Rosario, Santa Fe, Argentina
autoras: GABRIELA HUTT, MARIA BELEN MARMIROLI - gabyta_rc@hotmail.com

Estadistica general

Bibliografia: cuadernillos de Metodos Estadisticos de la UNR

~~UNIDAD 1: Papel de la estadística en la toma de decisiones~~

(cuadernillo 1 y 2)

~~La estadística y el método científico~~: La estadística es la ciencia de los datos, y está asociada al método científico porque ambos nos proveen un conjunto de principios y procedimientos que tienden a reducir la incertidumbre mediante la obtención, resumen y análisis de la información facilitando la toma de decisiones.

~~Hipótesis estadísticas:~~
Hipótesis nula: es la afirmación de que nada esta sucediendo, no existe diferencia ni discrepancia con lo que se afirma.(H_0)
Hipótesis alternativa: es la afirmación que el investigador espera que sea cierta. Es lo contrario a la hipótesis nula. (H_1)
La hipótesis que se investiga por lo común es la que se expresa como H_1 y se llega a la conclusión de que la hipótesis investigada es verdades cuando se rechaza la H_0
Una hipótesis será rechazada si se puede demostrar estadísticamente que lo datos observados son muy poco probables de ocurrir si la teoría fuera cierta. Aquí se dice que los datos observados son estadísticamente significativos.
En estadística se recomienda decir "no rechazo de H_0" en lugar de "acepto H_0" porque si se acepta la hipótesis nula se corre el riesgo de cometer un error de tipo 1. El no rechazo de H_0 significa solamente que la evidencia muestral no es lo suficientemente fuerte como para llevar a su rechazo.
Errores:

- Tipo 1) Es el error que se comete cuando se rechaza H_0, siendo ella cierta. (e_1)
- Tipo 2) Es el error que se comete cuando no se rechaza H_0 y siendo ella falsa, consecuentemente H_1 cierta. (e_2)

Para evitar cometer errores se debería minimizar Pe_1 (Probabilidad de cometer un e_1) y Pe_2 (Probabilidad de cometer e_2) a la vez hasta llevarlos a 0, esto es imposible. Pero se han desarrollado dos técnicas de minimizar las probabilidades lo mas posible.
Técnicas:

1) Se elige un número alfa (nivel de significación) entre 0 y 1 ($0 \leq \alpha \leq 1$). Con el nivel de significación se determina la región de rechazo o regla de decisión. Y debe cumplirse que $Pe_1 < \alpha$
2) P-Valué: es la probabilidad asociada a la información. La regla de decisión es rechazar H_0 si P-Valué $\leq \alpha$. Cuanto mayor sea P-Valué mayor es la evidencia a favor de H_0.

Ensayos de Hipótesis: Clasificación.

- Unilateral: (>,<)
    - Por derecha (>)
    - Por izquierda (<)
- Bilateral (≠)

~~Inferencia estadística:~~ es el proceso de extraer conclusiones sobre la población basándose en la información de una muestra extraída de esa población.

Población: Conjunto total de objetos o individuos que presentan características comunes observables, definidas en un cierto tiempo y lugar.
Puede ser: Finita o Infinita.
Cuando estudio toda la población a cada elemento de la misma se lo denomina **unidad**.
Al número de elementos que conforman la población se los simboliza con **N**
A la medida descriptiva que se calcula a partir de todas las unidades de la población se la denomina **Parámetro** y es fijo (porque se analizan a todos los elementos)
Muestra: Parte representativa de la población que se selecciona para ser estudiada ya que la población es demasiado grande como analizarla en su totalidad.
Puede ser:

- **Representativa:** los elementos que la integran se encuentran en la misma proporción que en la población.
- **Aleatoria**: es cuando cada elemento de la población tiene la misma probabilidad de ser seleccionado

Al número de elementos que conforman a la muestra se los simboliza con **n**
A la medida descriptiva que se calcula usando todas las unidades de la muestra de la población se le denomina **Estadística** y es variable (porque depende la muestra)

Causas por las cuales se recurre a muestro

- Analizar a la población resulta muy costoso por la relación costo/beneficio
- Analizar a la población completa lleva mucho tiempo
- Al analizar el objeto de estudio se lo destruye, por lo cual si analizamos a toda la población nos quedamos sin unidades.
- La población a analizar es infinita, por lo cual es imposible analizarla en su totalidad
- La población a analizar es inaccesible

Sesgo del Muestreo:

Un método de muestreo es sesgado si produce resultados que difieren sistemáticamente de los verdaderos de la población.
Ejemplos:

Muestra conveniente es la formación de una muestra seleccionando las unidades que convienen ya sea por el resultado buscado o por la comodidad de acceso a ellas.

Muestra voluntaria es la formación de una muestra donde la decisión de participar en la muestra reside en las unidades.

Sesgo de selección es la tendencia sistemática sobre el procedimiento de muestreo para excluir o incluir a cierto tipo de unidades.

Sesgo de no respuesta es la distorsión que se logra cuando un gran número de unidades seleccionadas para la muestra no responden o se niegan a responder.

Sesgo de respuesta es la distorsión que se logra por la forma de preguntar o el comportamiento del entrevistador puede afectar la respuesta.

Clasificación del muestreo

Muestreo	No probabilístico	Por juicio Por conveniencia
	Probabilístico	Simple al azar Sistemático Estratificado Por conglomerado

Muestreo probabilístico es el método de muestreo donde a cada unidad de la población se le asigna una chance no negativa de ser seleccionada.

Simple al azar: Consiste en la elección de las unidades que formaran la muestra directo de la población siendo esta solamente enumerada. Se caracteriza por que todas las muestras de tamaño n posibles tiene la misma chance de ser seleccionadas y es

imparcial (probabilístico) Para ser representativo es necesario tener una muestra de gran tamaño. La población debe ser homogénea en base a la característica de estudio.
Métodos de selección: * Bolillero (población pequeña); * Nº Aleatorio (población mediana y grande)

Estratificado: Se obtiene dividiendo la población en subgrupos o estratos mutuamente excluyentes y sacando una muestra simple al azar o sistemática de las unidades dentro de cada estrato. Los elementos dentro de cada estrato son homogéneos respecto a la característica bajo estudio, los estratos entre sí son heterogéneos.
Se tomará una muestra más grande en estrato que en otro cuando en el primero haya mayor variabilidad entre las unidades que lo componen. En caso contrario se tomara una muestra proporcional – 10% -
La variabilidad de las unidades dentro de cada estrato es menor comparada con la variabilidad existente entre los estratos
La información es obtenida por separa para cada estrato, luego es combinada a a través del promedio ponderado para obtener una estimación de toda la población.
El método de muestreo estratificado es in sesgado cuando se usa el promedio ponderado.
Promedio General Ponderado:

$$\frac{\left(\text{Unid. en el trato1}\right)}{N}\left(\text{Estimación promedio Estrato 1}\right) + \frac{\left(\text{\# Unid. en el estrato 2}\right)}{N}\left(\text{Estimación promedio Estrato 2}\right)$$

Sistemático: Es para poblaciones homogénea. No necesito tener ordenados y numerados a todos los elementos, alcanza con tener los primeros del 1 al 10.
Se necesita conocer N y n para determinar el intervalo o salto K ya que la relación es:
K = N / n
Característica: No todas las muestras de tamaño n tienen la misma probabilidad de ser elegidas.
Método: Para seleccionar una muestra de tamaño n, con saltos de taño K. Enumero a los k primeros elementos, luego elegido de ellos uno al azar y desde allí sumo K.

Conglomerados: La población es divida en subgrupos que se caracterizan por ser homogéneos entre sí pero los elementos que componen a cada grupo entre sí son heterogéneos. Además cada elemento pertenece solo a un conglomerado o cluster.
Método: Dado que los grupos son homogéneos entre sí, y representan a toda la variabilidad de posible existencia en la población, se selecciona uno o dos por el método simple al azar y luego se estudia a fondo todos los elementos que los compone.
Aclaraciones:
- Una muestra por conglomerados no es una muestra simple al azar porque todas las muestras del mismo tamaño (no de tamaño n), tienen la misma chance de ser seleccionada.

- Una muestra por conglomerados no es una muestra estratificada porque en los estratos se estudia a todos los estados, mientras que por conglomerado se estudia algunos.
- La variabilidad de las unidades dentro de cada conglomerado es mayor comparada con la variabilidad entre conglomerados

UNIDAD 2: Organización y descripción de los datos. (Cuadernillo 3)

Significados
UNIDAD: elemento u objeto que observamos. Cuando es una persona se denomina Sujeto.
OBSERVACIÓN: información o característica registrada para cada unidad.
VARIABLE: característica que varia de unidad a unidad en la muestra o población. El resultado es un dato numérico y también va a ir variando.
CONJUNTO DE DATOS: conjunto de observaciones sobre una o mas variables.

Clasificación de las variables:

- **Variable Cualitativa o Categórica:** Son aquellas que se clasifican a las unidades en categorías. Las categorías pueden tener o no un orden. Las observaciones hechas sobre variable cualitativas se denominan "datos categórico"
- **Variable Cuantitativa o Numérica:** Son aquellas cuyas observaciones provienen de procesos de medición o conteo (finito, infinito numerable) Las operaciones aritméticas definidas sobre tales variables tienen significado. Son datos mesurables. Ej. Edad, peso, altura, ingreso, cantidad de autos, etc.
    - Variable Cuantitativa **Discreta:** es aquella en la cual puede contar número de posibles valores. Ej. N° de pisos de un edificio, N° de llamadas telefónicas, cantidad de hijos, etc.
    - Variable Cuantitativa **Continua:** es aquella en la cual puede asumir cualquier valor sobre un intervalo dado. Ej. Peso, altura, edad

DISTRIBUCIÓN DE UNA VARIABLE: Esta dada por el conjunto de valores posibles de esa variables y la frecuencia con la que ocurren cada uno de esos valores. Puede ser representada gráficamente, numéricamente y con un modelo.
Provee los posibles valores que una variable puede tomar y cuan frecuentemente ocurren estos valores. La distribución de una variable muestra el modelo de variación de la misma.
Tabla de frecuencias: Es un agrupamiento de datos en categorías mutuamente excluyentes dando el numero de observaciones en cada categoría. La tabla de frecuencias muestra el numero de elementos correspondientes a cada una de las categorías (fj – Frecuencia absoluta) y la proporción o porcentaje de artículos en cada clase es decir la fracción de elementos pertenecientes a esa clase (hj – Frecuencia Relativa = Frecuencia absoluta de la clase/n siendo n las observaciones) También muestra la frecuencia relativa porcentual (h% = hj x 100
Por lo tanto hay:
xi = variable
fj = frecuencia absoluta
hj = frecuencia relativa
hj% = frecuencia relativa porcentual
Fj = frecuencia absoluta acumulada
Hj = Frecuencia absoluta acumulada

Distribución por intervalos: * N° Intervalos: $\sqrt{n} = j$

* Amplitud: $(x_{max} – x_{min}) / j$

Graficación de variables cualitativas: Cuando se trata de variables cualitativas el calculo de frecuencias acumuladas no tiene sentido. Los gráficos mas comunes para variables cualitativas son los gráficos de sectores (torta) y el grafico de barras o pictograma.
También se puede graficar por medio del diagrama de pareto para el cual se debe calcular las frecuencias acumuladas.

DIAGRAMA DE PARETO: proporciona mas información visual que los diagramas de barras y de sectores circulares cuando la variable categórica tiene muchas categorías.
Es un tipo especial del grafico de barras horizontal donde las respuestas categorizadas se grafican en orden descendente de frecuencias y se combinan con un polígono acumulado en la misma escala.
El eje vertical de la izquierda contiene las frecuencias o %, el eje vertical derecho las categorías de Interés y el de la izquierda los % acumulados de 0 a 100 y las barras separadas uniformemente son del mismo ancho.
Lo importante al ver este diagrama se buscan las magnitudes de las distintas en las alturas de las barras que corresponden a las categorías adyacentes decrecientes y los % acumulados de las mismas.

GRAFICO DE BARRAS:
muestra el % de items que salen en cada categoría. Muestra una barra para cada categoría, el ancho de la barra no tiene importancia pero debería ser uniforme. Las barras pueden ser verticales u horizontales. Puede ser usado para representar 2 categorías cuantitativas al mismo tiempo lo que se llama grafico de barras compuesto.

PICTOGRAMA:
Las barras son reemplazadas por diagramas relacionados con algún tópico ejemplo casas, personas.

Graficación de variables cuantitativas: Cuando se encuentra una distribución de frecuencias de variables cuantitativas comúnmente se las grafica por medio de:

1- **Grafico de frecuencia**
2- **Grafico de tallo y hoja**
3- **Histograma**
4- **De caja o bigotes**

1-GRAFICO DE FRECUENCIA: es una manera rápida de mostrar la distribución de los datos sobre una recta. Cada punto o valor de la variable esta representada por una X, sobre la escala adecuada. Puede ser horizontal o vertical. La frecuencia o numero de valores que se repiten será representada en otra escala.
Pasos para su construcción:
1-dibujar una recta
2-marcar los valores max y min sobre un eje real
3-completar la escala p/ los n° con incrementos igualmente espaciados.
4-marcar cada valor observado c/una X sobre la escala adecuada
5-si hay 2 o mas items c/el mismo valor debemos apilarlos verticalmente.
Si una o mas observaciones están alejadas del resto, estas se denominan valores extremos. Un conjunto de datos separados del resto de los datos forman una concentración, conglomerado o racimo.
Un claro, brecha o gap esta dado por la distancia entre las observaciones.
FORMA DE LAS DISTRIBUCIONES: Sirve para ver la forma gral., centro aproximado de distribución y cualquier desviación de la forma gral.

2-DIAGRAMA DE TALLO Y HOJA: es una forma rápida de mostrar la distribución de un conjunto de datos con un N° relativamente pequeño de unidades. Ventaja: retiene los valores reales de la variable.
Hoja = ultimo digito
Tallo = lo que esta antes de la hoja
No se usa cuando tengo muchos datos, las hojas se ordenan de menor a menor a mayor, y si comparo 2 procesos de menor a mayor en espejo lo que sirve para comparar las disimetrías.
3- HISTOGRAMA: opción para mostrar la distribución de una variable cuantitativa cuando la cantidad de datos es grande, no mantiene los valores numéricos actuales. Muestra la distribución de una variable a través de la frecuencia o porcentaje del total de valores que hay en todo el rango de la variación.
Pasos para su construcción:
1-identificar el min y max color observado de la variable, calcular el rango (Xn –Xi)
2-dividir el rango en clases o intervalos de igual amplitud (las clases deben cubrir el total del rango de los valores, sin superponerse)
3-contar el N° de observaciones que caen en cada clase = frecuencia absoluta
4-dibujar el eje horizontal y marcar las clases sobre el
5-en el eje vertical se puede representar la frecuencia absoluta, la proporción o %.
6-dibujar un rectángulo (barra vertical) sobre cada clase con la altura igual a la frecuencia, proporción o %.

4-GRAFICO DE SERIE TIEMPO: grafica las observaciones contra el tiempo o en el orden n el que se obtuvieron. Los puntos consecutivos se conectan con líneas para ayudarnos a apreciar si la distribución es pareja o parece cambiar con el tiempo.
Patrones que debemos encontrar en un grafico de este tipo:
TENDENCIA: creciente o decreciente, cambios en la ubicación del centro, cambios en la variación o dispersión.
COMPONENTE ESTACIONAL O CICLO: patrones del comportamiento que se repiten con regularidad.

UNIDAD 3: Medidas resumen (cuadernillo 4)

Medidas descriptivas: Los datos sobre variables cuantitativas se pueden resumir con nitidez en tres formas:
* De disposición central o tendencia central
    - Media
    - Mediana
    - Moda
* De dispersión
    - Rango
    - Rango intercuartil
    - Desvío estándar
    - Coeficiente de variación
* De forma

De disposición central o tendencia central
Me dan una idea de donde esta el centro de la distribución, equilibrio, siempre va acompañada de una medida de dispersión. Son valores alrededor de los cuales las observaciones tienden a agruparse y permiten localizar el “centro” de la distribución.

MEDIA ARITMETICA: (x) se obtiene sumando las observaciones y dividiendo por el Nº de observaciones (n). Es sensible a las observaciones extremas.

$X = \sum(x\ f_j) / n$
Propiedades:
1) Todo conjunto de datos tiene una media aritmética
2) Para calcular la media se toman la totalidad de los datos, por lo cual si dentro de los datos observados existen valores extremadamente chicos o grandes la media aritmética no resulta representativa.
3) Es una medida útil para la comparación de poblaciones
4) Cada conjunto de datos tiene una sola media aritmética.

MEDIANA: (Mna) conjunto de n observaciones ordenadas de menor a mayor, es un valor tal que la mitad de las observaciones es menor o igual a ese valor, y la otra mitad de las observaciones es mayor o igual a ese valor. Se deben ordenar los valores observados de menor a mayor y si el Nº es impar la mediana será la suma /2 de los que ocupen la posición central y si es par la posición central.
Propiedades:
1) La mediana es resistente, no cambia o cambia muy poco con las observaciones extremas.
2) La mediana es única.

Mna de orden: ayuda a ubicar el lugar de la mediana. Primero se la calcula (Mna orden = n+1 / 2) y luego hay que fijarse en la frecuencia absoluta acumulada (Fj) para obtener el valor de la Mna.

MODA: (Mo) valor de la variable que ocurre con mayor frecuencia o sea el valor que tiene la frecuencia mas alta entre todas las observaciones. Si hay dos modas se llama bimodal. A veces la moda no es usada como medida de centro dado que el valor mas frecuente podría estar lejos del centro de la distribución sin embargo se tiene en cta. Para datos cualitativos.

CUARTILES (Q_1, Q_2, Q_3, Q_4): Son medidas de posición no central. Los cuartiles son medias descriptivas que parten a los datos ordenados en cuatro partes.
- Primer cuartil (Q_1): Es un valor tal que lel 25% de las observaciones son menores y el 75 % de las mismas son mayores. Para calcularlo se debe calcular el cuartil de orden (Q_1 de orden = n+1 / 4) luego me fijo en Fj y obtengo el valor del cuartil 1.
- Segundo cuartil (Q_2): Coincide con la mediana
- Tercer cuartil (Q_3): es un valor tal que el 75% de las observaciones son menores y el 25 5 son mayores. Para calcularlo primero se debe el tercer cuartil de orden (Q_3 de orden = n+i * ¾) y luego me fijo en Fj y así obtengo el tercer cuartil.
- Cuarto cuartil (Q_4) Coincide con el total de la muestra y es = al xi máximo.

Posiciones relativas de la media aritmética, el modo y la mediana:
- **Distribución simétrica**: tanto la media aritmética, el modo y la mediana coinciden , es decir dan igual.
- **Distribución asimétrica**: Se recomienda calcular la mediana. No es aconsejable calcular la media aritmética porque no es representativa, y no es aconsejable calcular el modo porque la distribución puede ser bimodal o que este no resulte representativo de l muestra.

De variación o dispersión
Son útiles pero a menudo dan una interpretación incompleta de los datos describen la dispersión que se encuentra en los datos. Un valor grande en ella indica mayor variación. Si los datos provienen de una muestra las medidas de variación se llaman estadísticas. Si

los datos constituyen la población entera, las medidas de variación serán parámetros. La notación para representar la desviación estándar de una muestra diferirá de la de la desviación estándar de la población.

RANGO: (Rgo) es la diferencia entre la mayor y menor valor de las observaciones de un conjunto de datos. El hecho de que utilice los valores extremos puede causar una distorsión del modelo real de la variación (no considera la forma en que se distribuyen los datos entre los valores mas pequeños y los mas grandes)

RANGO INTERCUARTIL: (RI): mantiene una idea de rango pero no esta influenciado por los valores extremos. Se dividen los datos ordenados en cuatro partes iguales y ver la distancia de las dos partes extremas.
Para dividir los datos, 1 se halla la mediana (Q2) y luego la mediana por mitades nuevamente (Q1 y Q3). RI=Q3-Q1
Considera la dispersión de la mitad central de los datos.

DESVIO INTERCUATILICO (DI): Indica la amplitud promedio del 50% central de las observaciones respecto de la mediana.
DI = Q1 – Q3 / 2

VARIANCIA: Es la medida aritmética de las desviaciones de la media elevados al cuadrado. S^2

DESVIO ESTANDAR: Hace uso de todas las observaciones para su calculo. Establece la forma en que los valores varían con respecto a la media. Es la raíz cuadrada del promedio de los cuadrados de las desviaciones de las observaciones con respecto a la media. (distancia promedio entre las observaciones y el promedio). 1 hay que hallar la variancia que es el promedio de los cuadrados de las desviaciones de las observaciones con respecto a la media aritmética. La desviación estándar es la raíz cuadrada positiva de la varianza.
* $S = \sqrt{1 / (n-1) \sum (x_j - x)^2}$
* Por intervalos: $S = \sqrt{1 / (n-1) \sum (x_j . f_j - n\, x)^2}$

COEFICIENTE DE VARIACIÓN (CV) Medida de dispersión relativa que es de muy simple calculo. Si trabajo c/ una muestra esta medida es CV= desvío estándar / media aritmética X100 (desvío/media %) es muy útil porque permite comparar varios grupos en cuanto a su dispersión o variabilidad. Cuanto más pequeño, menos disperso. A mas grande mas disperso (mas heterogéneo). Es función de la media, varia de 0 a 100.

DIAGRAMA DE CAJA: grafico constituido por 5 medidas resumen (X max, Xmin, Media, Q1 y Q3). Provee una simplificación del conjunto entero de datos. Provee una medida del centro a través de la mediana y medidas de dispersión a través del RI y R. La distancia de los cuartiles a la mediana puede indicarnos asimetría, la cual es chequeada mejor mediante el histograma o el de tallo y hoja.

Nota: en este grafico puedo tener una idea aproximada de las medidas de forma, dispersión y posición. La medida de posición me la da el medio (mediana), p/ la media y desvío necesito todos los datos y en este grafico puedo sacar el RI. A mas longitud de la caja mayor dispersión hay y a menor log menor disp. La simetría o disimetría la da la distancia entre Q1 y Q3.

REGLA DEL PULGAR P/IDENTIFICAR POTENCIALES VALORES EXTREMOS. Se halla 1.5 X RI, se trazan las barreras internas (Q1-1.5xRI ;
Q3-1.5xRI) las observaciones que caen fuera de las barreras internas son consideradas como potenciales valores extremos.

UNIDAD 4: Probabilidad. (Cuadernillo 6)

Probabilidad: es la interpretación de la frecuencia relativa, aplicada en las condiciones exactamente repetibles. "es la proporcion de veces que ocurriria el evento si el proceso se repitiera muchas veces bajo las mismas condiciones" esto implica la frecuencia relativa del largo plazo del resultado.
Proporcion de veces que ocurre en el largo plazo un suceso, o sea la frecuencia relativa con la que ocurre el evento.
es una medida de la posibilidad que tiene un suceso de presentarse ante la repetición de un suceso aleatorio. La probabilidad debe estar entre 0 y1.
Existen tres enfoques de probabilidad:

PROBABILIDAD	OBJETIVA	CLASICA O A PRIORI: su método se apoya por entero en el razonamiento abstracto, no efectúa experimentos reales porque la lógica se considera suficiente para todas las respuesta. P(A): $\frac{\text{nº de resultados favorables de que ocurra el suceso A}}{\text{nº de sucesos posibles}} = \frac{n}{N}$
		FRECUENCIA O POSTERIORI: Su método se apoya en el calculo de la probabilidad basado en los resultados brindados por el experimento aleatorio, es decir debe ser llevado a cabo. P(A): $\frac{\text{nº de veces que se presenta el resultado favorable}}{\text{nº de veces que se repite el experimento}}$
	SUBJETIVA: Se aplica a casos aislados y se basa en la creencia de la persona sobre un hecho presento o no. Para dicha afirmación la persona se basa en sus conocimientos previos.	

Proceso o Experimento aleatorio: es un proceso repetible del que se conoce el conjunto de resultados posibles, pero no puede predecirse con seguridad un resultado exacto. Hay un patron de comportamiento predecible a largo plazo, (que la frecuencia relativa de un resultado se acerque a un valor constante).

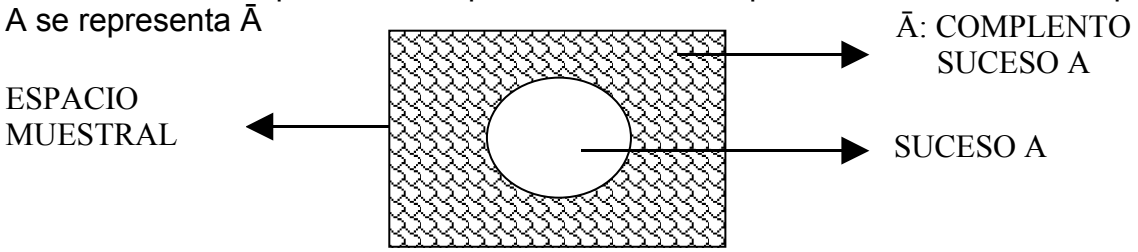
Simulación: es la imitacióndel comportamiento del azar usando artificios como generadores de numeros aleatorios, tablas de numeros al azar, etc, a traves de la cual se puede estimar la probabilidad de un evento.

Pasos básicos para calcular una probabilidad por simulación:
1-especificar un modelo para los sucesos elementales y fenómeno aleatorio subjetivo.
2-delinear como simular un suceso elemental y como representar una repetición del proceso aleatorio.
3-simular muchas repeticiones y estimar la probabilidad de un evento con la frecuencia relativa.-

Los elementos necesarios para hablar de probabilidad son:
- Espacio muestra: es una lista de todos los posibles resultados de un experimento aleatorio cuyo suceso gobierna todos los resultados aleatorios de un experimento (S)

- Evento: subconjunto del espacion muestra y se denota con imprentas mayúsculas A; B; C... se dice que ha ocurrido un evento A si al repetirse una vez el proceso aleatorio, ocurre cualquiera de los resultados de A.
- Union de dos eventos: AoB = AUB zona que contiene los resultados que estan en el evento A o B o en ambos. "al menos uno de los 2 ha ocurrido"
- **Intersección de dos eventos**: AyB = AunionB esta formado solo por los resultados que stan en ambos eventos A y B "ambos".
-

-Complemento de un suceso: dado el suceso A, el complemento de A se define como el suceso formado por todos los puntos maestrales que no están en A. El complemento de A se representa Ā



- Eventos Disjuntos: que no tienen resultados en comun = mutuamente excluyentes a intersección B es distinto de 0. no pueden ocurrir al mismo tiempo.

Reglas de la Probabilidad:
A cualquier evento A se le asigna un numero P (A) llamado probabilidad del suceso A cuando el S contiene un numero finito de Resultados posibibles hay otra técnica para asignar probabilidades a los eventos:

- asignar un aprobabilidad a c/u de los resultados individuales, entre 0 y 1 tal que su suma sea 1.
- La probabilidad de cualquier evento es la suma de las probabilidades de los resultados que componen dicho evento.

Si los sucesos del espacio muestral son equicomprobables, la probabilidad de n evento A P (A) es la proporcion de resultados del espacio muestral S que forman parte del evento A.

Reglas Basicas para asignar probabilidades a eventos

1- 0<P(A)<1 una probabilidad siempre es un numero entre 0 y 1 (cuando el suceso no puede ocurrir nunca y 1 cuando lo hace siempre)
2- P (S) = 1 si sumamos las probabilidades de c/u de los resultados individuales del espacio muestral (S), la probabilidad total es = 1.
3- P (A) = 1 – P (complemento A) regla del complemento cualquier evento y su respectivo complemento sus conjuntos disjuntos, si los unimos tendremos (S) y P (S) =1
4- Regla de la Suma: P (de que ocurra al menos uno AoB) = P(AUB) = PoB= P(A) + P(B) . P (AyB) si dos eventos A y B no tienen resultados en comun (disjuntos) entonces la Probabilidad de la que uno o el otro ocurra, es simplemente la suma de las probabilidades individuales implica que A y B son eventos disjuntos entonces P (AoB)= P (AUB)= P(A) + P (B).
5- Probabilidad Condicional: es la probabilidad condicional del evento A dad que el evento B ya haya ocurrido. P (A/B) = P(AYB)/P(B); para todo P(B)>0. Para sacar P (AyB) = P(AinterseccionB) REGLA DE LA MULTIPLICACIÓN. Entonces P(AinterseccionB)= P(A) . P(B/A) Base: para que ambos eventos ocurran, primero debe ocurrir uno, por ejemplo A y luego de esto, el evento B tamibien debe ocurrir.

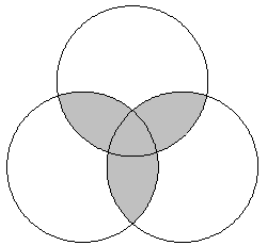
Regla de la suma: (al menos 1) Dado dos sucesos A y B quiere decir que se puede darse el suceso A o B o ambos.

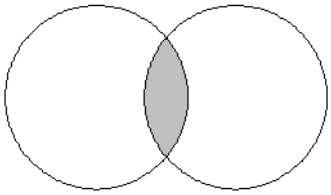
A U B = A o B = A + B = al menos 1

- Sucesos no mutuamente excluyentes:

P (A U B) = P(A) + P(B) – P(A ∩ B)

P(A U B U C) = P(A) + P(B) + P(C) – P(A ∩ B) – P(A ∩ C) – P(B ∩ C) – P (A ∩ B ∩ C)



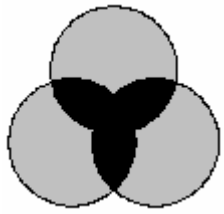


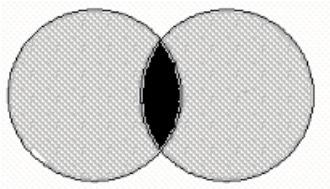
Puede ocurrir entonces debe ser eliminado

- Sucesos mutuamente excluyentes:

P (A U B) = P(A) + P(B)

P(A U B U C) = P(A) + P(B) + P(C)





Nunca ocurrirá

Puede ocurrir por ello se calcula la probabilidad

Regla de la multiplicación: (y)

- Sucesos dependientes: es cuando la ocurrencia de uno se ve afectada por la ocurrencia del otro.

P(A ∩ B) = P(AB) = P(A y B) = P(A) . P(B/A)

P(A ∩ B ∩ C) = P(ABC) = P(A) . P(B/A) . P(C/A y B)

- Sucesos independientes: es cuando la ocurrencia de uno no se ve afectada por la ocurrencia del otro.

P(A ∩ B) = P(AB) = P(A y B) = P(A) . P(B)

P(A ∩ B ∩ C) = P(ABC) = P(A) . P(B) . P(C)

EXTRACIONES CON REPOSICIÓN GENERAN SUCESOS INDEPENDIENTES.
EXTRACIONES SIN REPOSICIÓN GENERAN SUCESOS DEPENDIENTES.

Probabilidad condicional:

- La probabilidad de que habiendo ocurrido A ocurra B
  P(B/A) = P(A ∩ B) / P(A)
- La probabilidad de que habiendo ocurrido B ocurra A
  P(A/B) = P(A ∩ B) / P(B)

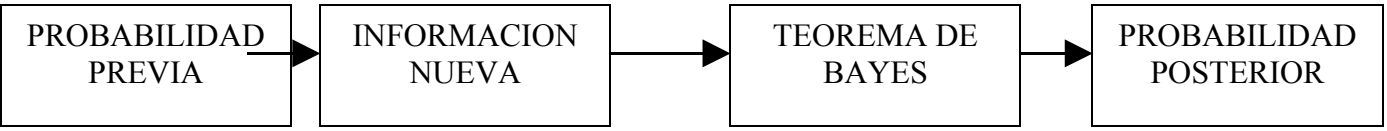
Justificación de la independencia:

P(A U B) = P(A) . P(B)

P(A/B) = P(A)

P(B/A) = P(B)

Teorema de Bayes (Corrección de probabilidades)



Probabilidad previa: $P(A) + P(\bar{A}) = 1$
Información nueva: $P(R/A)$; $P(R/\bar{A})$
Teorema de Bayes: $P(A/R) = \dfrac{P(A) \cdot P(R/A)}{P(A)\,P(R/A) + P(\bar{A})\,P(R/\bar{A})}$

Eventos Independientes: si saber que B ocurrio no cambia la P(A) es decir P(A/B)=P(A) A y B (independientes) si y solo si, P(A/B)=P(A), o P(B/A)=P(B). Si ambos eventos son independientes la regla gral. de multiplicación dice que la P de que ambos eventos ocurran juntos es el producto de las P individuales. Si B y A son independientes, P (AinterseccionB)=P(A)xP(B)

Partición y Teorema de Bayes: cuando para obtener la P (A) total se deben combinar las probabilidades de A en cada subconjunto. Los B1, B2, B3...son particiones de S si y solo si son mutuamente excluyentes y la union entre ellos es S (osea si cada resultado individual de S pertenece exactamente a uno de los eventos)
P(A)= P(A interseccion B1) + P (A intersección B2) + P (A intersección B3) luego usando la regla de la multiplicación para 2 eventos y la aplicamos a las 3 P
P(A)=P(A/B1) P(B1) + P(A/B2) P(B2) + P(A/B3) P(B3)
*Particion: sucesos complementarios cuyas probabilidades suman 1.

LEY DE LA PROBABILIDAD TOTAL

6- Si B1,B2,B3 forman una partición de S entonces

P(A)=P(A/B1) P(B1) + P(A/B2) P(B2) + P (A/B3) P(B3)

TEOREMA DE BAYES

7- Supongan que B1 Y B2 conforman una partición de S, que A es otro evento y que conocemos las P condicionales P(A/B1) Y P(A/B2) entonces la regla de Bayes nos da la P condicional:

P(B1/A)=P(A/B1) x P(B1) / P(A/B1) x P(B1) + P (A/B2) P(B2)

Variables aleatorias: es una funcion real definida en un espacio muestral. el resultado numerico de un proceso aleatorio.

1- variables aleatorias discretas: puede tomar un numero finito o infinito numerable.
2- Variables aleatorias continuas: puede tomar cualquier valos de un intervalo o conjunto de intervalos.

Probabilidad de Variables aleatorias discretas: si X es una variable aleatoria discreta que toma los valores x1; x2; x3;... xk entonces la distribución de probabilidad de x se puede presentar por una formula, tabla o grafica que indique las P (x) correspondiente a cada uno de los valores de x.
P(x)=P(X=X) si y solo si 1) P(X) >0 para todo (x,y) 2) la suma de todos los P(x)=1
La media de una variable aleatoria discreta x es el punto de equilibrio del grafico de bastones, es el valor esperado de x y se calcula asi: si x es una variable aleatoria discreta que toma valores x1, x2, x3, xk con probabilidades P1, P2, P3, PK entonces la media, o valor esperado de X=E(x)=Mx= X1 P1+ x2 P2 + x3 P3 + ...+xk Pk
La Variancia de una variable aleatoria discreta: X(desvio al cuadrado) es una medida de dispersión de los posibles valores respecto de la media.
Si x es una variale aleatoria discreta que toma valores x1,x2,x3,xk con P1,P2,P3,PK entonces la variancia de x es Var de X= desvio2 X= E(x-M)2=E(x2)-(E(x))2
Desvio: SD (x) = desvios x = raiz cuadrada E(X2) – E(x)2

VARIABLES ALEATORIAS CONTINUAS:
Es una variable aleatoria por que toma valores en algun intervalo, o union de intervalos, se dice continua.
Se asigna una P a cada valor individual o a la suma de las probabilidades seria eventualmente mayor que 1. asignamos probabilidades a intervalos de resultados y representamos esas probabilidades como areas debajo de una curva llamada curva de probabilidad.
El comportamiento de una variable aleatoria continua x esta dada por una funcion f(x) / P(x) tome valores en un subconjunto A) = el area debajo de la funcion f(x) sobre el subconjunto A.
La f (curva de densidad) debe cumplir 1) f(x)>0 2) el area debajo de la f (x) es 1
Media: es el punto de equilibrio de la f(x) de densidad P.

DISTRIBUCIÓN BINOMIAL:
Ensayo de Bernoulli: son experimentos en donde se pueden presentar solamente 2 resultados, éxito o fracaso, donde la probabilidad de éxito se simboliza P. Esta es la base de la distribución binomial de variables aleatorias.
VARIABLES ALEATORIAS BINOMIALES X:
Es el numero de exitos en n repeticiones independientes de Bernoulli donde c/ prueba tienen una probabilidad de éxito p.
1-el experimento consiste en n repeticiones idénticas e independientes.
2-cada prueba tiene 2 posibles resultados (éxito o fracaso)
3-la probabilidad de éxito p, permanece constante p/ cada prueba
4-la variable aleatoria binomial x es el numero de exitos en n pruebas XnB(n;p)

DISTRIBUCIÓN BINOMIAL
Formula
Ver en el cuadernillo

UNIDAD 5: Distribuciones de probabilidad. (Cuadernillo 5)

Modelo: Representación simplificada de un objeto del mundo real o de un fenómeno. Habitualmente están simplificados y representan solo ciertas características sobresalientes del objeto o fenómeno. Algunas técnicas estadísticas para la toma de decisiones requieren que hagamos unos supuestos sobre la población de la que se obtuvieron los datos. Estos supuestos generalmente se establecen en términos de un modelo para la población.
Supongamos que queremos estudiar una población y que cada unidad de la misma toma un valor para una variable determinada. Un modelo estadístico para esa población resumirá como se distribuyen los valores de la variable en la población. El resumen tendrá la forma de un modelo matemático que describe la relación entre los valores posibles de la variable estudiada y la proporción de elementos de la población que toman esos valores.
Los modelos nos ayudan a comprender que tipo de resultados se pueden esperar y que tan a menudo ocurren, facilitando así la toma de decisiones inteligentes entre teorías competitivas.
Ordenan y facilitan la comprensión de grandes masas de datos y sirven como marco de inferencia útil para la comparación, y determinar si una observación es poco usual o no.

x ~ f(x) (variable continua)	**Modelo Normal** **Modelo Uniforme**

x ~ f(x) (variable discreta)	**Modelo Binomial** Modelo de Wason Modelo de Geometria

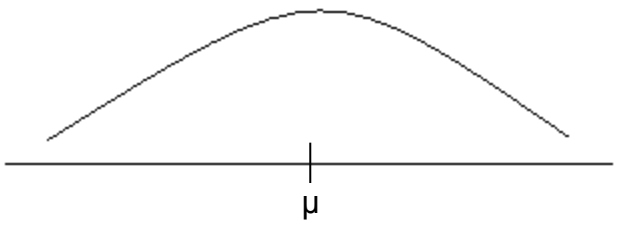
MODELOS PARA VARIABLES CONTINUAS

Curva de densidad: Es una función no negativa que describe la forma gral. De la distribución. El área total debajo de la curva es igual a 1 y las proporciones se miden como áreas debajo de las curvas de densidad.

Distribución normal: Es una función normal simétrica, campánulas y centrada en la media (M) (la media, mediana y moda coinciden ya que son simétricas y uniformes las distrib.) la dispersión esta determinada por el desvió estándar ( ).
La forma es descendiente hacia ambos lados de la media en forma cóncava hacia abajo al principio y luego hacia arriba. El punto en el que cambia la curvatura se llama punto de inflexión.
La distancia de la media a la proyección del punto de inflexión sobre el eje horizontal es la desviación estándar.
Cuando mas pequeño es el desvió mas alto será el pico de la campana porque los valores están mas centrados alrededor de la media.

Se distribuye normal con parámetros: $\mu = x$ (media aritmética)
$\sigma = S$ (desvío estándar)

Es decir : $x \sim N(\mu ; \sigma)$



μ +/- σ	68%
μ +/- 2σ	95%
μ +/- 3σ	99.7 %

Notación: x= N (M, T) la variable X esta normalmente distribuida con la media M y un desvió estándar T.

Forma de calcular el área debajo de una distribución Normal:
Si $x \sim N(\mu ; \sigma)$ se la debe estandarizar en $Z = (x - \mu) / \sigma$; y que se distribuye N (0;1), una vez estandarizada se busca el nº en la tabla de Z y allí se haya el valor %.
El signo + o – del valor de z indica si se esta ubicando a la derecha o a la izquierda de la media.

Percentiles de la distribución normal. Es el valor de la variable que acumula un P% de probabilidad.
Calculo: consiste en encontrar el valor de la variable x que represente el P%, para ello se parte de $Z = (x - \mu) / \sigma$, donde se conocen los 2 parámetros y el valor de Z, luego mediante el despeje se halla el valor de x. Para hallar el valor de Z se debe buscar en el interior de la tabla en P%.

Distribución uniforme: Es una curva plana de forma rectangular que cubre un intervalo determinado.

Se dice que una variable aleatoria continua se puede modelar por una función de densidad uniforme que viene dada 1/ (b-a), donde el campo de variación es [a;b]

Área: (b - a) . c = 1 ➡ c = 1/(b – a)

Promedio Poblacional: μ = $E_{(x)}$ = (a + b) / 2