargo, siempre va a haber un margen de error. **Nunca se dice que se acepta la hipótesis nula o la alterna**, puesto que aceptar es mucho más que no rechazar.

B. La región crítica

Simplemente observando la media de la muestra de 25 cajas es posible tener una idea si la media de la población es la planteada por H_0. Si la media de la muestra fuese 367.8 gr los investigadores estarían dispuestos a decir que la media de la población es 368 gr como propone H_0, ya que la diferencia entre la media de la muestra y la propuesta para la población no es muy grande. Sin embargo, si la media de la muestra es 320 gr., entonces sería muy difícil no rechazar H_0.

Región de rechazo y región de no rechazo

Para determinar **operacionalmente** cuando se rechaza y cuando no se rechaza la hipótesis nula es necesario estudiar la metodología de la prueba de hipótesis basándose en la distribución muestral de la media. La distribución muestral de la media generalmente sigue la distribución normal y es importante determinar cuál debe ser la **región de rechazo** (rejection region) o región crítica y cuál la **región de no rechazo** (nonrejection region). Generalmente la región de no rechazo cubre el 95% del área alrededor de la media y la región crítica el 5% de los extremos. El trabajo estadístico previo consiste en determinar con qué valor de X comienza la región crítica.



z = valor crítico

zona blanca = región de no rechazo

zona gris = región de rechazo

Nivel de confianza y nivel de significación

Si H_0 es cierta, en otras palabras, si la media de la población es 368 y se tomaran todas las posibles muestras de 25 cajas, entonces el 95% de esas muestras tendría su media entre los valores críticos. Esta región de no rechazo generalmente es de 95% ó 99%. El porcentaje correspondiente a la región de no rechazo se denomina el **nivel de confianza** y se escribe (1-α) donde α representa el porcentaje correspondiente a las colas (tails) y el porcentaje correspondiente a la región de rechazo (α) se llama el **nivel de significación** (level of significance)

C. El error en las pruebas de hipótesis

Cuando se decide rechazar o no una hipótesis nula se dan cuatro posibles situaciones:

1. H_0 es cierta y se rechaza (**error tipo I**)

2. H_0 es cierta y no se rechaza

3. H_0 es falsa y se rechaza (**1- β, poder)**

4. H_0 es falsa y no se rechaza (**error tipo II;** β)

Con las opciones 2 y 3 se ha tomado la decisión correcta, pero con las opciones 1 y 4 se ha cometido un error. Cada uno de estos errores es diferente y es por sus consecuencias que podemos detectar la diferencia.

Error tipo I

H_0 es cierta y se rechaza

Ejemplo:

H_0: Una medicina sirve para curar una enfermedad.

Si H_0 es cierta pero se rechaza, se condena a los enfermos a seguir sin una buena medicina.

Error tipo II

H_0 es falsa y no se rechaza

H_0: Una medicina sirve para curar una enfermedad.

Si H_0 es falsa y no se rechaza se está condenando a los enfermos a pagar por una medicina que no va a curarlos.

¿Cuál error es más importante? Depende de la situación.

H_0 : Un método muy caro de enseñanza promueve el aprovechamiento

Error tipo I:

H_0 es cierta y se rechaza.

No se implanta el método y no se promueve el aprovechamiento

Error tipo II:

H_0 es falsa y no se rechaza.

Se implanta el método y se gasta mucho sin lograr nada.

D. Pasos en la prueba de hipótesis

Ejemplo

[Cuando σ_X (desviación estándar de la población) es conocida]

En una fábrica se producen 10,000 cajas de cereal por día y se sabe que la media es 368 gr. y la desviación estándar es 15 gr. Se toma una muestra de 25 cajas y su media es 372.5 gr. ¿Podría decir el experto en control de calidad que la máquina está funcionando correctamente?

Pasos 1 y 2: PLANTEAR LAS DOS HIPÓTESIS

$$H_0: \mu_x = 368$$

$$H_1: \mu_x \neq 368$$

Paso 3: DETERMINAR EL NIVEL DE SIGNIFICACIÓN α

$$\alpha = 0.05$$

Paso 4: SELECCIONAR EL TAMAÑO DE LA MUESTRA

n = 25

Paso 5: SELECCIONAR LA PRUEBA ADECUADA

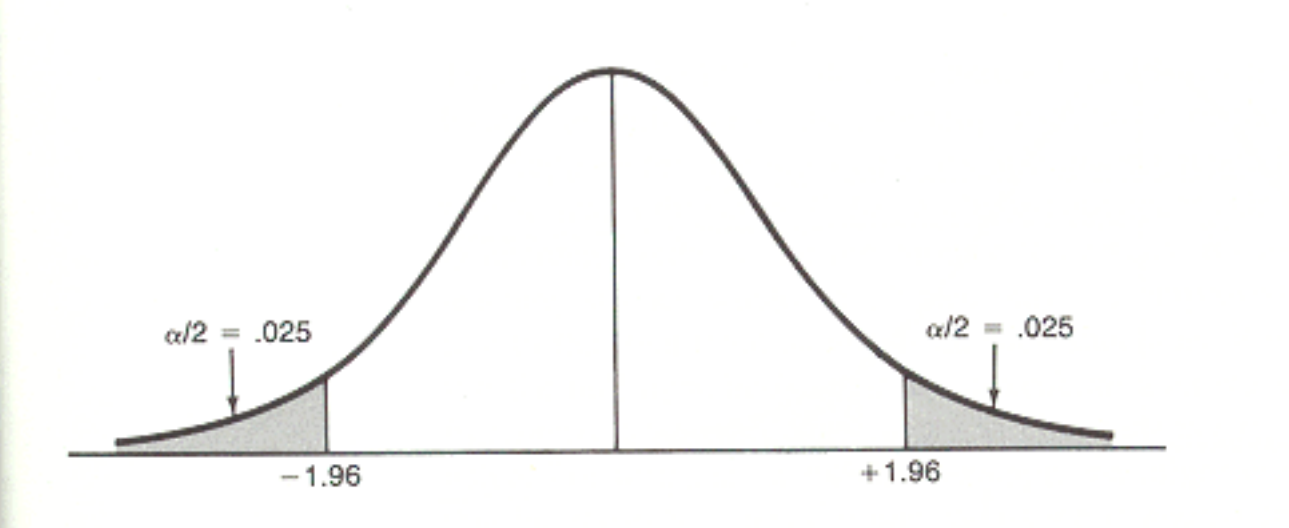
En este caso se conoce σ_X por lo tanto se usa la prueba z donde z_0 (z observada) es

$$z_o = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$$

Paso 6: DETERMINAR LOS VALORES CRÍTICOS

Para tener 5% como área de rechazo y 95% como área de no rechazo es necesario encontrar 0.4750 del área en la tabla. Esta área corresponde a los valores críticos

$z_c = 1.96$ y $z_c = -1.96$



Paso 7: ESCRIBIR LA REGLA DECISIONAL

Si $z_0 < -1.96$ ó si $z_0 > 1.96$ entonces se rechaza H_0

Si $-1.96 < z_0 < 1.96$ entonces no se rechaza H_0

z_0 se denomina la z observada o el valor observado de la estadística.

Paso 8: COMPUTAR z_o

$$z_o = \frac{\overline{X} - \mu_{\overline{X}}}{\frac{\sigma_X}{\sqrt{n}}} = \frac{372.5 - 368}{\frac{15}{\sqrt{25}}} = 1.50$$

Paso 9: DETERMINAR SI LA ESTADÍSTICA ESTÁ EN LA ZONA DE RECHAZO O NO

z_0 no se encuentra en la zona de rechazo, por lo tanto no se rechaza la hipótesis nula.

Paso 10: EXPRESAR LA DECISIÓN EN TÉRMINOS DEL PROBLEMA

Con un nivel de significación de 5% se puede decir que **NO HAY EVIDENCIA PARA CONCLUIR QUE LA MEDIA DE LAS CAJAS ES DIFERENTE DE 368 gr**.

Por lo tanto la máquina continúa funcionando adecuadamente y no hay que detener la producción para arreglar la máquina.

III. El p-value en la computadora

Con el advenimiento de las computadoras la prueba de hipótesis se ha facilitado. Ya no es necesario:

a. calcular la media de la muestra

b. determinar los valores críticos (paso 6)

c. indicar la regla decisional (paso 7)

d. computar la estadística (paso 8)

Todos estos pasos quedan sustituidos por la determinación del valor de p (p-value) que hace la computadora. Sin embargo es necesario interpretar correctamente este valor de p.

Significado de p

Si se toma como cierta la hipótesis nula, el valor de **p** representa la probabilidad de obtener una estadística igual o más alejada de la media (propuesta por H_0) que el valor obtenido en la muestra.

Ejemplo:

Cuando $\alpha = 0.05$ y $p = 0.02$ entonces se rechaza H_0

p = 0.02 significa que si la media de la población es 368 grs, entonces la probabilidad de obtener una muestra con una media de 380 grs (como la de la muestra) es muy pequeña, sólo 0.02. En otras palabras, la estadística se encuentra en la zona de rechazo.

Si p = 0.06 entonces no se rechaza H_0 puesto que la estadística se encuentra en la zona de no rechazo. En otras palabras, siempre que **p < α se rechaza H_0** y siempre **que p > α no se rechaza H_0**.

Visualmente se puede describir **p** como el área bajo los extremos de la curva.

IV. El intervalo de confianza

Otra forma de trabajar el mismo problema es utilizando el intervalo de confianza [**CI**] en vez de la prueba de hipótesis.

El propósito este método el razonamiento es determinar a qué poblaciones puede corresponder una muestra semejante a la que se ha obtenido. En vez de suponer que H_0 es cierta, simplemente se toma un intervalo alrededor de la media de la muestra.

En otras palabras, se construye un intervalo de confianza [**CI**] alrededor de la estadística observada. Es posible tener un cierto grado de confianza que este intervalo de confianza incluye el parámetro de la población. Partiendo de dónde se encuentra la media de la muestra se hacen inferencias sobre la media de la población.

A. La estimación de parámetros

Se utiliza la siguiente fórmula para determinar los valores críticos del intervalo

de confianza:

B. La prueba de hipótesis y el intervalo de confianza

Ejemplo:

En la fábrica donde se producen 10,000 cajas de cereal al día se sabe que la media es 368 gr. y la desviación estándar es 15 gr. Si se toma una muestra de 25 cajas y la media es 372.5 gr., ¿podría el experto en control de calidad decir que la máquina está funcionando correctamente?

El intervalo determinado por 372.5 es

$$372.5 \pm (1.96)(15/5)$$

$$372.5 \pm 5.88$$

Por lo tanto la media de la población debe estar en el intervalo

$$366.62 \leq \mu_{\bar{x}} \leq 378.38$$

Puesto que la media hipotética de la población es 368 gr., es posible concluir que:

Con un nivel de confianza de 95% se puede decir que **No hay evidencia partiendo de la muestra escogida para concluir que la media de las cajas es diferente de 368 gr.**

Es importante recordar que la muestra de 25 cajas es sólo una de las muchas muestras que se pudieron haber escogido, por lo tanto no hay pruebas, sólo se sabe que la muestra escogida apoya la hipótesis nula. Esta falta de pruebas es la que lleva al experto en control de calidad a seguir tomando muestras.

C. Interpretación del intervalo de confianza

Cuando se obtiene un intervalo de confianza de 95% como el anterior se puede decir que hay un 95% de confianza de que el intervalo contiene la media de la población. Sin embargo si se hubiese tomado otra muestra es casi seguro que el intervalo habría sido diferente puesto que la media de la muestra hubiera sido otra. En el ejemplo anterior:

$$372.5 \pm 5.88$$

Por lo tanto la media de la población parece estar en el intervalo

$$366.62 \leq \mu_x \leq 378.38$$

Si la media de la muestra hubiese sido 370 entonces el intervalo hubiera sido

$$370.5 \pm 5.88$$

$$364.12 \leq \mu_x \leq 375.88$$

El 95% de confianza quiere decir que si se construyeran todos los intervalos de todas las muestras de tamaño 25, entonces el 95% de estos intervalos contendrían la media de la población y 5 % no la contendría. Por lo tanto quizá el intervalo obtenido es uno del 5% y ha habido un error, pero es más posible que se haya obtenido uno del 95 %.

Es importante notar que **NO** se puede decir que hay una probabilidad de 95% de que el intervalo de confianza contenga la media. Eso es falso, el intervalo contiene la media de la población o no la contiene. **No se habla de probabilidad sino de confianza**.

El intervalo de confianza también puede verse como la forma de hacer la prueba a muchas hipótesis al mismo tiempo. Cualquier valor dentro del intervalo puede ser una hipótesis nula que se puede sostener, y cualquier valor fuera del intervalo seria una hipótesis nula que no se puede sostener.

D. El intervalo de confianza y la precisión estadística

La precisión estadística es la exactitud con la que se puede predecir un parámetro partiendo de una estadística.

Si se comparan los siguientes dos ejemplos se puede ver la relación entre los dos conceptos

Ejemplo 1:

En la fábrica donde se producen 10,000 cajas de cereal al día se toma una muestra de 25 cajas y la media es 372.5 gr., ¿cuál es el intervalo para un nivel de confianza de 95%?

El intervalo determinado por 372.5 es

$372.5 \pm (1.96)(15/5)$

372.5 ± 5.88

Por lo tanto la media de la población debe estar en el intervalo

$$\mathbf{366.62} \leq \mu_{\mathbf{x}} \leq \mathbf{378.38}$$

Ejemplo 2: En la fábrica donde se producen 10,000 cajas de cereal al día se toma una muestra de 25 cajas y la media es 372.5 gr., ¿cuál es el intervalo para un nivel de confianza de 90%?

El intervalo determinado por 372.5 es

$372.5 \pm (1.64)(15/5)$

372.5 ± 4.92

Por lo tanto la media de la población debe estar en el intervalo

$$\mathbf{367.58} \leq \mu_{\mathbf{x}} \leq \mathbf{377.42}$$

1. Si se reduce el nivel de confianza se aumenta la precisión estadística puesto que el intervalo correspondiente se ha hecho más pequeño. Al revés, si se aumenta el nivel de confianza se reduce la precisión estadística.

E. El tamaño de la muestra y la precisión estadística

Ejemplo 1:

En la fábrica donde se producen 10,000 cajas de cereal al día se toma una muestra de 25 cajas y la media es 372.5 gr., ¿cuál es el intervalo para un nivel de confianza de 95%?

El intervalo determinado por 372.5 es

372.5 ± (1.96)(15/5)

372.5 ± 5.88

Por lo tanto la media de la población debe estar en el intervalo

366.62 $\leq \mu_{\mathbf{x}} \leq$ **378.38**

Ejemplo 3:

En la fábrica donde se producen 10,000 cajas de cereal al día se toma una muestra de 49 cajas y la media es 372.5 gr., ¿cuál es el intervalo para un nivel de confianza de 95%?

El intervalo determinado por 372.5 es

372.5 ± (1.96)(15/7)

372.5 ± 4.2

Por lo tanto la media de la población debe estar en el intervalo

368.3 $\leq \mu_{\mathbf{x}} \leq$ **376.7**

2. Si se aumenta el tamaño de la muestra se aumenta la precisión estadística puesto que el intervalo correspondiente se ha hecho más pequeño. Al revés, si se reduce el tamaño de la muestra se reduce la precisión estadística.

3. Conclusión: Si se quiere aumentar la precisión sin sacrificar el nivel de confianza, sólo queda aumentar la muestra.

V. Pruebas de hipótesis de una cola. (one-tailed test)

En todos los ejemplos previos el interés ha sido determinar si la media de las cajas de cereal en la población es 368 gr. ó no. Pero a veces el propósito puede ser diferente y se busca saber si la media es más o menos de 368 gr. En el caso de un grupo de consumidores de cereal el interés es determinar que las cajas no tengan menos cereal del que se anuncia. Si, por el contrario, el

controlador de calidad representa los intereses del dueño de la compañía entonces su interés es que no se ponga cereal de más, lo que causaría pérdidas a la fábrica.

Por lo tanto es de suma importancia en todas las pruebas de una cola determinar la hipótesis nula. El razonamiento debe siempre partir de que la hipótesis alterna que es aquella que concuerda con los intereses del investigador. Es importante recordar que se sustenta la hipótesis alterna cuando se logra acumular evidencia contra la hipótesis nula.

Caso del consumidor

Un consumidor de cereal quiere acumular evidencia de que las cajas no se llenan lo suficiente. Por lo tanto su hipótesis alterna es que en la caja hay menos de lo que dice la compañía.

Por lo tanto contra la hipótesis alterna de

H_1: μ_x < 368 (los consumidores pierden)

Monta la hipótesis nula de

H_0: $\mu_x \geq$ 368

Por lo tanto él necesita evidencia contra el reclamo de la fábrica de que se echa tanto o más cereal:

Paso 1 y 2: H_0: $\mu_x \geq 368$ (el proceso está correcto)

H_1: $\mu_x < 368$ (los consumidores pierden)

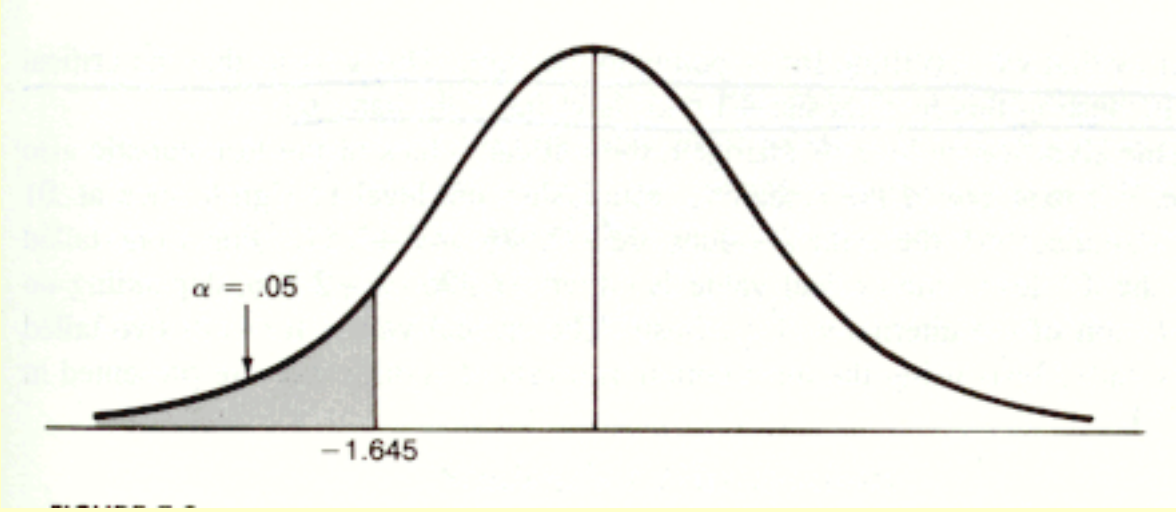
Paso 3: $\alpha = 0.05$

Paso 4: n = 25

Paso 5: En este caso donde σ_X se conoce se usa la prueba z

$$z_o = \frac{\overline{X} - \mu_{\overline{X}}}{\frac{\sigma_X}{\sqrt{n}}}$$

Paso 6: Para tener 5% del área en la zona de rechazo y 95% en la zona de no rechazo tenemos que hallar 0.45 del área en la tabla.



Esta área corresponde al valor crítico

$z_c = -1.645$

Paso 7: Si $z_0 < -1.645$ entonces se rechaza H_0

Si $z_0 > -1.645$ entonces no se rechaza H_0

Paso 8:

$$z_o = \frac{\overline{X} - \mu_{\overline{X}}}{\frac{\sigma_X}{\sqrt{n}}} = \frac{372.5 - 368}{\frac{15}{\sqrt{25}}} = 1.50$$

Paso 9: z_0 no cayó en la zona de rechazo, por lo tanto no se rechaza la hipótesis nula.

Paso 10: Con un nivel de significación de 5% se puede decir que **No hay evidencia para concluir que la media de las cajas es menos de 368 gr**.

Caso del ejecutivo

Un ejecutivo de la fábrica no quiere perder dinero y su planteamiento es que se está poniendo más cereal de la cuenta y por lo tanto su hipótesis alterna es:

H_1: $\mu_x > 368$ (la compañía pierde cereal)

Paso 1 and 2:

H_0: $\mu_x \leq 368$ (el proceso funciona)

H_1: $\mu_x > 368$ (la compañía pierde cereal)

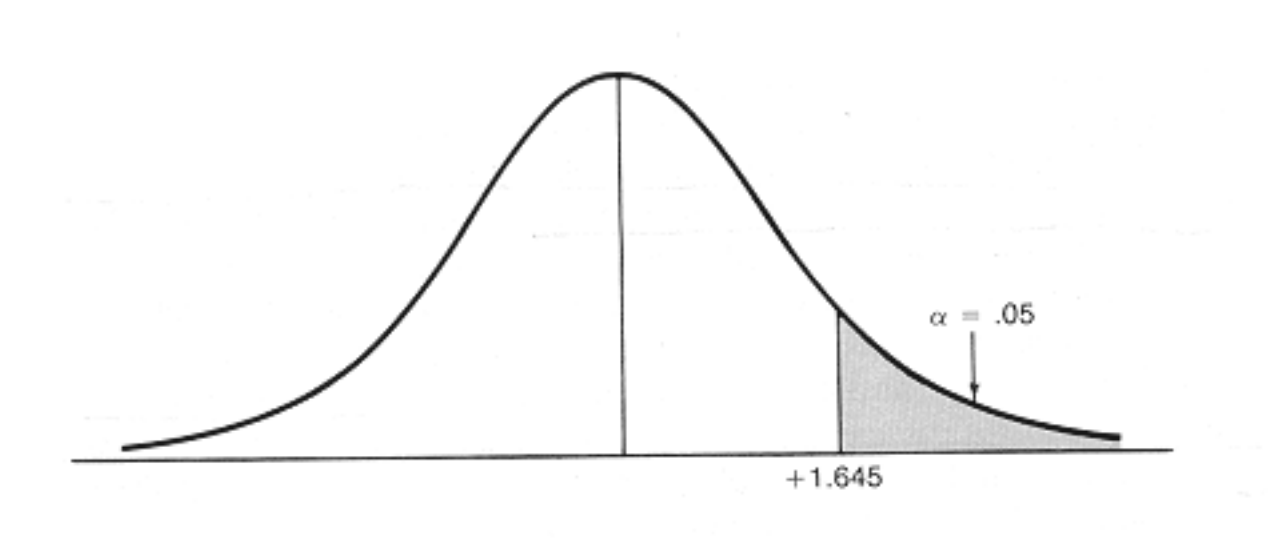
Paso 3: $\alpha = 0.05$

Paso 4: n = 25

Paso 5: En este caso σ_x se conoce y se usa la prueba z

$$z_o = \frac{\overline{X} - \mu_{\overline{X}}}{\frac{\sigma_X}{\sqrt{n}}}$$

Paso 6: Para que haya 5% del área en la zona de rechazo y 95% en la zona de no rechazo tenemos que encontrar 0.45 del área en la tabla.



Esta área corresponde al valor crítico

$z_c = 1.645$

Paso 7: Si $z_0 > 1.645$ entonces se rechaza H_0

Si $z_0 < 1.645$ entonces no se rechaza H_0

Paso 8:

$$z_o = \frac{\overline{X} - \mu_{\overline{X}}}{\frac{\sigma_X}{\sqrt{n}}} = \frac{372.5 - 368}{\frac{15}{\sqrt{25}}} = 1.50$$

Paso 9: z_0 no cayó en la zona de rechazo, por lo tanto no se rechaza la hipótesis nula

Paso 10: Con un nivel de significación de 5% se puede decir que **No hay evidencia para concluir que la media de las cajas es más de 368 gr.**

Lectura:

Hinkle capt. 8,9 pp.188-204; pp.217-226

Actividades:

Hinkle pp. 213 # 1,4a-e,6,7,8,13

Asignación:

Pruebas de hipótesis

MENU 6390

EDUC 6390: Estadística aplicada en la educación
Andrés Menéndez Raymat, Ph.D.
Conferencia 14
Prueba de hipótesis cuando σ_X es desconocida

Bosquejo

I. Introducción

En la unidad anterior se hizo la prueba de hipótesis cuando se conocía la desviación estándar de la población. Pero esto ocurre muy raramente. Por lo general, en los experimentos que se llevan a cabo se conoce la desviación estándar de la muestra solamente. En esos casos es necesario utilizar otra prueba estadística.

II. La prueba de t para la media

A. La familia de distribuciones t

Cuando no se conoce la desviación estándar de la población siempre se puede usar la desviación estándar de la muestra, pero dividida entre la raíz cuadrada de n.

$$\frac{s}{\sqrt{n}}$$

pero no se pueden usar las tablas de la distribución normal. Cuando se hace la sustitución, sobre todo para muestras pequeñas, la distribución muestral es diferente de la normal. Realmente se usa toda una familia de distribuciones semejante a la familia de las distribuciones normales, pero que tienen más área bajo los extremos y menos en el centro. (Hinkle, p. 207)

Sin embargo, a medida que el tamaño de la muestra aumenta las **distribuciones de t** se parecen más a la normal.

Por esta razón cuando el tamaño de la muestra es de más de 120 se deja de utilizar la tabla de t y se regresa a la tabla de z.

Sin embargo la tabla de t no es tan detallada como la de z. Sólo aparecen los valores críticos de ciertas áreas de rechazo. (Hinkle p.637, Tabla C3)

Cada una de las distribuciones de t está asociada con el tamaño de la muestra por medio de los llamados grados de libertad (degrees of freedom).

El grado de libertad de la distribución corresponder a (n-1) donde n representa el tamaño de la muestra.

Mientras más pequeño es el tamaño de la muestra, hay más área bajo la curva en los extremos de la misma. Por lo tanto, a medida que aumentan los grados de libertad, las distribuciones de t se parecen más a la distribución normal. A partir de 120 la diferencia es tan mínima que se utiliza la normal.

Para las pruebas de hipótesis cuando se desconoce la desviación estándar de la población se utiliza la misma metodología que se ha estudiado anteriormente, sólo cambia la estadística y la tabla. La estadística que se utiliza en estos casos es

$$t_o = \frac{\overline{X} - \mu_X}{\frac{s_X}{\sqrt{n}}}$$

B. Prueba t de una cola

Ejemplo:

En una fábrica la capacidad promedio de cierto tipo de batería es 140 amperes-horas. La distribución de la capacidad es normal. Una agencia de servicio al consumidor quiere chequear que la media prometida por la fábrica es correcta y para hacerlo escogen una muestra aleatoria de 20 baterías. Su interés primordial es asegurarse de que no se estafe a los consumidores. Los resultados son los siguientes:

137.4	140.0	138.8	139.1	144.4
139.2	141.8	137.3	133.5	138.2
141.1	139.7	136.7	136.3	135.6
138.0	140.9	140.6	136.7	134.1

La prueba de hipótesis en esta situación debe ser una prueba de una cola donde la media y la desviación estándar (obtenidas por medio de la calculadora) son s = 2.66; media = 138.47

Prueba de hipótesis

Paso 1 y 2:

H_o: $\mu_x \geq 140$ (no se estafa a los consumidores)

H_1: $\mu_x < 140$ (se estafa a los consumidores)

Paso 3: $\alpha = 0.05$

Paso 4: n = 20 por lo tanto hay n -1 = 19 d.f.

Paso 5: En este caso donde σ_x no es conocida se usa la prueba t

$$t_o = \frac{\overline{X} - \mu_x}{\frac{s_x}{\sqrt{n}}}$$

Paso 6: Para tener 5% del área en la zona de rechazo para una prueba de una cola y 19 df. Esta área corresponde al valor crítico t_c = -1.7291

Paso 7: Si t < -1.7291 entonces se rechaza H. Si t > -1.7291 entonces no se rechaza H_0.

Paso 8:

$$t_o = \frac{\overline{X} - \mu_x}{\frac{s_x}{\sqrt{n}}} = \frac{138.47 - 140}{\frac{2.66}{\sqrt{20}}} = -2.57$$

Paso 9: t_o cayó en la zona de rechazo, por lo tanto se rechaza la hipótesis nula

Paso 10: Hay evidencia para concluir que la capacidad promedio de las baterías es menos de lo que la compañía reclama.

III. Tipos de pruebas estadísticas de hipótesis

Uno de los aspectos más importantes de las estadísticas en la investigación es comprender que para poder utilizar una prueba estadística el experimento tiene que cumplir con los supuestos que exige la prueba.

Las pruebas de hipótesis pueden ser paramétricas, libres de distribución y noparamétricas.

A. Pruebas paramétricas

Requieren

1. variables medidas en la escala de razón o intervalar

2. análisis de un parámetro de la población y otros requisitos que dependen de la prueba en específico.

La robustez en las pruebas paramétricas

Una prueba paramétrica es robusta si a pesar de no cumplir cabalmente con todos los requisitos se puede emplear sin que deforme mucho las conclusiones. Cuando la prueba no es robusta es necesario utilizar otra prueba libre de distribución o no paramétrica.

B. Pruebas libres de distribución

1. La prueba estadística no depende de la forma de la distribución de la población

2. Los datos están en escala nominal u ordinal

C. Pruebas no paramétricas

No tienen que ver con los parámetros de la población

IV. Supuestos para la prueba de z

La prueba de z es paramétrica por lo tanto requiere que:

1. las variables se midan en la escala de razón o la escala intervalar

2. se lleve a cabo el análisis de un parámetro de la población

Además requiere que:

3. las observaciones sean independientes y seleccionadas aleatoriamente

4. la distribución de la población sea normal o que el tamaño de la muestra sea mayor de 30 para poder utilizar el Teorema Central del Límite.

V. Supuestos para la prueba de t

Es paramétrica por lo tanto requiere que:

1. las variables se midan en la escala de razón o la escala intervalar

2. se lleve a cabo el análisis de un parámetro de la población

Además requiere que:

3. las observaciones sean independientes y seleccionadas aleatoriamente

4. la distribución de la población sea normal

La **prueba de t es robusta** si la distribución de la población difiere un poco de la normal, pero el tamaño de la muestra es suficientemente grande (más de 30). Sin embargo, si el tamaño de la muestra es menor de 30 y la población no es normal, entonces es preferible olvidarse de la prueba y usar otra libre de distribución.

VI. El intervalo de confianza para la prueba de t

Con la prueba de t se utiliza el intervalo de confianza como con la prueba de z, pero la fórmula cambia para los límites del intervalo. Se utiliza la siguiente fórmula:

$$\overline{X} \pm t_{n-1}\left(\frac{s}{\sqrt{n}}\right)$$

Ejemplo:

En una fábrica la capacidad promedio de cierto tipo de batería es 140 amperes-horas. Un técnico de control de control de calidad quiere chequear que esto sea cierto y para hacerlo escoge una muestra aleatoria de 20 baterías. Para hacerlo debe hallar el intervalo de confianza de 95% que le permitiría decir que la fábrica está funcionando adecuadamente.

137.4	140.0	138.8	139.1	144.4
139.2	141.8	137.3	133.5	138.2
141.1	139.7	136.7	136.3	135.6
138.0	140.9	140.6	136.7	134.1

Si se hiciese una prueba de hipótesis en esta situación ésta debería ser una prueba de dos colas donde la media y la desviación estándar (obtenidas por medio de la calculadora) son s = 2.66; media = 138.47

Determinación del intervalo de confianza

Utilizando la siguiente fórmula

$$\overline{X} \pm t_{n-1}\left(\frac{s}{\sqrt{n}}\right)$$

el intervalo es:

$138.47 \pm (2.093)\ (2.66/4.47)$

138.47 ± 1.24

$137.23 < \mu_X < 139.71$

Por lo tanto la media de la población se espera que esté en ese intervalo

Como la media de la hipótesis nula no está en el intervalo de confianza se puede decir con un 95% de confianza que: **Hay evidencia para concluir que la media es diferente de 140 amperes hora.**

VII. Significación estadística y significación práctica

Cuando se rechaza una hipótesis nula, en el lenguaje técnico de las estadísticas se dice que

La diferencia entre el parámetro hipotético y la estadística de la muestra es estadísticamente significativa.

La pregunta que queda por hacer es si el hecho de que la diferencia sea estadísticamente significativa indica que esta diferencia tenga importancia práctica. Cuando la muestra es grande una pequeña diferencia entre estadística y parámetro puede llegar a ser significativa. Esto se debe a que al aumentar el tamaño de la muestra se logra más precisión estadística, pues el intervalo de confianza se hace más estrecho.

Por lo tanto es necesario, en términos del experimento que se lleva a cabo, determinar si esta diferencia es de valor práctico. Puede ser que cueste más parar la fábrica para reparar la máquina que permitir que siga funcionando con un ligero desperfecto. Estas preguntas no las puede responder la estadística, sino la situación en que se da el problema.

Lectura:

Hinkle capt. 8,9 pp.204-213; pp.217-227

Actividades:

Hinkle pp. 213 # 2,3,5,14; pp.227 ej. 6

MENU 6390

EDUC 6390: Estadística aplicada en la educación

Andrés Menéndez Raymat, Ph.D.

Conferencia 15

Pruebas de hipótesis para dos muestras

Bosquejo

I. Introducción

Hasta ahora se ha hecho la prueba de hipótesis para inferir el parámetro de la población partiendo de las estadísticas obtenidas en una muestra. En esta unidad por medio de la prueba de hipótesis se compararán estadísticas de dos muestras para hacer inferencias sobre los parámetros de sus respectivas poblaciones. Primero se trabajará con muestras que provienen de poblaciones independientes, luego con muestras que provienen de poblaciones dependientes.

II. Prueba de t para la diferencia entre dos medias cuando las varianzas de las poblaciones son iguales (Pooled Variance T-test)

A. Si se conocen las desviaciones estándar de las poblaciones

Situación:

Se quiere determinar si las medias de dos poblaciones independientes son diferentes cuando se conocen las desviaciones estándar de las poblaciones y las muestras son grandes.

Población 1:

Media μ_1

Desv. Est. σ_1

Tamaño de la muestra n_1

Población 2:

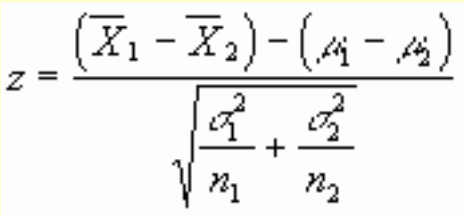
Media μ_2

Desv.Est. σ_2

Tamaño de la muestra n_2

De acuerdo con el Teorema Central del Límite, si la muestra es grande, la estadística que se usa cuando se conoce la varianza de la población tiene una distribución normal.

En el caso de dos muestras la estadística z que se utiliza se computa siguiendo la siguiente fórmula:



$$z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

donde:

= medias de las muestras

μ = medias de las poblaciones

σ^2 = varianzas de las poblaciones

n = tamaños de las muestras

B. Si no se conocen las desviaciones estándar de las poblaciones

El problema, en la mayoría de los casos, es que no se conoce la desviación estándar de las poblaciones. Sólo se conocen las desviaciones estándar de las muestras. Si esto ocurre, es necesario asegurarse de que el estudio cumple con los siguientes supuestos antes de seguir el método que se presentará a continuación:

1. Supuestos

1. Las muestras se seleccionan aleatoriamente.

2. Las muestras son independientes (ie. Las observaciones en una muestra no tienen nada que ver con las observaciones en la otra muestra)

3. Las poblaciones tienen una distribución normal

4 . Las varianzas de las poblaciones son iguales (**homogeneidad de varianzas**)

Si hay el mismo número de observaciones en los dos grupos, la prueba es robusta y por lo tanto no hace falta realizar la prueba de homogeneidad de varianzas.

Tradicionalmente los dos primeros supuestos se logran seleccionando aleatoriamente los sujetos y asignando aleatoriamente la mitad al grupo control y la otra mitad al experimental.

2. Prueba de t de varianzas combinadas (pooled variance t-test)

i. Hipótesis

Es importante señalar que las hipótesis se pueden presentar de dos formas diferentes, como una comparación o como una diferencia comparable a cero.

ii. Dos colas

$H_0: \mu_1 = \mu_2$ ó $\mu_1 - \mu_2 = 0$

$H_1: \mu_1 \neq \mu_2$ ó $\mu_1 - \mu_2 \neq 0$

iii. Una cola

$H_0: \mu_1 \geq \mu_2$ ó $\mu_1 - \mu_2 \geq 0$

$H_1: \mu_1 < \mu_2$ ó $\mu_1 - \mu_2 < 0$

o viceversa

iv. Estadística

$$t = \frac{\left(\overline{X_1} - \overline{X_2}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

donde la varianza combinada es

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1)(n_2 - 1)}$$

esta prueba de t tiene una distribución t con (n_1 + n_2 - 2) grados de libertad.

3. Ejemplo

Compara los promedios de dividendos en la bolsa de valores de NY y la de Londres.

Asume que $\sigma_{NY} = \sigma_L$

NY: $n_1 = 21$; $_1 = 3.27$; $s_1 = 1.30$

Londres: $n_2 = 25$; $_2 = 2.53$; $s_2 = 1.16$

Es neceario chequear si se cumple con los supuestos:

1. Las muestras se seleccionaron aleatoriamente.

2. Las muestras son independientes (ie. Las observaciones en una muestra no tienen nada que ver con las observaciones en la otra muestra)

3. Las poblaciones tienen una distribución normal

4 . Las varianzas de las poblaciones son iguales

Paso 1 y 2:

H_0: $\mu_1 = \mu_2$ ó $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 \neq \mu_2$ ó $\mu_1 - \mu_2 \neq 0$

Paso 3: $\alpha = 0.05$

Paso 4: $n_1 = 21$; $n_2 = 25$

Paso 5: Prueba que se debe usar (varianza combinada) pooled-variance t-test con

$n_1 + n_2 - 2 = 21 + 25 - 2 = 44$ df

Paso 6: Valores críticos para las zonas de rechazo y no rechazo.

Para tener 5% del área en la zona de rechazo hay que buscar los valores críticos para la prueba de dos colas bajo 0.025. Los valores críticos corresponden a -2.0154 y 2.0154

Paso 7: Regla decisional

Si $t_0 < -2.0154$ ó si $t_0 > 2.0154$ se rechaza H_0

Si $-2.0154 < t_0 < 2.0154$ **NO** se rechaza H_0

Paso 8: Computar t_0 después de computar la varianza combinada

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1)(n_2 - 1)}$$

$$S_p^2 = \frac{(20)(1.3)^2 + (24)(1.16)^2}{21 + 25 - 2}$$

$$t_0 = \frac{\left(\overline{X_1} - \overline{X_2}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$t_0 = \frac{(3.27 - 2.53) - (0)}{\sqrt{1.51\left(\frac{1}{21} + \frac{1}{25}\right)}} = \frac{0.74}{0.3637} = 2.03$$

Paso 9: t_0 cayó en la zona de rechazo por lo tanto se rechaza H_0

Paso 10: Con un 5% de significación podemos decir que la evidencia apoya la conclusión de que hay diferencias entre las medias de los dos grupos.

4. Intervalo de confianza

De igual forma que se hizo con una sola muestra se puede construir un intervalo de confianza alrededor de la estadística y determinar si la diferencia entre los parámetros se halla dentro del intervalo de confianza. La fórmula para el intervalo de confianza es:

$$CI_{95} = \left(\overline{X}_1 - \overline{X}_2\right) \pm t_c\left(s_{\overline{X}_1 - \overline{X}_2}\right)$$

donde la diferencia entre las medias es = 3.27 - 2.53 = 0.74;

t_c = 2.0154 (valor crítico de t)

$$s_{\overline{X}_1 - \overline{X}_2} = \sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{1.510\left(\frac{1}{21} + \frac{1}{25}\right)} = 0.3637$$

Por lo tanto $CI_{95} = 0.74 \pm (2.0154)(0.3637) = 0.74 \pm 0.733$

(0.007, 1.473)

Como el punto cero que corresponde a la diferencia entre las medias en la hipótesis nula no se encuentra en el intervalo, se rechaza la hipótesis nula en favor de la alterna y se concluye: **Con un 5% de significación podemos decir que la evidencia apoya la conclusión de que hay diferencias entre las medias de los dos grupos**. Otra forma de decirlo, (probability statement)

La probabilidad de que la diferencia observada entre las medias de las muestras haya ocurrido al azar, si en efecto la hipótesis nula fuese cierta es menos de 0.05

III. Prueba de t para la diferencia entre dos medias cuando las varianzas de las poblaciones no son iguales (Separate Variance t-test)

Cuando no se puede asumir que las dos poblaciones de las que se tomaron las muestras tienen varianzas iguales o homogéneas entonces se tiene que utilizar otro método que fue desarrollado por Satterthwaite para bregar con diferentes varianzas. El método es idéntico al anterior con la diferencia de que la fórmula incorpora las dos varianzas y es necesario utilizar una fórmula adicional para determinar los grados de libertad.

$$t_0 = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$\Delta = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

donde Δ = grados de libertad y Δ siempre se aproxima a la parte integral, nunca se redondea.

Se va a utilizar el mismo ejemplo anterior pero no se va a asumir que las varianzas de las dos poblaciones son iguales.

1. Ejemplo

Compara los promedios de dividendos en la bolsa de valores de NY y la de Londres.

Asume que $\sigma_{NY} \neq \sigma_{L}$

NY: $n_1 = 21$; $_1 = 3.27$; $s_1 = 1.30$

Londres: $n_2 = 25$; $_2 = 2.53$; $s_2 = 1.16$

Pasos 1 y 2:

$H_0: \mu_I = \mu_2$ ó $\mu_I - \mu_2 = 0$

$H_1: \mu_I \neq \mu_2$ ó $\mu_I - \mu_2 \neq 0$

Paso 3: $\alpha = 0.05$

Paso 4: $n_1 = 21$; $n_2 = 25$

Paso 5: Prueba que se debe usar

En este caso se usa la prueba de t para varianzas diferentes (separate-variance t-test) con Δ grados de libertad

$$\Delta = \frac{\left(\frac{1.698}{21} + \frac{1.353}{25}\right)^2}{\frac{\left(\frac{1.698}{21}\right)^2}{20} + \frac{\left(\frac{1.353}{n_2}\right)^2}{24}}$$

$\Delta = 40.58$

Por lo tanto, los grados de libertad son 40

Paso 6: Valores críticos para las zonas de rechazo y no rechazo

Para tener 5% del área en la zona de rechazo hay que buscar los valores críticos para la prueba de dos colas bajo 0.025 con 40 df. Los valores críticos corresponden a -2.0211 y 2.0211

Paso 7: Regla decisional

Si $t_0 < -2.0211$ ó si $t_0 > 2.0211$ se rechaza H_0

Si $-2.0211 < t_0 < 2.0211$ **NO** se rechaza H_0

Paso 8: Computar t_0

$$t_0 = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t_0 = \frac{(3.27 - 2.53) - (0)}{\sqrt{\frac{1.698}{21} + \frac{1.353}{25}}}$$

$t_0 = 2.01$

Paso 9:

t_0 no cayó en la zona de rechazo por lo tanto **NO** se rechaza H_0

Paso 10:

Con un 5% de significación podemos decir que no existe evidencia para concluir que haya diferencias entre las medias de los dos grupos.

2. Intervalo de confianza

De igual forma que se hizo cuando las varianzas eran iguales se puede construir un intervalo de confianza alrededor de la estadística y determinar si la diferencia entre los parámetros se halla dentro del intervalo de confianza. La fórmula para el intervalo de confianza es:

$$CI_{95} = \left(\overline{X}_1 - \overline{X}_2\right) \pm t_c \left(s_{\overline{X}_1 - \overline{X}_2}\right)$$

donde la diferencia entre las medias es = 3.27 - 2.53 = 0.74; $t_c = 2.0211$ (valor crítico de t con 40 grados)

$$s_{\bar{X}_1-\bar{X}_2} = \sqrt{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)} = \sqrt{\left(\frac{1.698}{21}+\frac{1.353}{25}\right)} = 0.3674$$

Por lo tanto

$CI_{95} = 0.74 \pm (2.0211)(0.3674)$

$= 0.74 \pm 0.7426$

(-0.0026, 1.4826)

Como el punto cero que corresponde a la diferencia entre las medias en la hipótesis nula se encuentra en el intervalo, no se rechaza la hipótesis nula en favor de la alterna y se concluye:

Con un 5% de significación podemos decir que la evidencia apoya la conclusión de que no hay diferencias entre las medias de los dos grupos. Otra forma de decirlo, (probability statement)

La probabilidad de que la diferencia observada entre las medias de las muestras haya ocurrido al azar, si en efecto la hipótesis nula fuese cierta es mayor de 0.05

IV. Prueba F para la diferencia entre dos varianzas

Se han obtenido resultados contradictorios en las dos pruebas de hipótesis. Por lo tanto es imprescindible determinar cuál es la prueba apropiada. Nótese que es más fácil rechazar cuando las varianzas son iguales que cuando no lo son. Para determinar si las varianzas de la población son iguales es necesario hacer la prueba de homogeneidad de varianzas que utiliza otro tipo de estadística, la **prueba F**.

A. La estadística

Esta prueba se basa en la razón

$$\frac{S_1^2}{S_2^2}$$

que sigue una distribución que no se ha estudiado hasta el momento, la **distribución F** que tiene una tabla diferente a las que hasta ahora se han consultado y que para colmo no es simétrica. Esta distribución depende de dos conjuntos de grados de libertad, uno para el numerador y otro para el denominador. La estadística es:

$$F = \frac{S_1^2}{S_2^2}$$

donde n_1 - 1 son los grados de libertad del numerador y n_2 - 1 son los grados de libertad del denominador.

B. Ejemplo:

Compara las varianzas de los dividendos de las bolsas de NY y Londres.

NY: $n_1 = 21$; $\bar{x}_1 = 3.27$; $s_1 = 1.30$

Londres: $n_2 = 25$; $\bar{x}_2 = 2.53$; $s_2 = 1.16$

Pasos 1 y 2:

H_0: $\sigma_1^2 = \sigma_2^2$

H_1: $\sigma_1^2 \neq \sigma_2^2$

Paso 3: $\alpha = 0.05$

Paso 4: $n_1 = 21$; $n_2 = 25$

Paso 5: Seleccionar la prueba adecuada

En este caso estamos usando la distribución F con 20 y 24 grados de libertad en el numerador y el denominador.

Paso 6: Determinar los valores críticos para separar la región de rechazo de la

de no rechazo.

Para tener 5% en el área de rechazo es necesario hallar los valores críticos para una prueba de dos colas bajo 0.025 para 20 y 24 grados de libertad. El valor crítico corresponde a 2.33 para la cola superior. Pero para hallar el valor crítico de la cola inferior es necesario hacer un trabajo adicional. Hay que utilizar el recíproco de los valores críticos de la cola superior con los grados de libertad invertidos.

$$F_{I(n_1-1),(n_2-1)} = \frac{1}{F_{u(n_2-1),(n_1-1)}}$$

$$F_I = \frac{1}{2.41} = 0.415$$

Por lo tanto el valor crítico en la cola de la izquierda es 0.415

Este trabajo con la cola inferior se puede eliminar si tenemos buen cuidado de poner siempre en el paso 8 la desviación estándar mayor en el numerador.

Paso 7: La regla decisional

Si $F_0 < 0.415$ ó si $F_0 > 2.33$ rechazar H_0

Si $0.415 < F_0 < 2.33$ no rechazar H_0

Paso 8: Computar F_c

$$F = \frac{S_1^2}{S_2^2}$$

F = 1.698/1.353 = 1.25

Paso 9: Determinar si el valor observado de F cayó o no en la zona de rechazo y tomar la decisión estadística. No cayó en la zona de rechazo, por lo tanto no se rechaza la hipótesis nula.

Paso 10: La decisión

No hay evidencia de una diferencia entre las varianzas de los dos grupos, por lo tanto se puede utilizar la prueba de t para dos medias cuando las varianzas son homogéneas.

C. Supuestos

Esta prueba de homogeneidad de varianzas asume que las dos poblaciones tienen distribuciones normales. La prueba de F no es robusta bajo este supuesto, especialmente si las muestras tienen tamaños diferentes.

V. Prueba de t para dos muestras dependientes

Hasta ahora hemos estado trabajando con muestras tomadas de dos poblaciones independientes, que no tienen ninguna relación una con la otra. Ahora vamos a concentrarnos en situaciones en que el primer grupo está relacionado con el segundo. Esto ocurre cuando los individuos de las muestras han sido pareados o el mismo individuo ha sido examinado en dos ocasiones diferentes.

Ejemplos:

Puede ser el mismo individuo que ha tomado dos pruebas diferentes (pre y post), hermanos, hombre y mujer del mismo país, maridos y mujeres, el mismo objeto vendido bajo dos condiciones diferentes, etc. Es necesario notar que obligatoriamente se tiene que tener el mismo número de observaciones en cada muestra. Cuando es el mismo individuo que se mide en dos ocasiones diferentes se llama una **prueba de medidas repetidas**. En estos casos lo importante, más que las medidas en sí, es la diferencia entre las medidas. La diferencia entre parear y repetir medidas puede verse en el siguiente ejemplo

Parear: El pareo ocurre cuando se toman pares de cajas de cereal que son idénticas y se llena cada una en una máquina diferente.

Medidas repetidas: La medidas repetidas ocurren cuando se toma una caja de cereal, se llena en una máquina, se vacía y se vuelve a llenar en la otra máquina.

El objetivo de la prueba de t para muestras dependientes es estudiar las diferencias con más precisión, puesto que la variabilidad que se debe a la diferencia entre los sujetos se reduce al ser el mismo sujeto o sujetos

semejantes los que se miden.

A. La prueba de z

En todos los problemas de muestras dependientes lo primero que hay que hacer es determinar la diferencia entre todos los pares y hallar el promedio de las diferencias que se expresan con la letra mayúscula "D" y dos suscritos, el primero indica la muestra y el segundo la pareja dentro de las muestras. Asi $D_{\mathbf{5}} = X_{15} - X_{25}$ se refiere a la diferencia entre la primera y la segunda medida en el **quinto** sujeto o pareja de sujetos.

$D_1 = X_{11} - X_{21}$; $D_2 = X_{12} - X_{22}$; $D_3 = X_{13} - X_{23}$; $D_4 = X_{14} - X_{24}$

$D_i = X_{1i} - X_{2i}$

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}$$

se forma añadiendo todas las diferencias y dividiendo entre el número de diferencias. Las diferencias se forman seleccionando un individuo de una muestra y su pareja de la otra. Si se conoce la desviación estándar de la población entonces se puede utilizar la estadística z

$$z = \frac{\overline{D} - \mu_D}{\frac{\sigma_D}{\sqrt{n}}}$$

pero como la desviación estándar de la población casi nunca se conoce por lo general se utiliza

B. La prueba t para la diferencia entre las medias de muestras dependientes (t-test for the Mean Difference)

1. Supuestos

1. La distribución de la población de diferencias es normal

2. La selección para la pareja es aleatoria

Esta **prueba es robusta** con respecto a la normalidad de la población si el tamaño de la muestra es grande y la distribución es simétrica. H_0 significa que no hay diferencia en la medias de las dos poblaciones

La estadística utilizada es

$$t_0 = \frac{\overline{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}$$

donde = promedio de las diferencias de los pares en las muestras y μ_D = promedio de las diferencias de los pares en las poblaciones; s_D = desviación estándar de las diferencias

2. Ejemplo

En una empacadora de cereal el jefe de producción tiene dos máquinas para llenar cajas y quiere compararlas para determinar cuál desperdicia más cereal y eventualmente tomar una decisión sobre la máquina que va a dejar funcionando en la empacadora. En este ejemplo el jefe de producción tiene que probar las 10 cajas diferentes correspondientes a los diferentes tipos de cereal que se empacan en la fábrica. Toma dos cajas de cada tipo y obtiene los siguientes datos:

Cantidad (en gramos) de cereal derramado en una muestra de 10 tipos de cajas empacadas por dos máquinas diferentes

	Tipo de máquina		
Tipo de cereal	Nueva	Vieja	Diferencias
1	12.73	13.89	-1.16
2	9.75	10.32	-0.57
3	13.78	17.01	-3.23
4	8.37	10.43	-2.06
5	11.71	11.39	+0.32
6	15.47	17.99	-2.52
7	14.56	16.02	-1.46
8	11.74	11.90	-0.16

9	9.76	13.11	-3.35
10	12.47	13.88	-1.41

Hay que recordar en el momento de la decisión final cómo fue que se hizo la resta. (A lo que derramaba la nueva se le restó lo que derramaba la vieja). Por lo tanto si la **diferencia es estadísticamente significativa, esto querría decir que la vieja es mejor**.

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = -1.56$$

$$\sum_{i=1}^{10} D_i^2 = 38.1676$$

Pasos 1 y 2:

H_O: $\mu_D \geq 0$ ó $\mu_{nueva} \geq \mu_{vieja}$

Esta sería la hipótesis en el caso de que el jefe quiera acumular evidencia para señalar que la nueva desperdicia menos cereal que la vieja y así justificar la compra

H_1: $\mu_D < 0$; $\mu_{nueva} < \mu_{vieja}$

Paso 3: $\alpha = 0.05$

Paso 4: n = 10

Paso 5: En este caso se utiliza la distribución de t con 9 df

Paso 6: valores críticos

Para tener un 5% del área en la zona de rechazo tenemos que hallar los valores críticos para la prueba de una cola con 9 df. El valor crítico corresponde a -1.833

Paso 7: Regla decisional

Si $t_0 < -1.8331$ se rechaza H_0

Si $t_0 \geq -1.8331$ no se rechaza H_0

Paso 8: Cómputo de t_0

$$t_0 = \frac{\overline{D} - \mu_D}{\frac{s_D}{\sqrt{n}}} = \frac{-1.56 - 0}{\frac{1.24}{\sqrt{10}}} = -3.978$$

Paso 9: El valor observado de t cayó en la zona de rechazo por lo tanto se rechaza la hipótesis nula.

Paso 10: Con un 5% de significación hay evidencia de que la máquina nueva derrama menos cereal que la vieja. Aquí no se utiliza el intervalo de confianza pues la prueba es de una sola cola.

Lectura:

Hinkle captítulo 11 pp.251-285

Actividades:

1. Utilizando los datos de los ejercicios Hinkle pp. 285-289 asume que el error estándar en los siguentes ejercicios es:
2a error estándar = 0.39
3a error estándar = 1.22
4a error estándar = 2.17
5a error estándar = 1.11
a. Haz la prueba de hipótesis utilizando Excel para obtener el valor de p en los ejercicios 3a,4a y 5a
b. En el ejercicio 5a determina si las varianzas de la población son homogéneas o no utilizando la prueba de F antes de hacer la prueba de t.

MENU 6390

EDUC 6390: Estadística aplicada en la educación
Andrés Menéndez Raymat, Ph.D.
Conferencia 16
Prueba no paramétrica de hipótesis: Ji-cuadrada

I. Pruebas paramétricas y no paramétricas

En la estadística que se ha utilizado hasta este momento del curso, por lo general, se ha estudiado **una sola** característica o **variable** de cada sujeto. Solamente en las tablas de contingencia para agrupar datos y más tarde en las secciones sobre correlación y regresión se ha estudiado la relación entre dos variables. La estadística inferencial analizada también se ha limitado al estudio de una sola variable. Las pruebas de hipótesis (tanto para una como para dos muestras) han sido todas paramétricas y de una forma u otra han asumido como modelo estadístico, la distribución normal.

En la conferencia 14 se había hablado de las características de las diversas pruebas de hipótesis que se presentan a continuación:

A. Pruebas paramétricas

Propiedades

1. Requieren que las variables se midan con la escala intervalar o de razón

2. Se relacionan con el estudio de un parámetro de la población (media, varianza, etc.)

B. Pruebas libres de distribución

Propiedades

1. Son pruebas de hipótesis que no dependen de la forma de la distribución de la población.

2. Los datos no asumen una escala intervalar o de razón. Es suficiente que la escala sea nominal u ordinal para poder utilizar estas pruebas.

C. Pruebas no paramétricas

Propiedades

1. No se relacionan con el estudio de un parámetro de la población.

Nota

Por lo general, cuando se hace referencia a pruebas no paramétricas se puede estar hablando tanto de las no paramétricas como de las libres de distribución (distribution free tests)

Las dos indicaciones más importantes que se deben tener en cuenta para utilizar una prueba no paramétrica son que:

1. la distribución de la población no sea normal

2. la escala de medición de la variable en cuestión sea categórica.

Los métodos no paramétricos son menos poderosos que los paramétricos. Esto quiere decir que es más difícil rechazar la hipótesis nula con las pruebas no paramétricas.

Por esa razón los estadísticos, por lo general, recurren a los métodos no paramétricos sólo cuando los datos no cumplen con los supuestos paramétricos. Sin embargo, hay problemas de investigación en los que las variables categóricas son las indicadas y por lo tanto sólo un método no paramétrico es el indicado. Los pasos en las pruebas de hipótesis no paramétricas son los mismos de las paramétricas. Los cambios, por lo general, se limitan a cambios en la fórmula para obtener el valor observado y en la tabla que se utiliza. Pero se habla de hipótesis nula, nivel de significación, error tipo I y tipo II, etc.

II. Las pruebas de hipótesis Ji-cuadrada

La prueba de hipótesis no paramétrica más común utiliza la estadística **JI CUADRADA (χ^2)** que tiene una distribución muy parecida a la distribución **JI CUADRADA (χ^2)**. Fue desarrollada por Karl Pearson en 1900 y se utiliza en varias situaciones diferentes.

En este capítulo se utilizará esta prueba de hipótesis para una o dos variables.

A. Prueba Ji-cuadrada de ajuste

Cuando el estudio tiene que ver con una sola variable el nombre que recibe la prueba de hipótesis es **Prueba Ji-cuadrada de ajuste** (Goodness of fit test) o JI CUADRADA (χ^2) de una vía**.** En este caso el interés del investigador se concentra en comparar las frecuencias de los niveles de la variable en la muestra con las frecuencias de los niveles de la variable en la población.

B. Prueba Ji-cuadrada de independencia

Cuando el estudio tiene que ver con dos variables el nombre que recibe la prueba es **Prueba Ji-cuadrada de independencia** (Testing independence). En este caso el interés del investigador se dirige a analizar si las frecuencias en los diferentes niveles de las dos variables indican que existe o no una relación entre las dos variables. Algunos autores le llaman a esta **Prueba Ji-cuadrada de homogeneidad**

III. La distribución JI CUADRADA (χ^2)

Al igual que las distribuciones z y t, **JI CUADRADA (χ^2)** es una familia de distribuciones cuya forma depende de los grados de libertad de la distribución.

Propiedades

Las distribuciones ji cuadrada tienen un sesgo positivo por lo que la zona de rechazo para las pruebas de hipótesis son siempre de una cola a pesar de que siempre la hipótesis es no direccional. Los valores son siempre positivos y el valor mínimo posible es cero. Esto ocurre puesto que, como se verá más adelante, en el cómputo de la fórmula hay un paso en que se cuadran los valores obtenidos haciéndose todos positivos. La zona donde se encuentra el cero (izquierda de la distribución) corresponde siempre a la zona de no rechazo. No es necesario analizar el valor crítico negativo. Si el valor observado es mayor que el valor crítico se rechaza la hipótesis nula. A medida que los grados de libertad aumentan, la distribución se hace más simétrica y con más de 30 grados de libertad comienza a parecerse a la distribución normal. (Hinkle, p.577)

IV. La prueba de ajuste (Goodness of fit test)

A. Supuestos

Se utiliza la prueba de ajuste cuando el estudio cumple con los siguientes supuestos:

1. Los sujetos están categorizados con respecto a una sola variable que puede tener dos o más categorías.

2. Cada sujeto aparece una sola vez y en una sola categoría.

3. Cada asignación a una categoría es independiente de cualquier otra asignación. (El que un sujeto se asigne a una categoría no tiene nada que ver con cómo se asigna otro sujeto)

4. Los cómputos se hacen con todos los sujetos del estudio.

5. La frecuencia esperada ($\mathbf{f_e}$) en cada celda es igual o mayor de 5.

B. Lógica del análisis

Ejemplo

La pregunta de investigación es si un dado está cargado o no. Para determinarlo se lanza el dado en múltiples ocasiones y se anota la frecuencia en que cada uno de los seis valores del dado ocurrió.

Si el dado no está cargado se espera que las frecuencias de cada uno de los seis valores sean iguales. Esta frecuencia esperada se indica como f_e.

De 60 tiradas del dado en 10 de ellas se espera el valor 1, en 10 el valor 2, etc.

Esta situación raramente ocurre, pues está presente el aspecto aleatorio del experimento. Sin embargo, es posible decir que mientras más cercanas sean las frecuencias observadas (f_o) a las frecuencias esperadas, más seguro se puede estar de que el dado no está cargado. La fórmula que indica cuanto se parecen las frecuencias esperadas a las observadas es la de la estadística χ^2

C. Fórmulas

La fórmula para computar χ^2 **observada** es

Ejemplo

En la tabla que aparece a continuación se encuentran, en la primera columna, las frecuencias observadas de un dado que se lanza 120 veces. En la segunda

columna se incluyen las frecuencias esperadas y en la tercera la estadística ji-cuadrada.

Valor	f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2/f_e$
1	15	20	-5	25/20
2	27	20	7	49/20
3	20	20	0	0/20
4	30	20	10	100/20
5	10	20	-10	100/20
6	18	20	-2	4/20
Total	120			278/20

χ^2 observada = 278/20 = 13.9

Para determinar si esta estadística es demasiado grande se utiliza la tabla de la distribución. Si cae en la zona de rechazo, la conclusión es que no ha ocurrido al azar y que el dado está cargado.

D. La prueba de hipótesis usando χ^2

La prueba de hipótesis sigue los pasos conocidos:

1. Determinar la hipótesis del investigador y la nula y seleccionar el nivel de significación α

2. Seleccionar la prueba que se va a utilizar y los grados de libertad que corresponden a **c-1** donde **c** son los niveles de la variable o las categorías.

3. Determinar el valor crítico, las zonas de rechazo y no rechazo

4. Computar el valor observado

5. Tomar la decisión e interpretar los resultados.

H_o indica que no hay diferencia entre las frecuencias observadas de los valores en la muestra y las frecuencias esperadas.

No hay una forma convencional de escribir la hipótesis nula. Se usan palabras

H_o: Las frecuencias observadas son iguales a las observadas.

La implicación de H_o es que si las frecuencias observadas no son iguales esto se debe a un error de muestreo.

H_1 es la hipótesis alterna y siempre indica que las frecuencias observadas son diferentes de las esperadas.

H_1: Las frecuencias observadas **NO** son iguales a las observadas.

Ejemplo

Se lanza una moneda 200 veces y salen 92 caras y 108 cruces. ¿Está cargada la moneda?

Esta es una prueba de ajuste, pues el investigador está interesado en determinar si las frecuencias de caras y cruces obtenidas en la muestra corresponden a una población donde la probabilidad de obtener cara o cruz es 0.5 en cada tirada. Si la probabilidad no fuese 0.5, entonces se podría decir que la moneda está cargada.

¿Cumple el ejemplo con los supuestos de la prueba de ajuste?

1. Los sujetos están categorizados con respecto a una sola variable que puede tener dos o más categorías.

2. Cada sujeto aparece una sola vez y en una sola categoría.

3. Cada asignación a una categoría es independiente de cualquier otra asignación. (El que un sujeto se asigne a una categoría no tiene nada que ver con cómo se asigna otro sujeto)

4. Los cómputos se hacen con todos los sujetos del estudio.

5. La frecuencia esperada ($\mathbf{f_e}$) en cada celda es igual o mayor de 5.

valor	f_o	f_e
cara	92	100
cruz	108	100

Pasos 1 y 2: Hipótesis

La redacción de las hipótesis depende de la interpretación del problema por el investigador. Hay variedad, pero, por lo general, habrá una referencia a las frecuencias de los niveles de las variables. La hipótesis alterna es la hipótesis del investigador. Las hipótesis siempre se refieren a la población, aunque en este caso no se hable de parámetros.

La prueba es no direccional, pues la hipótesis alterna indica que las frecuencias obtenidas difieren de las esperadas.

Dos colas:

H_o: Las frecuencias (de cara y cruz) en la población son iguales. El dado no está cargado.

H_1: Las frecuencias(de cara y cruz) en la población no son iguales. El dado está cargado.

Paso 3: Nivel de significación α; $\alpha = 0.05$

Paso 4: Selección de la prueba que se va a utilizar

Como se cumple con los supuestos de la prueba de ajuste se debe utilizar la prueba χ^2 de una vía con c -1 grados de libertad donde **c** es el número de categorías. En este caso 2 -1 = 1 df

Paso 5: Valores críticos para las zonas de rechazo y de no rechazo.

En el caso de $\alpha = 0.05$ y 1 df, el valor crítico es 3.841 (Hinkle, p.638, Tabla C.4).

Paso 6: Regla decisional

Por lo tanto si $\chi^2_o \geq 3.84$, se rechaza H_o

si $\chi^2_o < 3.84$, **NO** se rechaza H_o

Paso 7: Cómputo de χ^2_o

$f_e = N (.5) = 200(0.5) = 100$

Paso 8: Determinar si la estadística cayó en la zona de rechazo o en la de no rechazo y tomar la decisión estadística.

χ^2_o cayó en la zona de NO rechazo, por lo tanto **NO se rechaza** la hipótesis nula.

Paso 9: Expresar la decisión en términos del problema

Con un nivel de significación de 0.05 podemos decir que **NO hay suficiente evidencia para concluir que** el dado estuviera cargado. La probabilidad de que las frecuencias observadas hayan ocurrido al azar, si en efecto la hipótesis nula fuera cierta es mayor de 0.05.

Se pueden hacer pruebas de hipótesis utilizando χ^2 cuando hay más de dos categorías o niveles para la variable.

Ejemplo

En una muestra de 864 sujetos las frecuencias obtenidas para cada ocupación fueron:

ocupación	f_o
agricultura	145
obreros	310
empleados del gobierno	305
profesionales	78
ejecutivos	26
Total	864

Un sociólogo opina que la distribución de los sujetos en varias ocupaciones en un área específica del país es la siguiente:

ocupación	f_e	f_e
agricultura	20%	172.8
obreros	30%	259.2

empleados del gobierno	30%	259.2
profesionales	15%	129.6
ejecutivos	5%	43.2

Determina si la hipótesis del sociólogo es correcta.

Paso 1 y 2: Hipótesis

Dos colas:

H_0: Las frecuencias en la población corresponden a la hipótesis del sociólogo.

H_1: Las frecuencias en la población NO corresponden a la hipótesis del sociólogo.

Paso 3: Nivel de significación α; $\alpha = 0.05$

Paso 4: Selección de la prueba que se va a utilizar

Como la variable es nominal se debe utilizar la prueba χ^2 de una vía con c -1 grados de libertad donde **c** es el número de categorías. En este caso 5 -1 = 4 df

Paso 5: Valores críticos para las zonas de rechazo y de no rechazo.

El problema con la distribución de χ^2 es que la distribución no es simétrica y cambia según los grados de libertad. En el caso de 4 grados de libertad (Hinkle, p.577).

En el caso de $\alpha = 0.05$ y 4 df, el valor crítico es 9.488 (Hinkle, p.638, Tabla C.4).

Paso 6: Regla decisional

Por lo tanto si $\chi^2_o \geq 9.488$, se rechaza H_o

si $\chi^2_o < 9.488$, **NO** se rechaza H_o

Paso 7: Cómputo de χ^2_o

f_e = 864 (.20) = 172.8 (agricultura)

f_e = 864 (.30) = 259.2 (empleados del gobierno)

f_e = 864 (.15) = 129.6 (profesionales)

f_e = 864 (.05) = 43.2 (ejecutivos)

Paso 8: Determinar si la estadística cayó en la zona de rechazo o en la de no rechazo y tomar la decisión estadística.

χ^2_o cayó en la zona de rechazo, por lo tanto **se rechaza** la hipótesis nula.

Paso 9: Expresar la decisión en términos del problema

Con un nivel de significación de 0.05 podemos decir que NO hay suficiente evidencia para concluir que la hipótesis del sociólogo sea correcta. La probabilidad de que las frecuencias observadas hayan ocurrido al azar, si en efecto la hipótesis nula fuera cierta es menor de 0.05.

V. La prueba de independencia (Testing independence)

A. Supuestos

Se utiliza la prueba de independencia cuando el estudio cumple con los siguientes supuestos:

1. Los sujetos están categorizados con respecto a dos variables que pueden tener dos o más niveles.

2. Cada sujeto aparece una sola vez y en un solo nivel de cada variable.

3. Cada asignación a un nivel es independiente de cualquier otra asignación. (El que un sujeto se asigne a un nivel no tiene nada que ver con cómo se asigna otro sujeto)

4. Los cómputos se hacen con todos los sujetos del estudio.

5. La frecuencia esperada ($\mathbf{f_e}$) en cada celda es igual o mayor de 5.

B. Lógica del análisis

Ejemplo

La pregunta de investigación es si el aprovechamiento en matemática está relacionado con el género. Una variable es género (hombres y mujeres) y la otra aprovechamiento (bajo nivel, nivel y sobre nivel). La tabla de contingencia presenta las frecuencias observadas.

	bajo nivel	nivel	sobre nivel	Total
Hombres	46	10	65	121
Mujeres	55	4	38	97
Total	101	14	103	218

Si el aprovechamiento en matemática no está relacionado con el género se espera el mismo porcentaje de hombres que de mujeres en cada nivel de la variable aprovechamiento. La tabla de frecuencias esperadas es

	bajo nivel	nivel	sobre nivel	Total
Hombres	(101)(121)/218	(14)(121)/218	(103)(121)/218	121
Mujeres	(101)(97)/218	(14)(97)/218	(103)(97)/218	97
Total	101	14	103	218

	bajo nivel	nivel	sobre nivel	Total
Hombres	56.06	7.77	57.17	121
Mujeres	44.94	6.23	45.83	97
Total	101	14	103	218

(Es importante observar que los totales no cambian). La frecuencia esperada se encuentra en la segunda tabla.

Es posible decir que mientras más cercanas sean las frecuencias observadas (f_o) a las frecuencias esperadas, más seguro se puede estar de que el género y el aprovechamiento son independientes uno de otro. En otras palabras, es posible esperar los mismos porcentajes de hombres que de mujeres en cada

nivel de la variable aprovechamiento. La fórmula que indica cuanto se parecen las frecuencias esperadas a las observadas es la estadística χ^2

C. Fórmulas

La fórmula para computar χ^2 **observada** es

f_o	f_e	f_o-f_e	$(f_o-f_e)^2$	$(f_o-f_e)^2/f_e$	
46	56.06	-10.06	101.2036	1.805273	
55	44.94	10.06	101.2036	2.251972	
10	7.77	2.23	4.9729	0.640013	
4	6.23	-2.23	4.9729	0.798218	
65	57.17	7.83	61.3089	1.072396	
38	45.83	-7.83	61.3089	1.337746	
				7.905618	Ji-cuadrada

χ^2 observada = 7.91

Para determinar si esta estadística es demasiado grande se utiliza la tabla de la distribución. Si cae en la zona de rechazo, la conclusión es que las dos variables están relacionadas

D. La prueba de hipótesis usando χ^2

La prueba de hipótesis sigue los pasos conocidos:

1. Determinar la hipótesis del investigador y la nula y seleccionar el nivel de significación α

2. Seleccionar la prueba que se va a utilizar y los grados de libertad que corresponden a **c-1** donde **c** son los niveles de la variable o las categorías.

3. Determinar el valor crítico, las zonas de rechazo y no rechazo

4. Computar el valor observado

5. Tomar la decisión e interpretar los resultados.

H_o indica que no hay diferencia entre las frecuencias observadas de las celdas correspondientes de hombres y mujeres en la muestra y las frecuencias

esperadas, por lo tanto las variables género y aprovechamiento no están relacionadas. No hay una forma convencional de escribir la hipótesis nula. Se usan palabras

H_0: Las variables género y aprovechamiento están relacionadas en la población.

La implicación de H_0 es que si las frecuencias observadas no son iguales esto se debe a un error de muestreo. H_1 es la hipótesis alterna y siempre indica que las frecuencias observadas son diferentes de las esperadas.

H_1: Las variables género y aprovechamiento están relacionadas en la población.

Ejemplo

La pregunta de investigación es si el aprovechamiento en matemática está relacionado con el género. Una variable es género (hombres y mujeres) y la otra aprovechamiento (bajo nivel, nivel y sobre nivel). La tabla presenta las frecuencias observadas.

	bajo nivel	nivel	sobre nivel	Total
Hombres	46	10	65	121
Mujeres	55	4	38	97
Total	101	14	103	218

Esta es una prueba de independencia, pues el investigador está interesado en determinar si las frecuencias obtenidas en la muestra son las esperadas.

El estudio cumple con los supuestos de la prueba de independencia puesto que:

1. Los sujetos están categorizados con respecto a dos variables que pueden tener dos o más niveles.

2. Cada sujeto aparece una sola vez y en un solo nivel de cada variable.

3. Cada asignación a un nivel es independiente de cualquier otra asignación. (El que un sujeto se asigne a un nivel no tiene nada que ver con cómo se asigna otro sujeto)

4. Los cómputos se hacen con todos los sujetos del estudio.

5. La frecuencia esperada ($\mathbf{f_e}$) en cada celda es igual o mayor de 5.

Algunos autores aceptan la prueba si no hay más de un 20% de las celdas con f_e menor de 5.

Paso 1 y 2: Hipótesis

La redacción de las hipótesis depende de la interpretación del problema por el investigador. Hay variedad, pero, por lo general, habrá una referencia a la relación entre las dos variables.

La hipótesis alterna es la hipótesis del investigador. Las hipótesis siempre se refieren a la población, aunque en este caso no se hable de parámetros.

La prueba es no direccional, pues la hipótesis alterna indica que las frecuencias obtenidas difieren de las esperadas.

Dos colas:

H_0: No hay asociación entre las dos variables en la población.

H_1: Las dos variables están relacionadas en la población.

Paso 3: Nivel de significación α; $\alpha = 0.05$

Paso 4: Selección de la prueba que se va a utilizar

Como se cumple con los supuestos de la prueba de independencia se debe utilizar la prueba χ^2 de independencia con (r-1)(c-1) grados de libertad donde c es el número de columnas y r el número de filas. En este caso (1)(2) = 2 df

Paso 5: Valores críticos para las zonas de rechazo y de no rechazo.

En el caso de $\alpha = 0.05$ y 2 df, el valor crítico es 5.991 (Hinkle, p.638, Tabla C.4).

Paso 6: Regla decisional

Por lo tanto si $\chi^2_o \geq 5.991$, se rechaza H_0

si χ^2_o **<** 5.991, **NO** se rechaza H_o

Paso 7: Cómputo de χ^2_o

χ^2_o **= 7.91**

Paso 8: Determinar si la estadística cayó en la zona de rechazo o en la de no rechazo y tomar la decisión estadística.

χ^2_o cayó en la zona de rechazo, por lo tanto **se rechaza** la hipótesis nula.

Paso 9: Expresar la decisión en términos del problema

Con un nivel de significación de 0.05 podemos decir que **hay suficiente evidencia para concluir que** el aprovechamiento en matemáticas está relacionado con el género. La probabilidad de que las frecuencias observadas hayan ocurrido al azar, si en efecto la hipótesis nula fuera cierta es menor de 0.05.

Lecturas

Hinkle capt. 21 pp.574-581

Actividades

Hinkle pp. 596-97 ej. 2,3,8

Bosquejo 16
MENU 6390