Unidad 1: Datos estadísticos y Etapas para su análisis

¿Qué es la estadística?

Método de la recolección en masa de datos que serán utilizados para la toma de decisiones frente a la incertidumbre.

La estadística trata de la selección, análisis y uso de datos con el fin de resolver problemas

Existen dos ramas de la estadística:

Descriptiva: Se encarga de la recopilación, análisis y organización de las variables (datos) para su posterior interpretación.

¿Cómo?

Mediante tablas, gráficos y también dependiendo de donde se saquen las variables se lo denominara como **parámetro** en caso de **población** o **estadístico** en caso de **muestra.**

Inferencia: Técnica que nos permite obtener una generalización o conclusión de un parámetro de la población mediante estadígrafos con un cierto agregado de incertidumbre.

Diferencia entre población y muestra

Población: Es la totalidad de individuos de la cual se desea obtener información, de los cuales podemos denotar que son infinitos o finitos.

Cuando hacemos una medición completa de cada uno de los elementos se le dice **Censo.**

Muestra: Es un subconjunto seleccionado de la población y a la medición de cada uno de estos elementos se lo denomina Muestreo.

Generalmente realizar un censo es costoso por lo cual se suele recurrir a un muestreo. Hay que tener en cuenta el hecho de que al seleccionar un muestreo hay que hacerlo mediante **"Procedimientos de muestreo"** para así no elegir una muestra no representativa.

Datos Estadísticos

Una información es dato estadístico cuando es un conjunto o conjuntos de valores factibles que puedan ser comparados, analizados e interpretados. Por ende los datos aislados carecen de interés.

Las medidas que se toman de la información se les llaman variables y podemos encontrar las siguientes.

Variable Cuantitativa: Dato de tipo numérico que son el resultado de un proceso de **conteo** a lo cual se la conocerá como **"variable discreta"** o de **medición** que se denominara como **"variable continua"**.

Variable Cualitativa: Dato expresado en palabras ya que es imposible medirlas solo clasificarse porque representan una categoría.

Unidades estadísticas o de análisis

Son cada uno de los elementos que pertenecen a un conjunto. Por ejemplo si se quiere analizar la inflación de cada una de las provincias del país la unidad estadística serian las provincias.

Unidad de relevamiento

Lugar donde se encuentran las unidades estadísticas.

Escala de medida

Podemos encontrar cuatro tipos que son**:**

- Nominal: Consiste en categorías cualitativas en la que se registran los números de observaciones las cuales son mutuamente excluyentes ósea que dos proposiciones no pueden ser verdaderas a la vez y además no incluyen un orden lógico.

- Ordinal: Presenta categorías cualitativas mutuamente excluyentes, con un orden lógico.

Como se puede observa tanto la ordinal como la nominal trabajan con categorías las cuales, es necesario recalcar, respetan los siguientes principios: Mutuamente excluyente, Único principio clasificatorio (No mezclar variables) y Colectivamente exhaustiva (Que se puedan clasificar todas las unidades del conjunto).

- De Intervalos: Consiste en datos cuantitativos, se suele utilizar cuando se toman valores numéricos de algunos elementos y se pueden establecer los intervalos entre esas medidas. Los intervalos por regla deben tener la misma distancia y además cada escala tiene un punto cero definido arbitrariamente.

- De Razón: Consiste en datos cuantitativos, que tiene intervalos con distancias constantes, tiene un punto cero no arbitrario y además la razón en los números tiene algún significado.

Operaciones permitidas	
Nominal	Igualdad
Ordinal	Mayor o menor
De Intervalos	Suma y resta
De Razón	Multiplicación y división

Siempre teniendo en cuenta que “De razón” cuenta con las operaciones anteriores, “De Intervalos” con las anteriores pero no con la “De razón” y así sucesivamente.

Generalmente se puede descender de una escala a otra, pero nunca subir en la jerarquía salvo de ordinal a Intervalos.

Las variables según el tipo de unidad de referencia

Podemos encontrar:

- Individuales: Hace referencia a las características de una unidad de análisis o individuo. Ejemplo: Notas de un alumno.

- Agregadas: Hace referencia a las características del agregado en las que actúan los individuos la cual es independiente de los mismos. Ejemplo: Fracaso escolar de individuos en una escuela pública o privada.

- Mixta: Hace referencia a las características agregadas pero que son basadas en las características de los individuos. Ejemplo: Un colegio es de clase media por que sus alumnos son de un estrato económico medio.

Las variables según el papel que cumplen en la investigación

Podemos encontrar:

- Independientes: Variables toman o ya vienen con valores que influyen a otras haciendo que estas cambien de valor.
- Dependientes: Variables sujetas a otras.
- Control: Sirven para comprender la relación entre una variable dependiente e independiente. Esta se aplican en casos como:

1. Para averiguar porque entre dos variables hay dependencia.
2. Para verificar si la relación es casual o de dependencia
3. Para probar las relaciones en todas las circunstancias o si solo se manifiesta en situación concretas

Etapas del método científico en el análisis de datos

Es necesario para obtener respuestas congruentes durante el análisis de la información. Consta de 5 pasos:

1. Formular hipótesis: Formular claramente lo que se quiere estudiar

2. Diseño del experimento: Decidir entre hacer un censo o un muestreo

3. Recopilación de datos estadísticos: Etapa donde se recogen los datos de las fuentes. Entre los datos a recopilar encontramos la siguiente ramificación:

    - Datos primarios: Son los que hay que recoger mediante el uso de técnicas como los grupos de interés, entrevistas Teléfono, Cuestionario por corre, Puerta en puerta, Abordaje en un centro comercial, Entrevistas

    - Datos secundario: Son los que ya están disponibles por ejemplo en una biblioteca.

    Luego tenemos los métodos de relevamiento que son:

    - Estáticos: Los datos son obtenidos a una fecha determinada

- Dinámicos: Los datos corresponden a las operaciones que se realizan de forma continua atreves del tiempo.

4. Tabulación y descripción de resultados: Se refiere a la organización y presentación de la información recogida para la interpretación de los mismos. De los cuales podemos ramificar dos opciones:

    - Escrito: Esto es solo para cuando tenemos un número reducido de datos que presentar

    - Distribución de frecuencias: Cuando hay una cantidad de masiva de datos se utilizan diagramas y gráficos.

5. Generalización o Inferencia Final: Este último paso es cuando se decidió trabajar con una muestra, ya que el problema formulado anteriormente siempre va a referirse a una población necesitamos inferir los datos extraídos de la muestra mediante el uso de la estadística inferencia.

Unidad 2: Organización y presentación de datos estadísticos

Tablas estadísticas

Podemos encontrar dos tipos:

- Tablas de contingencia: Utilizadas para variables cualitativas dentro de las cuales podemos encontrar (Diagrama de barras, circulo radiado, Zonas, Diagrama de pareto)

- Tablas de distribución de frecuencias: Utilizadas para variables cuantitativas dentro de las cuales podemos encontrar muchas como para ser descritas en una oración :v

Las partes principales de una tabla son:

- Titulo: Descripción del contenido de la tabla
- Encabezamiento: Titulo de los datos presentados
- Conceptos o Columna Matriz: Clasificaciones que aparecen en las hileras de las tablas, a la izquierda por lo general.
- Cuerpo: Formado por datos estadísticos
- Notas del encabezamiento: Aclaran ciertos aspectos que no han sido puestos en el titulo
- Notas al pie: Aclaraciones
- Fuente: Lugares de donde se ha obtenido la información.

Para construir una tabla es necesario tener en cuenta:

- Simplicidad
- Solo un tema por vez en la tabla

- Ordenamiento de las clasificaciones mediante el uso de la cronología, geografía, cuantitativo y cualitativo.
- Destacar de cifras importantes.
- Redondear cifras cuando no se necesita exceso de detalle
- Agregar anotaciones

Formas de agrupar variables cuantitativas

Antes que nada hay que saber distinguir entre:

- Valores posibles: Son todos los valores que puede asumir
- Valores realmente observados: Son los valores que de entre los posibles han sido efectivamente obtenidos

Según el comportamiento (Discreto o continuo) y la cantidad de valores observados podemos agruparlas en base a los siguientes tipos:

Series simples o datos no agrupados

Cuando contamos con una cantidad pequeña de valores y no se repiten con independencia del tipo de variable.

x_1	Primer valor
x_2	Segundo valor
x_3	Tercer valor
x_n	Valor enésimo

Cabe aclarar que al ser x minúscula nos estamos refiriendo a que los valores se obtuvieron de un muestreo, en caso de ser X mayúscula nos referimos a un censo. Lo mismo para la cantidad de elementos dentro de la muestra es “n” y para la cantidad de elementos de una población es “N”

Datos agrupados o en lista

Cuando la cantidad de datos a analizar es muy grande se hace uso de tablas que se denominan como distribuciones de frecuencia cuyo objetivo es compactar y simplificar la información para su mayor entendimiento.

Las distribución de frecuencias pueden ser: Unidimensionales (Solo una variable), Bidimensionales (relación entre dos variables), Multidimensionales (relación entre más de dos variables).

Solo utilizaremos las Unidimensionales las cuales pueden ser:

Titulo: En Lista

y_i	n_i	h_i	N_i	H_i
y1	**n1**	**h1**	**n1**	**h1**
y2	**n2**	**h2**	**n1 + n2**	**h1 + h2**
y3	**n3**	**h3**	**n1 + n2 + n3 = n**	**h1 + h2 + h3 = 1**
m	**n**	**1**		

Los componentes de la tabla son:

- **Y_i**: Es la variable la cual representa los valores observados
- **m**: Es la cantidad de valores distintos observados en la variable
- Frecuencia absoluta **(n_i)**: Cantidad de veces que se presento un valor observado en la variable los cuales sumados deben ser igual a n
- **n**: Es el número de elementos de la población o muestra (En caso de ser “n” es muestra, “N” es población)
- Frecuencia relativa **(h_i)**: Proporción en la que se presenta el valor obtenido como el cociente entre ni/n y la suma de todos ellos deber ser igual a 1.
- Frecuencia absoluta acumulada **(N_i)**: Suma sucesiva de las frecuencias absolutas hasta llegar al número de elementos totales n
- Frecuencia relativa acumulada **(H_i)**: Suma sucesiva de de las frecuencias relativas hasta llegar a 1.

En Intervalos

Para casos en los cuales el número de observaciones es grande y el número de valores distintos que asumió la variable es casi tan grande como el total de observaciones es necesario el uso de clases de intervalos.

Antes de armar la tabla de distribuciones es necesario definir los intervalos:

1. Definimos el recorrido de la serie "**R"** utilizando los valores que se nos dieron como datos: R = Vmax – Vmin
2. Definimos la amplitud del intervalo "**C**" lo que nos va a dar como resultado la distancia entre el valor mínimo y el máximo de cada intervalo:

   C = R/ N° de intervalos

   - **El N° de intervalos es arbitrario y siempre tiene que ser impar.**
   - **En caso de darnos un número decimal tenemos que redondearlo para arriba hasta llegar a un numero divisible por dos y entero.**
3. Calcular los limites que aseguren que tanto el valor más chico como el más grande de la serie estarán dentro de algún intervalo de clase.

   - Primero calcular el recorrido Ampliado "**R'**": R' = C * N° de intervalos
   - Hacer la diferencia entre el recorrido ampliado y el recorrido de la serie: R' – R = X
   - Obtener los limites: $\mathbf{L_{inf}}$**= Vmin - X** o $\mathbf{L_{max}}$**= Vmax + X** pero solo uno se puede alterar, el otro limite debe quedar intacto.

Titulo: En Intervalo

y'i-1 – y'i	yi	ni	hi	Ni	Hi
y'0 – y'1	y1	n1	h1	n1	h1
y'1 – y' 2	y2	n2	h2	n1+n2	h1+h2
y' 2 – y'3	y3	n3	h3	n1+n2+n3 = n	H1+h2+h3 = 1
		n	1		

Los componentes de la tabla:

- Marca de clase **(Y_i)**: Es el punto medio del intervalo y se calcula como la suma de los extremos divido 2: (y'i-1 + y'i) / 2.
- Frecuencia absoluta **(n_i)**: Cantidad de veces que se presentaron entre esos intervalos.
- **n**: Es el número de elementos de la población o muestra (En caso de ser "n" es muestra, "N" es población).
- Frecuencia relativa **(h_i)**: Proporción de valores comprendida en cada intervalo.
- Frecuencia absoluta acumulada **(N_i)**: Suma sucesiva de las frecuencias absolutas hasta llegar al número de elementos totales n
- Frecuencia relativa acumulada **(H_i)**: Suma sucesiva de de las frecuencias relativas hasta llegar a 1.

Si se quisiera conocer la frecuencia absoluta podemos usar:

- Hojas de cálculo: Se representa cada unidad dentro de una clase con una raya diagonal
- Forma de asiento: Las clases se anotan al inicio de cada columna y debajo de ellas se va poniendo el valor que está comprendido en la misma. Al final de cada columna se pone las m cantidades dentro de la misma

Formas de agrupar variables cualitativas

Distribución categórica o tablas de contingencia

Se suelen agrupar en categorías, soliendo variar en el hecho de que haya subcategorías. Siempre y cuando respetando los principios de la colectividad exhaustiva y lo de mutuamente excluyente, además es necesaria la definición específica de cada categoría para evitar la violación de uno de los principios.

Representaciones graficas

Gráficos Lineales (Para variables en lista)

- Bastones: Se trabaja con variables discretas y con el uso de frecuencias relativas o absolutas. Se hace uso de este tipo de grafico cuando tenemos un agrupamiento de variables en lista. El eje "X" representa los valores de la variable y el "Y" los valores de cualquiera de las dos frecuencias.

- Acumulativo de frecuencias: Se trabaja de la misma manera que el grafico de bastones, las diferencias que trae es que se utilizan las variables acumulativas y el grafico queda en forma de escalera.

Gráficos de superficie (Para variables en intervalos)

- Histograma(n,h): Se trabaja con variables continuas y con el uso de frecuencia absoluta o relativa. Los ejes deben comenzar obligatoriamente en 0 haciendo a veces necesario el uso de interrupciones de escala.

- Polígono de frecuencias: Se suele hacer uso del histograma para trazar este tipo de grafico, ya que se traza una línea desde las marcas de clase (centro) de cada rectángulo. Es útil cuando se necesita hacer comparaciones con otros histogramas, el único defecto es el hecho de que no abarca todas las frecuencias ósea no son proporcionales.

- Curva suave: Se suele presentar cuando el histograma tiene una amplitud muy pequeña, este tipo de grafico tiene la característica de representar la verdadera distribución de la población de la que se extrae la muestra y son a menudo requeridos en la estadística inferencial.

- Escalonado(N,H): Trabaja con variables continuas y hace uso de las frecuencias acumuladas operando de manera similar al histograma

- Ojiva(N,H): Representa las distribuciones acumuladas en forma de un diagrama de líneas en las cuales encontramos la ojiva menor y la mayor. La ojiva menor se utilizan los limites superiores de cada clase y utiliza la columna de la frecuencia absoluta de manera ascendente mientras que la ojiva mayor utiliza el límite inferior de cada clase y utiliza la columna de frecuencia absoluta/relativa acumulada de manera descendente

Gráficos especiales(Para variables categoricas):

- Diagrama de barras horizontales(n,h): Hace uso de variables cualitativas y trabaja con frecuencias absolutas o relativas siendo estas representadas por el eje x mientras que las categorías son representadas por el eje y. El ancho de cada barra es puramente arbitrario y la orientación de las barras puede también ser vertical

- Circulo radiado: Representa variables categóricas (cualitativas) utilizando también frecuencias absolutas o relativas pero pasadas a porcentaje.

- Zonas: Utilizado para comparar varios fenómenos usualmente en un orden cronológico, geográfico, etc.

- Diagrama de Pareto: es una gráfica para organizar datos de forma que estos queden en orden descendente, de izquierda a derecha y separados por barras. Teniendo como objetivo el analizar las causas, estudiar resultados y planear mejoras continuas

Unidad 3: Descripción de datos estadísticos

Medidas descriptivas

Representan el comportamiento de un conjunto de datos dándonos una idea de las tendencias y la dispersión de los mismos como la medida de posición, de asimetría, de puntiagudez y de dispersión

Medidas de posición

Son medidas que me dan la tendencia central de la distribución de datos como:

Media (**M(X) o M(x)**): Suma aritmética de todos los valores observados divido la cantidad de lo mismos que indica la localización del centro de la distribución

Fórmula para cada caso	
Serie simple	$\frac{\sum_{i=1}^{n} xi}{n}$
Variables en lista	$\frac{\sum_{i=1}^{n} yi * ni}{n}$
Variables en intervalos	$\frac{\sum_{i=1}^{n} y'i * ni}{n}$

Propiedades: Siendo “k” una constante y “X” e “Y” variables.

- M(k) = k
- M(k * Y) = k *M(Y)
- M(k ± Y) = k ± M(Y)
- M(X ± Y) = M(X) ± M(Y)

Mediana (**Me o me**): Valor que divide al conjunto de datos en dos, que no supera a no más de la mitad de las observaciones y es superado por no más de la mitad de las observaciones.

Las formas para calcular varían en cada caso

- **Serie simple**: Se ordenan los valores que asumió la variable de manera ascendente o descendente, en caso de ser par la cantidad de datos se suman los dos valores centrales y se divide por dos, y en caso de ser cantidad impar el valor de la mediana queda a la vista ya que se encuentra en el centro dividiendo la serie de datos en dos.

- **Variable en lista**: Para buscar la mediana es necesario contar con la frecuencia absoluta(n_i) y la acumulada(N_i) de la misma, luego:

    a. Buscar el primer valor que supere el cociente n/2 en la frecuencia absoluta acumulada, lo llamaremos "Nj".

    b. Una vez establecido "Nj" ver si el valor anterior a ese "Nj-1" es menor o igual al cociente n/2.

    c. En caso de ser menor la mediana es el valor de la variable Yi que está en la misma fila Nj o en caso de ser igual entonces la mediana es (Yi +Y_{i-1}) / 2

- **Variable en intervalos**: Se trabaja de la misma manera que en lista, solo que se tienen en cuenta un parde cosas mas:

    1. En caso de ser Nj-1 < n/2 se utiliza la siguiente formula

$$me = (Y_{i-1}) + C * \frac{\frac{n}{2} - (N_{i-1})}{nj}$$

    2. En caso de ser Nj-1 = n/2 la mediana es igual a (Yi-1)

Moda (**Md o md**): Valor de la variable que más se repite

Formulas para cada caso	
Serie simple	Basta con ordenar los valores y ver qué valor se repite más veces, en caso de haber más de uno que se repite igual cantidad de veces se los toma a todos ellos como moda o si no hay valores que se repiten no hay moda
Variable en lista	Es necesario ver la frecuencia absoluta para determinar qué valor Yi se repite más veces
Variable en intervalos	Se consulta también la frecuencia absoluta, la clase que tenga un mayor "n_i" será la clase modal y su marca de clase es la moda.

Medidas de Dispersión

Son medidas que indican la concentración o la dispersión de los datos con respecto a la promedio de la distribución, estas medidas son a menudo complemento de las medidas de posición ya que por sí solas no caracterizan totalmente una distribución o también sirven para hacer comparaciones entre conjuntos. Por ejemplo si hay mucha dispersión respecto al promedio la medida de posición no representa un valor altamente significativo del conjunto

Recorrido: Es la diferencia entre el valor mínimo y el máximo del conjunto de datos. Esta medida es muy sensible a errores ya que si aparece un valor fuera de lo común ya sea muy grande o pequeño no es posible usarla como medida de dispersión ya que no revela nada acerca de la distribución ordinara de los datos.

Varianza: Se define como la media aritmética de los cuadrados de las desviaciones con respecto a la media aritmética

Formulas para cada caso	
Serie siemple	$\frac{\sum_{i=1}^{n} yi^2}{n} - M(Y)^2$
Variable en lista	$\frac{\sum_{i=1}^{n} yi^2 * ni}{n} - M(Y)^2$
Variable en intervalo	$\frac{\sum_{i=1}^{n} y'i^2 * ni}{n} - M(Y)^2$

Nota: En la inferencia estadística se suele usar como denominador n – 1 para corregir una variabilidad.

Propiedades: Siendo k una constante y “Y” e “X” variables

- V(k) = 0
- V(k*Y) = k² * V(Y)
- V(k+Y) = V(Y)
- V(-k+Y) = V(Y)

Desviación Estándar: Es la raíz cuadrada positiva de la varianza y es utilizada para comparar dos o más conjuntos de datos siempre con respecto a la media aritmética y respetando la unidad de los elementos

Formas de calculo	
Serie simple, v. en lista y en intervalo	$Ds = \sqrt{V(Y)}$

Propiedades: Siendo k una constante y “Y” e “X” variables

- Ds(Y) >= 0
- Ds(k) = 0

- Ds(k + Y) = Ds(Y)
- Ds(Y - k) = Ds(Y)
- Ds(k*Y) = k * Ds(y)
- Ds(X ± Y) = $\sqrt{V(Y) \pm V(X)}$

Coeficiente de variación: Es la razón entre la desviación estándar y la media aritmética la cual nos dará como resultado un porcentaje. Esta medida de dispersión o variabilidad es muy útil a la hora de comparar dos o más conjuntos con diferentes dimensionabilidad ósea que ambos conjuntos tienen diferentes magnitudes o unidades

La forma de cálculo es:

$$Cv = \frac{Ds}{M(y)}$$

Apreciación del Coeficiente de variación:

Coeficiente de variación	Apreciación
26% o mas	Muy heterogéneo
Entre 16% y 26%	Heterogéneo
Entre el 11% y el 16%	Homogéneo
Entre el 0% y el 11%	Muy homogéneo

Si el Cv <= 20% se dice que el promedio es representativo.

Medidas de asimetría

Cuando dos conjuntos de datos coinciden en sus medidas de posición y dispersión se suele recurrir a las medidas de asimetría que muestran de manera grafica la tendencia central de cada una de ellas.

- Simetría: Es cuando las medidas de posición son iguales (mediana, moda, media) lo cual forma un grafico con la campana de Gauss

- Asimetría Negativa o a la izquierda: Es cuando las media es menor a la mediana y esta menor a la moda (Media < mediana < moda)

- Asimetría positiva o a la derecha: Es cuando la meda es mayor a la mediana y esta mayor a la moda (Media > Mediana > Moda)

Medidas de puntiagudez

Indica la velocidad con la que sube o baja la curva en una distribución de datos la cual se calcula con el coeficiente de curtosis de Fisher

La forma de cálculo es:

$$k = \frac{M(Y)^4}{V(Y)^4} - 3$$

En base al resultado diremos que:

- Mesocurtica **(K = 0)**: Esto presenta que el grado de concentración es medio alrededor de los valores centrales de la variable.
- Leptocurtica **(K > 0)**: Alto grado de concentración alrededor de los valores centrales de la variable.
- Platicurtica **(K < 0)**: Grado reducido de concentración alrededor de los valores centrales de la variable.

Unidad 4: Teoría de probabilidades

Esta teoría como lo indica su nombre se basa en las probabilidades que es la que nos ayuda a medir los riesgos asociados a la toma de decisiones

Experimento aleatorio

Es aquel que conduce a dos o más resultados posibles. Este puede estar definido por una sola prueba o por un conjunto de pruebas bajo las mismas condiciones

Espacio probabilístico o muestral

Conjunto de resultados posibles (elementos o puntos de muestra) de un experimento aleatorio simbolizado por la letra Ω (Omega).

Los puntos de muestra deben ser Mutuamente excluyentes y colectivamente exhaustivos es decir por lo menos uno de los eventos debe ocurrir cuando se realiza un experimento.

Omega también puede ser considerado como el conjunto de productos cartesianos y estos pueden ser representados por un sistema de coordenadas cartesianas o un diagrama de árbol.

En la mayoría de los casos el espacio probabilístico está compuesto por un número finitos de puntos no obstante se pueden presentar casos como: Infinidad numerable de punto e infinidad no numerables de puntos.

Evento o Hecho

Subconjunto del espacio probabilístico denotado con la letra "**E**" y los tipos de eventos son:

- Evento simple o elementales: Cuando está formado solo por un punto o elemento del espacio probabilístico. En base a esto

podríamos decir que el espacio muestral es un conjunto de eventos simples.

- Evento compuesto: Cuando está formado por más de un punto o elemento del espacio.

- Evento cierto: Contiene todos los eventos simples del espacio probabilístico, es decir, es igual al espacio probabilístico.

- Evento imposible: Contiene el elemento vacio, es decir que no puede ocurrir nunca.

Y a su vez los eventos se clasifican en

- Eventos mutuamente excluyentes: Dos eventos o más que pertenecen a un mismo espacio probabilístico son mutuamente excluyentes si la intersección entre ellos da como resultado el espacio vacio

- Eventos no mutuamente excluyentes: Dos eventos o más que pertenecen a un mismo espacio probabilístico no son mutuamente excluyentes si la intersección entre ellos da como resultado un conjunto distinto del vacío. Además estos eventos presentan independencia o dependencia.

    1. **Independencia:** Cuando la ocurrencia o no de un evento en cualquier prueba no afecta la ocurrencia del evento en las pruebas siguientes
    2. **Dependencia:** Cuando la ocurrencia o no de un evento en cualquier prueba afecta a la ocurrencia del evento en las siguientes pruebas.

- Colectivamente exhaustivos: Dos eventos o más que perteneces a un mismo espacio probabilístico son C.E si la unión entre ellos da como resultado el espacio muestral en sí.

Nota: Un evento ha ocurrido si al menos uno de sus elementos se ha presentado.

Teoría de probabilidades

Se cuenta con tres tipos de teorías

- Clásica: Se basa en que todos los resultados de un experimento tienen la misma probabilidad de ser elegidos. Esto presenta un panorama ideal en el cual por ejemplo una moneda está perfectamente equilibrada (Que no presenta desbalance alguno) o que un dado no está modificado para así lograr que un numero tenga más oportunidad que otro, ante estos casos la teoría clásica da resultados erróneos

$$P\ (E1) = \frac{Casos\ favorables\ al\ evento}{Casos\ posibles\ (n)}$$

- Frecuencia: Se basa en repetir n veces una prueba siendo cada prueba igual a la anterior, es decir en las mismas condiciones. Por lo tanto si se repitió un experimento n veces y se obtuvieron "ni" resultados la razón entre ambos nos dará la probabilidad de que suceda ese hecho.

$$P\ (E1) = \frac{ni}{n} = \frac{Veces\ que\ se\ presenta}{Numero\ total\ de\ pruebas}$$

  Por lo tanto también podemos decir que la probabilidad va a estar siempre entre 0 y 1 (0 <= P (E1) <= 1)

- Subjetiva: Considera la probabilidad como una medida de confianza personal en una situación particular asignando un peso entre 0 y 1 a la ocurrencia de un evento según su grado de creencia de que este sea posible.

Axiomas y propiedades para la familia de eventos (F)

- Axioma F1: Si "X" es el evento imposible y "Z" el evento cierto, ambos pertenecen a la familia de eventos (F).
- Axioma F2: Dado un conjunto numerable de eventos A1,A2,A3… entonces la intersección entre ellos da como resultado un evento que pertenece a F.
- Axioma F3: Si A es un evento y ~A es su complemento, entonces ~A es un evento y pertenece a F
- Axioma F4: Si A y B son eventos, entonces la diferencia entre ellos también lo será y pertenecerá a F.

Axioma de los eventos y sus propiedades

- Axioma P1: Si A es un evento entonces este tiene asociada una probabilidad.
- Axioma P2: Conocida como la ley de la no negatividad, la probabilidad de un evento no es negativa P(A) >= 0.
- Axioma P3: La probabilidad de todo espacio muestral es 1.
- Axioma P4: Sea un conjunto de eventos numerables finitos o infinitos que sean mutuamente excluyentes, entonces la probabilidad de la unión de dichos conjuntos es igual a la suma de sus probabilidades.
- Propiedad P5: Si A y B e F / A C B → P(B - A) = P(B) – P(A)
- Propiedad P6: Si A e F → P(~A) = 1 – P(A)
- Propiedad P7: Si X es cualquier espacio muestral y P es cualquier función de probabilidad definida sobre X, entonces P(Ø) = 0 siendo Ø es evento imposible.
- Propiedad P8: La probabilidad de un evento es un número que va desde 0 a 1.
- Propiedad P9: Si A y B e F / A C B → P(A) < P(B)

- Propiedad P10: Sea un conjunto de eventos numerables finitos o infinitos que no sean necesariamente mutuamente excluyentes, entonces la probabilidad de la unión de dichos conjuntos es menor o igual a la suma de sus probabilidades.

Cálculo de probabilidades

Probabilidad Total (∪)

Es la probabilidad de la unión de eventos en los siguientes casos:

- Mutuamente Excluyentes: P(A ∪ B) = P(A) + P(B)
- No mutuamente excluyentes: P(A ∪ B) = P(A) + P(B) – P(A ∩ B)

Probabilidad Condicional (A/B)

La probabilidad de que ocurra un hecho A, dado que otro hecho B ha ocurrido la cual se define como:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Cabe aclarar que P (A/B) ≠ P (B/A)

Probabilidad Compuesta (A∩B)

Se presenta cuando se da la ocurrencia de eventos a la vez la cual se define como

$$P(A \cap B) = P(B) * P(A/B)$$

Probabilidad Marginal

Es la suma de las probabilidades conjuntas la cual se define como

$$P(A) = P(A \cap B) + P(A \cap D)$$

Independencia y Dependencia Estadística

La dependencia e independencia se da por las formas en la que se encara un proceso aleatorio de selección de elementos que puede alterar o no el número de componentes del espacio probabilístico. En caso de ser con reposición todos los eventos serian independientes unos con otros, pero si es sin reposición se genera eventos dependientes.

Teorema de Bayes

Es una fórmula que sirve para calcular probabilidades condicionales, estableciendo una relación entre probabilidades condicionales alternas, es decir se sabe que P (A/B) ≠ P (B/A) pero con el teorema de Bayes es posible llegar a una relación entre ambas estableciéndola de la siguiente manera.

$$P(A/B) = \frac{P(A)*P(B/A)}{P(B)}$$

Desarrollo:

Supongamos lo siguiente

- $P(A/B) = \frac{P(A \cap B)}{P(B)}$ ---> $P(A \cap B) = P(A/B) * P(B)$
- $P(B/A) = \frac{P(B \cap A)}{P(A)}$ ---> $P(B \cap A) = P(B/A) * P(A)$

$$P(A \cap B) = P(B \cap A)$$

Por lo tanto, al ser las probabilidades conjuntas iguales podemos decir

$$P(A/B) * P(B) = P(B/A) * P(A)$$

$$P(A/B) = \frac{P(A)*P(B/A)}{P(B)}$$

Hay que tener en cuenta, en el caso de haber varios eventos, que la formula puede ser rescrita como:

1) $$P(A/B) = \frac{P(A)*P(B/A)}{P(B \cap A1)+P(B \cap A2)+P(B \cap Ai)}$$

2) $$P(A/B) = \frac{P(A)*P(B/A)}{P(B/A1)*P(A1)+P(B/A2)*P(A2)+P(B/Ai)*P(Ai)}$$

En el punto 1 podemos decir que toda probabilidad marginal en este caso la de P (B) es la suma de sus probabilidades conjuntas P (B ∩ A) y en el punto 2 haciendo un despeje de la probabilidad condicional P (B / A) podemos obtener que la probabilidad conjunta P (B ∩ A) = P (B / A) * P (A)

Unidad 5: Variables Aleatorias y Distribuciones de Probabilidad

Variable Aleatoria

Función real valorada, definida sobre los eventos elementales de un espacio probabilístico, donde cada evento le corresponde un valor que asume la variable. Esta busca representar el conjunto de valores posibles de un espacio probabilístico mediante un nuevo conjunto numérico para así poder presentar probabilidades de presentación de los valores posibles.

Las podemos clasificar en variables aleatorias discretas o continuas.

Formas de representación

- Distribución de probabilidad (Función de cuantía): Muestra todos los valores posibles que puede asumir una variable aleatoria con sus correspondientes probabilidades**.** Lo que nos conlleva a decir que la función cuantía "P (Y = Xi) = P (Xi)" siempre va a sumir un valor numérico, que va a ser mayor a cero y que la sumatoria de todos ellos es igual a 1.

- Distribución de probabilidad acumulada: Utilizada cuando es necesario saber cuál es la probabilidad de que la variable aleatoria discreta asuma un valor menor o igual a un número determinado. Mediante la fórmula F(X) = P(X ≤ xi) = $\sum_{xi<X} \mathrm{P(Xi)}$

  Características:
  - Es una función monótona creciente
  - Los valores que puede asumir la función están entre 0 y 1
  - La probabilidad de que la variable asuma un valor menor a 0 es igual a 0
  - La probabilidad de que la variable asuma valores menores o iguales a todos los valores posibles va a ser igual a 1
  - Dado dos números enteros positivos A y B tales que A < B, la P(A ≤ X ≤ B) = P(B ≤ X) – P (X < A) = P (B ≤ X) – P(A ≤ A - 1)
  - Dado una variable aleatoria X y un numero entero positivo A, la P(A ≤ X) = 1 – P(X < A) = 1 – P(X ≤ A - 1)

- Función de densidad de probabilidad
  Esta función trabaja con variables aleatorias continuas que van a representar la probabilidad para un intervalo de valores [a, b] y no un valor puntual. La cual es de la siguiente manera

$$P(a \leq X \leq b) = \int_a^b fx * dx$$

O

$$P(x \leq a) = \int_{-\infty}^a fx * dx$$

Características:

- P[x = a] = 0
- $\int_{-\infty}^{\infty} fx * dx = 1$
- $P(\mathrm{a} \leq \mathrm{X} \leq \mathrm{b}) = F(b) - F(a)$
- $P(x \geq a) = 1 - F(a)$

Parámetros en la distribución de probabilidad

Estos son calculados en base a los valores posibles que una variable puede asumir y se utilizan para caracterizar un fenómeno

Esperanza Matemática: Es la media aritmética de las variables aleatorias, es decir es la suma de los productos de todos los posibles valores de la variable por sus respectivas probabilidades representada por la letra (Mu) o E (variable aleatoria)

Formula	
V. aleatoria Discreta	$E(x) = \sum_{i=1}^{N} xi * P(xi)$
V. aleatoria Continua	$E(x) = \int_{-\infty}^{\infty} x * f(x) * dx$

Propiedades

1) Dada una variable aleatoria X, una función de ella G(x) es también una variable aleatoria. Es decir en discreta y continua seria

$$E[G(x)] = \sum_{x} G(x) * f(x)$$

$$E[G(x)] = \int_{-\infty}^{\infty} G(x) * f(x) * dx$$

2) La esperanza matemática de una constante es la constante misma
3) E(c*X) = c * E(X)
4) E(c ± X) = c ± E(X)
5) E(X ± Y) = E(X) ± E(Y)
6) E(X * Y) = E(X) * E(Y)

Varianza: Permite conocer la concentración de los resultados alrededor de la esperanza

Formula	
V. aleatoria Discreta	$V(x) = \sum_{i=1}^{N} xi^2 * P(xi) - (E(x))^2$
V. aleatoria Continua	$E(x) = \int_{-\infty}^{\infty} x^2 * f(x) * dx - (E(x))^2$

Propiedades

1) V(X) ≥ 0
2) V(c) = 0
3) $V(c * X) = c^2 * V(X)$
4) V(c ± X) = V(X)
5) V(X ± Y) = V(X) ± V(Y) Solo cuando X e Y sean independientes

Desviación Estándar: Es la raíz cuadrada de la varianza al igual que en las medidas de dispersión.

Momento en las distribuciones de probabilidades

Los "momentos" reflejan el grado de asimetría y el grado de agudeza de una función de probabilidad con respecto a la media o esperanza. Estos se dividen en dos que son "Momentos naturales de orden k" y "Momentos centrado de orden k"

Momentos naturales de orden K: Tienen como objetivo calcular la esperanza de las variables aleatorias

Momentos Naturales de orden K	
V. aleatoria Discreta	$E(x^k) = \sum_{i=1}^{N} x_i^K * P(xi)$
V. aleatoria Continua	$E(x^k) = \int_{-\infty}^{\infty} x^k * f(x) * dx$

Momentos centrados de orden k: Cuando tienen un orden par miden dispersión y cuando tiene un orden impar miden asimetrías.

Momentos centrados de orden K	
V. aleatoria Discreta	$m_k = E(x^k) = [E(xi - E(X))]^K$
V. aleatoria Continua	$m_k = E(x^k) = \int_{-\infty}^{\infty} [xi - E(X)]^k * f(x) * dx$

Para el caso de la variable discreta según qué valor tome K representara distintas cosas

- K = 2: Representa la varianza
- K = 3: Utilizada para determinar si una distribución es simétrica o asimétrica. Entonces:
    - $E(x^3) = 0$: Simetría alrededor de la esperanza
    - $E(x^3) > 0$: Asimétrica hacia la derecha
    - $E(x^3) < 0$: Asimétrica hacia la izquierda

- K = 4: Representa la curtosis, agudeza o puntiagudez. Entonces:

    - $E(x^4) = 0$: Mesocurtica
    - $E(x^4) > 0$: Leptocurtica
    - $E(x^4) < 0$: Platicurtica.

Unidad 6: Modelos especiales de probabilidad de variables aleatorias discretas

Los modelos probabilísticos nos permite predecir la conducta de futuras repeticiones de un experimento.

- Modelo bipuntual o de Bernoulli: Modelo aplicable a variables aleatorias que pueden asumir dos valores (Éxito o fallo) los cuales tienen una de ellas la probabilidad "P" que es la probabilidad de éxito y la otra "Q" que es la probabilidad de fracaso que es igual a 1-P dado que esta es el complemento de "P".

Función de cuantía

$$P(X = xi) = P^X * (1 - P)^{1-X}$$

Si hacemos valor 0 a X obtendremos que "Q" y si hacemos valer 1 a X obtendremos "P".

Parámetros	
Esperanza	$E(x) = \sum_{x=0}^{1} x * P(x)$
Varianza	$V(x) = \sum_{x=0}^{1} x^2 * P(x) - (E(x))^2$
Desviación estándar	Ds = $\sqrt{PQ}$

Demostraciones:
La media es una variable aleatoria bipuntual es igual a la probabilidad de éxito

$E(x) = \sum_{i=0}^{1} x * P(x)$ = 0*Q + 1*P = P

La varianza es igual a la probabilidad de un éxito multiplicada por la de un fallo

$$V(x) = \sum_{x=0}^{1} x^2 * P(xi) - (E(x))^2 = (0^2*Q+1^2*P) - P^2 = P - P^2 = P*Q$$

Este modelo no posee tabla alguna ya que la prueba es realizada una sola vez.

- Modelo Binomial: En caso de tener n pruebas todas realizadas bajo las mismas condiciones e independientes entre sí y cada uno de ellas adopta el modelo bipuntual es decir que en cada prueba solo se presentan dos valores posibles mutuamente excluyentes y colectivamente exhaustivos podemos hacer uso del modelo binomial que será la distribución del numero de Éxitos X.

$$\boldsymbol{x \approx B(n, P)}$$

Además todas las pruebas deben tener igual probabilidad de éxito y fallo, es decir permanecer constantes. Para ello es necesario tener en cuenta el tipo de muestreo a realizar y el tamaño de la población

Población infinita: En cualquier tipo de muestre la probabilidad se va a mantener constante

Población finita: En caso de ser un muestreo con reposición la probabilidad se mantendrá constante, pero en caso contrario la muestra deberá ser menor al 5% de la población para así hacer imperceptible la variación.

Funciones	
Función de Cuantía	P(x = xi, n, P) = $C_n^x * P^x * Q^{n-x}$
Función Acumulada	P(x ≤ xi, n, P) = $\sum_{x=0}^{xi} C_n^x * P^x * Q^{n-x}$

Parámetros	
Esperanza	E(x) = n*P
Varianza	V(x) = n * P * Q
Desviación estándar	Ds(X) = $\sqrt{n * P * Q}$

- Modelo Hipergeometrico: En caso de que se nos encontremos con un muestreo sin reposición de una población finita siendo n > 5 % habrá que hacer uso de este modelo ya que encontraremos dependencia estadística entre las pruebas.

Funciones	
Función de cuantía	$\boldsymbol{P(X = x, N, X, n) = \frac{C_X^x * C_{N-X}^{n-x}}{C_N^n}}$
Función acumulada	$\boldsymbol{P(X \leq x, N, X, n) = \sum \frac{C_X^x * C_{N-X}^{n-x}}{C_N^n}}$

Este modelo hace uso de la estadística clásica la cual se caracteriza por obtener las probabilidades a razón de los casos favorables y lo casos posibles (C.F/C.P).
Los casos favorables en este caso se pueden escribir de esta manera

$$C_X^x * C_{N-X}^{n-x}$$

Los casos posibles se pueden escribir de esta manera

$$C_N^n$$

Llegando a conformar las funciones utilizadas para este método.

Parámetros	
Esperanza	**E(x) = n*P** siendo P = X/N
Varianza	**V(x) = n*P*Q** siendo Q = (1-(X/N))
Desviación estándar	**Ds(x) =** $\boldsymbol{\sqrt{n * P * Q}}$

Cabe aclarar de que cada uno de estos parámetros puede ser afectado por el **factor de corrección de poblaciones finitas** siendo este $\frac{N-n}{N-1}$, que cuando N -> ∞ tiende a ser 1.

- Modelo de proporción: Este modelo permite expresar la proporción de éxitos en vez de un número de éxitos variando su cálculo dependiendo del comportamiento de la variable aleatoria, es decir sea Binomial o Hipergeometrica.
  El cálculo proporcional variara en cualquiera de los dos casos siendo para la binomial **p(con sombrero) = x/n** y para la hipergeometrica **P = x/n**.

  Las funciones de cuantía y acumulación también están sujetas a su modelo, siendo las funciones de este modelo iguales.

  Parámetros para el caso Binomial

Esperanza

$E\ (P_{sombrero}) = E\ (\frac{x}{n}) = \frac{1}{n} * E(x) = \frac{1}{n} * n * P = \boldsymbol{P}$

Varianza

$V(P_{sombrero}) = V\left(\frac{x}{n}\right) = \frac{1}{n^2} * V(x) = \frac{1}{n^2} * n * P * Q = \frac{\boldsymbol{P}*(\boldsymbol{1}-\boldsymbol{P})}{\boldsymbol{n}}$

Desviación Estándar

$Ds(P_{sombrero}) = \sqrt{\frac{\boldsymbol{P}*(\boldsymbol{1}-\boldsymbol{P})}{\boldsymbol{n}}}$

Parametros para el caso HiperGeometrico

Esperanza
E ($P_{sombrero}$)= $E\left(\frac{x}{n}\right) = \frac{1}{n} * E(x) = \frac{1}{n} * n * P = \boldsymbol{P}$

Varianza con factor de corrección
V($P_{sombrero}$)= $V\left(\frac{x}{n}\right) = \frac{1}{n^2} * V(x) = \frac{1}{n^2} * n * P * Q = \frac{\boldsymbol{P}*(\boldsymbol{1}-\boldsymbol{P})}{\boldsymbol{n}} * \frac{\boldsymbol{N}-\boldsymbol{n}}{\boldsymbol{N}-\boldsymbol{1}}$

Desviación Estándar con factor de corrección
Ds($P_{sombrero}$) = $\sqrt{\frac{\boldsymbol{P}*(\boldsymbol{1}-\boldsymbol{P})}{\boldsymbol{n}} * \frac{\boldsymbol{N}-\boldsymbol{n}}{\boldsymbol{N}-\boldsymbol{1}}}$

Recordemos que el factor de corrección es agregado por el hecho de que la muestra pertenece a una población finita

- Modelo de Poisson: Modelo que proporciona la probabilidad de que ocurra un determinado número de veces “n” en un intervalo de tiempo, longitud, area, etc.

Funciones	
Función de cuantía	P(x=xi,μ = λ) = $\frac{e^{-\lambda} * \lambda^{x}}{x!}$
Función acumulada	P(x≤xi,μ = λ) = $\sum_{X=0}^{Xi} \frac{e^{-\lambda} * \lambda^{x}}{x!}$

Siendo λ = n * P número promedio de éxitos en cierto tiempo y espacio. Esta conjuntamente con la variable x designa al número de éxitos en una muestra de tamaño n.

Parámetros	
Esperanza	E(x) = n*P = λ
Varianza	V(x) = n*P = λ
Desviación estándar	Ds(x) = $\sqrt{n * P}$ = λ

Si realizamos un grafico de simetría la variable nos presentaría generalmente asimetría positiva, pero si el tamaño de la muestra aumenta esta tiende a la Simetría.

- Modelo uniforme discreto: Aplicable a un experimento con N resultados mutuamente excluyentes e igualmente probables. Es decir la probabilidad es una constante para todos los resultados de la prueba. Dando así lugar a su función de cuantia como

P(x = xi) = 1/N

Unidad 7: Modelos especiales de probabilidad de variables aleatorias continuas

Los modelos probabilísticos nos permite predecir la conducta de futuras repeticiones de un experimento.

- Modelo Uniforme Continuo: Una variable aleatoria continua cuyo valor solo puede encontrarse en un intervalo (a, b) tiene una distribución uniforme si su función de densidad es constante en dicho intervalo

Funciones	
Función de densidad	f(x) = $\frac{1}{b-a}$ $a \leq x \leq b$
Función acumulada	f(x) = $\frac{x-a}{b-a}$ $a \leq x \leq b$

Parámetros	
Esperanza	E(x) = $\frac{a+b}{2}$
Varianza	V(x) = $\frac{(b-a)^2}{12}$
Desviación estándar	Ds(x) = $\sqrt{\frac{(b-a)^2}{12}}$

Desarrollo de la función densidad

Supongamos una constante k y el intervalo (a,b).

Se sabe que una función de densidad f(x) es igual a 1 cuando tenemos la integral desde -∞ a ∞, es decir.

$$\int_{-\infty}^{\infty} f(x) * d(x) = 1$$

Ahora integramos a la constante k entre los limites (a,b)

$$\int_{a}^{b} k * dx = k * (b - a)$$

Dividimos ambos miembros por "**k * (b - a)"**

$$\int_{a}^{b} \frac{k}{k * (b - a)} * dx = \frac{k * (b - a)}{k * (b - a)}$$

Dando por resultado

$$\int_{a}^{b} \frac{1}{(b - a)} * dx = 1$$

Lo cual indica que la función" f(x) = $\frac{1}{(b-a)}$ "es una función de densidad constante.

Desarrollo de la función de acumulación

$$\int_a^x \frac{1}{(b-a)} * dx = \frac{x-a}{b-a}$$

La cual dará 0 cuando x ≤ a y 1 cuando x ≥ b

- Modelo exponencial: Modelo que busca responder el tiempo o espacio que puede haber entre la ocurrencia entre hechos o el tiempo o espacio que es necesario que transcurra para que se presente un hecho.
  Se suele decir que el modelo exponencial es el reverso del modelo de poisson ya que este se encarga del número de éxitos la exponencial busca el tiempo entre esos éxitos o el tiempo hasta que suceda uno de ellos.

Funciones	
Función de densidad	t(x) = $\lambda - e^{-\lambda x}$
Función acumulada	t(x) = P(x ≤ xi) = $1 - e^{-\lambda x}$

Parámetros	
Esperanza	E(x) = β = $\frac{1}{\lambda}$
Varianza	V(x) = β² = $\frac{1}{\lambda^2}$
Desviación estándar	Ds(x) = β = $\frac{1}{\lambda}$

- Modelo Normal: Teoría que sirve para explicar algunas variables aleatorias la relación entre los intervalos de sus valores y sus respectivas probabilidades la cual es muy útil para aproximar los modelos binomial, poisson, hipergeometrica cuando el tamaño de la muestra es grande.

  Característica

  El comportamiento muestral de un estadígrafo es independiente de cómo se comporte la población, para un n > 30.

- Modelo Normal General: Sea una variable aleatoria continua x diremos que tiene una distribución aproximadamente normal con media μ y desviación estándar σ, es decir X ~ N(μ,σ) si su función de densidad es

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  A tener en cuenta:

  - Toda el área bajo la curva es igual a 1
  - La curva es asintótica con el eje de las x
  - El área determinada por un intervalo debajo de la curva y por encima de las x es la probabilidad del mismo.
  - El dominio de f(x) es infinito
  - f(x) solo depende de μ y σ
  - μ Es el factor de traslación que mueve la grafica
  - σ Hace que sea más o menos aplanada la función

  La función de acumulación es:

$$F(x) = \int_{-\infty}^{x} f(x).dx$$

  Propiedades:

- Es simetrica con respecto al valor medio
- Es de forma acampanada
- Es de transformación lineal de escala
- Se puede realizar combinaciones lineales de variables aleatorias

- Modelo Normal Estándar: Sea una variable aleatoria continua x la cual es estandarizada a una nueva variable Z, se dice que Z ~ N(0,1) si su función de densidad es

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{-z^2}{2}}$$

La estandarización de la variable se realiza mediante lo siguiente

$$Z = \frac{x - \mu}{\sigma} = \frac{1}{\sigma} x - \frac{\mu}{\sigma}$$

Demostración de E (Z) = 0

$$E\left(\frac{1}{\sigma} x - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma} * \mu - \frac{\mu}{\sigma} = 0$$

Demostración de V (Z) = 1

$$V\left(\frac{1}{\sigma} x - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2} V(x) - 0 = \frac{1}{\sigma^2} * \sigma^2 = 1$$

Su función de acumulación es:

$$F(z) = \int_{-\infty}^{Z} f(z) * dz$$

- Regla Empirica de la varianza: Esta regla sirve para darnos la probabilidad que hay de obtener un valor x de una variable aleatoria continua en base a la media y la desviación para el caso de modelos normales.

    - Entre la media y la desviación estándar siempre va a ver la probabilidad del 68%.
    - Entre la media y dos veces la desviación estándar la probabilidad va a ser del 95 %.
    - Entre la media y tres veces la desviación estándar la probabilidad va a ser del 99%.

- Aproximación del Modelo Binomial al Modelo Normal: Supongamos que el tamaño de la muestra n aumenta de manera considerable (n ≥ 30) por lo cual si viésemos su función notaríamos que tiende a ser simétrica independientemente del valor de éxito (P) y del valor de fracaso (Q). En este caso se puede definir a la variable aleatoria discreta x que se comporta como binomial como una variable aleatoria continua z estandarizándola de la siguiente manera

$$Z = \frac{x - n*P}{\sqrt{n*P*Q}} \sim N(0,1)$$

Siendo en este caso **μ = n*p** y **σ =** $\sqrt{\boldsymbol{n} * \boldsymbol{P} * \boldsymbol{Q}}$

- Aproximación del modelo de Poisson al modelo Normal: Siguiendo la misma lógica que la de la aproximación anterior, haremos uso de el modelo de poisson cuando λ ≥ 21 estandarizándola de la siguiente manera

$$Z = \frac{x - n * P}{\sqrt{n * P}}$$

Siendo en este caso **μ = n*p** y **σ =**$\sqrt{\boldsymbol{n} * \boldsymbol{P}}$

- Aproximación del modelo Hipergeometrico al Modelo Normal:

$$\boldsymbol{Z} = \frac{\boldsymbol{x} - \boldsymbol{n}\frac{\boldsymbol{X}}{\boldsymbol{N}}}{\sqrt{\boldsymbol{n}\frac{\boldsymbol{X}}{\boldsymbol{N}} * \left(\boldsymbol{1} - \frac{\boldsymbol{X}}{\boldsymbol{N}}\right) * \frac{\boldsymbol{N} - \boldsymbol{n}}{\boldsymbol{N} - \boldsymbol{1}}}}$$

- Aproximación del Modelo Proporcional al Modelo Normal:

$$\boldsymbol{Z} = \frac{\boldsymbol{Psomb} - \boldsymbol{P}}{\sqrt{\frac{\boldsymbol{P} * \boldsymbol{Q}}{\boldsymbol{n}} * \frac{\boldsymbol{N} - \boldsymbol{n}}{\boldsymbol{N} - \boldsymbol{1}}}}$$

- Distribuciones de las pequeñas muestras: Suele haber situaciones en las cuales no contamos con una n grande, por lo tanto no podemos asegurar que la distribución se comporta como normal. Para ellos está el uso de las siguientes distribuciones de probabilidades con n pequeña.

- Grados de Libertad: Es el numero de parámetros a estimar representado por la operación n-1.

- Chi-Cuadrado(X^2): Se define como la suma de n variables aleatorias continuas independientes al cuadrado (Z_i^2) en donde cada variable Z_i~N(0,1)

  Es decir $\sum_{i=1}^{n} Z_i^2 = \left(\frac{x - \mu}{\sigma}\right)^2 = X_{(n)}^2$

  Características: Siendo $\emptyset = n - 1$
  - El campo de variación de esta distribución va desde el 0 hasta el + ∞

- La función de Chi cuadrado es positivamente asimétrica
- Si Ø ≥ 30 la función se aproxima a la distribución normal

Parámetros	
Esperanza	E($X^2_{(Ø)}$) = μ
Varianza	V($X^2_{(Ø)}$) = 2Ø
Desviación estándar	Ds(x) = $\sqrt{2Ø}$

- "t" de Student: Esta distribución se define como

$$t_{Ø} = \frac{Z}{\sqrt{\frac{X^2}{Ø}}} = \frac{x-u}{\sigma}$$

Y es aplicada cuando se desconoce la varianza poblacional y el tamaño de la muestra es pequeño

Características:

- La variable t asume valores de -∞ a ∞.
- Es unimodal y simétrica respecto a 0.
- Es más aplanada que la distribución normal, pues su varianza es ligeramente superior a 1.

Parámetros	
Esperanza	E(t) = 0
Varianza	V(t) = $\frac{Ø}{Ø-2}$

Desviación estándar	$Ds(t) = \sqrt{\frac{\emptyset}{\emptyset - 2}}$

Unidad 8: Teoría del Muestreo

La teoría del muestro que corresponde a la inferencia estadística tiene sus bases en dos aspectos

- Estimación de parámetros de población: Lo cual se logra partiendo de estimadores muéstrales y calculando la precisión de la estimación
- Docimasia o Prueba de hipótesis: Se utiliza la muestra para determinar si un supuesto sobre la población es verdadero o falso, midiendo los riesgos de cometer un error.

Razones para el muestreo

- Mayor exactitud: Suele suceder que un muestreo sea más exacto que un censo ya que en el muestreo se observan menores errores de observación y además los errores de estimación pueden ser calculados para así determinar un cierto grado de confianza para su uso.
- Costo: Un muestreo es generalmente menos costoso que un censo.
- Tiempo: Se tarda mucho menos en hacer un muestreo y en tabular sus datos.

Base teórica del muestreo

Los datos estadísticos poseen dos importantes características que son

- Diversidad: Hace referencia a la cantidad de características dentro una muestra que poseen los elementos los cuales las pueden tener en mayor o en menor medida. Aunque las variaciones son limitadas haciendo posible el muestreo de una cantidad pequeña.

- Regularidad o Uniformidad: Es la tendencia de las características conmensurable a concentrarse alrededor de una medida de tendencia central.

A la hora de comenzar a realizar un muestreo es necesario seguir las etapas de investigación científica por lo cual una vez establecido el problema a analizar recurrimos a diseñar las muestras que comprende:

  - Plan de Muestreo
  - Elección del estimador a utilizar

Procedimientos para la selección de Muestras

Para seleccionar una muestra representativa es necesario considerar dos criterios:

- Fiabilidad: Este criterio es dado por la varianza del estadígrafo mientras mayor sea la varianza menos fiable va a ser la muestra.
- Efectividad: Está ligado al costo que implica realiza el muestreo, un diseño de muestreo es más efectivo que otro solo si este resulta más barato pero tiene la misma calidad.

Tipos de procedimientos

Muestreo no probabilístico: Las unidades de la población que integran la muestra son seleccionadas en base al criterio del investigador lo cual implica que no haya aleatoriedad y por ende no se pueda hacer inferencia sobre la población de la cual fue extraída. Dentro del muestreo no probabilístico encontramos:

- Muestro de Criterio: El investigado selecciona las unidades que van a componer la muestra.

- Muestreo de la muestra disponible: La muestra queda constituida por una parte de la población que se encuentra convenientemente

disponible, es decir se compone de la parte que es más fácil de tomar.

- Muestreo por cuotas: Caso particular del muestreo de criterio en el cual el investigador establece pasos explícitos para obtener una muestra que sea similar a la población objetivo, ejerciendo controles sobre algunas características de sus elementos. A partir de un listado de la población se estiman el tamaño de la muestra.

Muestro probabilístico: Las unidades son seleccionadas en base a lo propuesto por la teoría de probabilidades permitiéndonos así conocer el error de muestreo, la evaluación en términos de probabilidad y la precisión del estimador, además este tipo de muestro nos permite realizar inferencias sobre la población dentro del cual podemos encontrar:

- Muestreo Aleatorio simple: Procedimiento en el cual se seleccionan n unidades de una población de tamaño N de tal manera que cada una de las muestras del mismo tamaño tengan las mismas probabilidades de ser seleccionadas. En caso de ser un muestreo con reemplazo la probabilidad de cada muestra será $1/N^n$ y de ser sin reposición la probabilidad de cada muestra será $1/C_N^n$. Este tipo de muestro debe ser usado en poblaciones pequeñas y homogéneas en donde se pueda identificar todos los elementos de la población (Colectivamente exhaustivo)

- Muestreo Aleatorio Estratificado: En casos de la población no ser homogénea en relación a la característica bajo estudio se crean los denominados "estratos" que son subconjuntos de la población que nos pueden brindar información si se quiere con distinta precisión. El procedimiento para realizar este tipo de muestro es:

    - Subdividir la población de cantidad N en estratos, es decir N = $\sum_{i=1}^{N} Ni$

- De cada estrato obtener una muestra aleatoria simple ni, siendo la muestra total la suma de todos ellos, es decir n = $\sum_{i=1}^{n} ni$

- Obtener los estadísticos de interés.

Es posible realizar una estratificación doble, que es cuando a los estratos se los divide en subestratos. Ahora, con respecto al tratamiento de cada estrato hay que tener en cuenta que los elementos dentro de cada uno deben ser los más homogéneos posibles para así conseguir buenos valores, pero los estratos entre sí deben ser lo más heterogéneos posibles.
Con respecto al tamaño de la muestra que se la denomina afijación hay tres formas de obtenerla:

- Afijación Igual: Donde todas las muestras son iguales, este tipo de afijación es aplicable cuando los estratos tienen igual participación en la población y sus desviaciones son parecidas

$$n_i = \frac{n}{r}$$

  Siendo n el tamaño de la muestra deseada y r la cantidad de estratos.

- Afijación Proporcional: Cada "n_i" posee en la muestra la misma proporción que cada "N_i "posee en la población, este tipo de afijación es aplicable cuando los estratos difieren en su participación en la población y sus desviaciones son parecidas.

$$n_i = \left(\frac{N_i}{N}\right) n$$

- Afijación Optima o de Neyman: Este tipo de afijación es la mejor que hay ya que maximiza la precisión del estimador de la muestra utilizando en la ecuación la desviación estándar de cada estrato, este tipo de afijaciones es aplicable cuando los estratos difieren entre sí o son iguales y sus desviaciones son disimiles.

$$n_i = \left(\frac{N_{i*\sigma_i}}{\sum N_{i*\sigma_i}} \right) * n$$

- Muestreo sistemático: Se selecciona un elemento de la población cada k elementos, después de haber ordenado los elementos de la misma de una manera específica. El punto de partida se selecciona al azar de entre los primeros k valores. El valor de k que significa "razón de muestreo" está representado por la siguiente fórmula:

$$k = \frac{N}{n}$$

  Este tipo de muestro es más eficaz que el muestreo aleatorio simple si los elementos en la población se asemejan más entre sí.
  El hecho de necesitar que la población sea ordenada en base a un criterio presenta una desventaja si la población es demasiada extensa.

- Muestreo por Conglomerados: Consiste en tomar pequeños grupos de la población que se denominan conglomerados, dentro de los cuales debe haber heterogeneidad pero entre conglomerados debe haber homogeneidad. Una vez realizado los mismos se toma de cada uno de ellos elementos al azar para así constituir una muestra global. La toma de elementos para la muestra se divide en rondas, si hubo varias rondas se dice que el muestreo es de etapas múltiples.

La ventaja es la reducción de costos a la hora de hacer el muestreo, el problema es que el grado de la varianza es muy alto lo cual se puede compensar aumentando el tamaño de la muestra.

Distribuciones en el muestreo

Antes que nada hay que recordar que los muestreos pueden ser con reposición o sin teniendo cada uno sus respectivas características:

Muestro con reposición	
Cantidad de elementos en el espacio probabilístico	N^n
Probabilidad de cada evento elemental	$1/N^n$

Muestreo sin reposición	
Cantidad de elementos en el espacio probabilístico	C_N^n
Probabilidad de cada evento elemental	$1/C_N^n$

Una vez aclarado eso podemos decir que hay tres tipos de distribuciones que son:

- Media Muestral(X con rayita arriba): Es el promedio de observaciones

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

La cantidad de medias muéstrales será la cantidad total del espacio probabilístico que se ve afectado por el tipo de muestreo realizado. Tengamos en cuenta que la media muestral es una constante para

una muestra, pero en el conjunto de muestras posibles la media muestral se vuelve una variable aleatoria pues no se sabe que muestra ha de presentarse.

El símbolo de la media muestral es una x con una raya arriba, por comodidad la nombro xr

Al ser la media muestral una variable aleatoria podemos calcular su esperanza, varianza y desviación estándar.

- Esperanza de la media muestral: Esta medida debe ser igual a la media poblacional tanto como si el muestreo se realizo con reposición o sin reposición. Su fórmula es:

$$E(xr) = \sum_{x=0}^{n} xr_i * P(xr_i)$$

O en caso de contar una tabla de distribución de frecuencia absoluta

$$E(xr) = \frac{\sum_{i=1}^{n} xr_i * n_i}{N^n \ o \ C_N^n}$$

Siendo n_i la frecuencia absoluta de cada media muestral y el denominador dependerá de si el muestreo es con reposición o sin reposición.

- Varianza de la media muestral: Esta es proporcional a la varianza poblacional y mientras mayor sea la muestra menos variabilidad tendrá la media muestral

$$\sigma^2(xr) = \sum_{x=0}^{n} {xr_i}^2 * P(xr_i) - [E(xr)]^2$$

O en caso de contar una tabla de distribución de frecuencia absoluta

$$\sigma^2(xr) = \frac{\sum_{i=1}^{n} {xr_i}^2 * n_i}{N^n \; o \; C_N^n} - [E(xr)]^2$$

Siendo n_i la frecuencia absoluta de cada media muestral y el denominador dependerá de si el muestreo es con reposición o sin reposición.

La relación que guarda la varianza de la media muestral con la varianza población es la siguiente:

Muestreo con reposición	$\sigma^2(xr) = \frac{\sigma^2(X)}{n}$
Muestreo sin reposición	$\sigma^2(xr) = \frac{\sigma^2(X)}{n} * \frac{N-n}{N-1}$

Siendo $\sigma^2(X)$ la varianza población y $\frac{N-n}{N-1}$ el factor de correcciones finitas

- Desviación estándar: Se calcula como la raíz cuadrada de la varianza

$$\sigma(xr) = \sqrt{\sigma^2(xr)}$$

Recordemos que la varianza está sujeta a si el muestreo es con reposición o sin reposición.
La relación que guarda la desviación estándar de la media muestral con la varianza población es la siguiente:

Muestreo con reposición	$\sigma(xr) = \frac{\sigma(X)}{\sqrt{n}}$
Muestreo sin reposición	$\sigma(xr) = \frac{\sigma(X)}{\sqrt{n}} * \frac{N-n}{N-1}$

Cabe aclarar que el factor de corrección de poblaciones finitas es justamente para poblaciones finitas, en caso de ser infinita la población no es necesario aplicarlo por más que el muestreo haya sido realizado sin reponer

- Proporción Muestral ($P_{sombrero}$ o p): Esta se define como la razón de la cantidad de éxitos en n pruebas sobre el tamaño de la muestra.
  Es decir: $\boxed{p = x/n}$
  La proporción muestral también va a ser una variable aleatoria y por ende podemos calcular su esperanza, varianza y desviación.

  - Esperanza de la proporción muestral: Siempre es igual a la proporción poblacional independientemente de si el muestreo fue con o sin reposición. Se define como

$$E(p) = \sum_{i=1}^{n} p_i * P(p_i)$$

O en caso de contar una tabla de distribución de frecuencia absoluta

$$E(p) = \frac{\sum_{i=1}^{n} p_i * n_i}{N^n \ o \ C_N^n}$$

Siendo n_i la frecuencia absoluta de cada media muestral y el denominador dependerá de si el muestreo es con reposición o sin reposición.

- Varianza de la proporción muestral: Se define como

$$\sigma^2(p) = \sum_{i=1}^{n} {p_i}^2 * P(p_i) - [E(p)]^2$$

O en caso de contar una tabla de distribución de frecuencia absoluta

$$E(p) = \frac{\sum_{i=1}^{n} {p_i}^2 * n_i}{N^n \ o \ C_N^n} - [E(p)]^2$$

Siendo n_i la frecuencia absoluta de cada media muestral y el denominador dependerá de si el muestreo es con reposición o sin reposición.

Con respecto a las relaciones que este guarda con la varianza poblacional es

Muestreo con reposición	$\sigma^2(p) = \frac{P(1-P)}{n}$
Muestreo sin reposición	$\sigma^2(p) = \frac{P(1-P)}{n} * \frac{N-n}{N-1}$

Siendo P la proporción poblacional.

- Desviación estándar de la proporción Muestral: Es la raíz cuadrada de la varianza de "p" es decir

$$\sigma(p) = \sqrt{\sigma^2(p)}$$

Con respecto a las relaciones que este guarda con la desviación estándar poblacional es

Muestreo con reposición	$\sigma(p) = \sqrt{\frac{P(1-P)}{n}}$

Muestreo sin reposición	$\sigma(p) = \sqrt{\frac{\mathrm{P}(1-\mathrm{P})}{n}} * \frac{N-n}{N-1}$

El factor de corrección para poblaciones finitas recibe el mismo trato que en la media muestral.

- Varianza muestral corregida($s^2_{sombrero}$): Indicara como toda varianza el grado de dispersión o concentración de las observaciones muéstrales respecto a su valor central y se define como

$$s^2{}_{sombrero} = \frac{\sum_{i=1}^{n}(x - xr)^2}{n-1}$$

La relación que establece la varianza muestral con la varianza poblacion es

$$s^2{}_{sombrero} = S^2 \frac{n}{n-1}$$

Se puede calcular la desviación estándar muestral también siendo esta la raíz cuadrada de la varianza muestral corregida

$$s_{sombrero} = \sqrt{s^2{}_{sombrero}}$$

Esta al igual que sus predecesoras es también una variable aleatoria con la diferencia de que solo podemos calcular la esperanza

  - Esperanza: Siempre es igual a la varianza poblacional independientemente de si el muestreo se realizo con o sin reposición. Se define como

$$E(s^2{}_{sombrero}) = \sum_{i=1}^{n} s^2{}_{sombrero.i} * P(s^2{}_{sombrero.i})$$

O en caso de contar una tabla de distribución de frecuencia absoluta

$$E(s^2{}_{sombrero}) = \frac{\sum_{i=1}^{n} s^2{}_{sombrero.i} * n_i}{N^n \; o \; C_N^n}$$

- Teorema central de limite: Cualquiera sea la distribución de la población en la medida que posea una varianza finita, la variable aleatoria Z = $\frac{xr-\mu}{\sqrt{xr}}$ ~N(0,1) a medida que n crezca indefinidamente.
  En general la normalidad de una distribución para la media muestral es el denominado teorema central del límite y establece que:

  - Cuando la población es grande y está distribuida normalmente la media muestral también será normal.
  - Cuando la población no está distribuida normalmente, la distribución de medias muéstrales se aproximara a una distribución normal si la muestra n es grande (n ≥ 30).
  - Si la distribución normal de las medias muestrales tienen la media igual al valor esperado de la muestra E(xr) y la desviación estándar entonces teóricamente estos pueden ser calculados por la media poblacional (μ = E(xr)) y la desviación estándar poblacional ($\sigma(X) = \frac{\sigma(xr)}{\sqrt{n}}$)

Unidad 9: Estimación Estadística

Como se vio en la unidad anterior a veces no se puede trabajar con la población por x limitaciones por lo cual se hace necesario el uso de estimadores. El método para obtener el valor o mejor dicho hacer la estimación recibe el nombre de estimador mientras que el resultado obtenido se llama estimación del parámetro. A partir de estos conceptos llamaremos entonces

Ø Al parámetro a estimar

$Ø_{sombrero}$ Al estimador que por razones cómodas solo para este resumen se llamara Øs el cual será función de las observaciones muéstrales y por esto implica que va a ser una variable aleatoria. Por otro lado un estimador es un estimador puntual de una característica de la población si proporciona un solo número como estimación y por otro lado está la estimación por intervalos en donde sitúa al parámetro entre dos valores para una confianza dada

Las propiedades de los buenos estimadores puntuales son:

- Insesgabilidad: Es cuando la esperanza del estimador es igual al parámetro a estimar, es decir :

  $$E(Øs) = Ø.$$

  Si no son iguales ambos valores la diferencia entre ellos se denomina sesgo dando pie al sesgo positivo $E(Øs) > Ø$ y al negativo $E(Øs) < Ø$. Un caso particular de esta es cuando la mediana muestral es igual a la media poblacional, eso solo sucede cuando la distribución de la población es normal

- Eficiencia: Esta propiedad se refiere a la variabilidad de los estimadores, dado por la varianza que nos da un grado de confianza sobre los mismos. Es decir supongamos que tenemos dos estimadores Ø1 y Ø2. Independientemente de si uno presenta un sesgo o no siempre se preferirá el que tenga menos varianza. Hay un método para elegir estimadores llamado Eficiencia relativa que es la razón entre la varianza de ambos estimadores

- Consistencia: Un estimador es consistente si a medida que el tamaño de la muestra se aproxima al tamaño de la población (n → N) si el estimador se aproxima al parámetro a estimar (Øs → Ø).

- Suficiencia: Es un estimador que utiliza toda la información posible de la muestra

La estimación puntual tiene el problema de que no puede otorgar precisión de la estimación obtenida como así puede darla la estimación por intervalos

Error, riesgo y tamaño de la muestra

Supongamos un parámetro a estimar Ø con un estimador Øs, el cual al tomar una muestra asume un valor particular Øs0 dando pie a la estimación puntal que se utilizara para estimar a Ø. Entonces el error de muestreo "e" es la diferencia entre (Øs0 – Ø)

$$e = (Øs0 - Ø)$$

Ahora para medir el riesgo de cometer un error mayor que "e" seria

$$Pr[(Øs0 - Ø) > e]$$

Pero antes hay que normalizar la variable particular, es decir $z = \frac{Øs0 - Ø}{\sigma(Øs)}$

La probabilidad anteriormente mencionada nos da el nivel de confianza con respecto a la distribución de valores con los que estamos trabajando.

- Relación entre error, riesgo y tamaño de la muestra: A la hora de ver las relaciones es necesario tener presenta que siempre se parte

de la función de la tipificación de una variable, es decir z = $\frac{\emptyset s0 - \emptyset}{\sqrt{\sigma(\emptyset s)}}$ y que el riesgo es función inversa de z.

- Riesgo:

$$Z = \frac{e * \sqrt{n}}{\sigma}$$

Si incrementamos el error, Z se incrementa y el riesgo disminuye.
Si incrementamos el tamaño de la muestra, Z se incrementa y el riesgo también disminuye.
Si incrementamos la desviación, Z disminuye y el riesgo aumenta.

- Error: Despejando "e" de la ecuación anterior

$$e = \frac{Z * \sigma}{\sqrt{n}}$$

Mientras mayor sea la desviación, mayor será el error
Si se incrementa n el error será más pequeño
Si incrementamos el riesgo disminuye z y disminuye el error

- Tamaño de la muestra: Despejando "n"

$$n = \frac{z^2 * \sigma^2}{e^2}$$

Si incrementamos el riesgo, z disminuye y por ende n también
Si incrementamos la varianza, n aumenta
Si incrementamos el error menor tamaño de la muestra.

Estimación por intervalos

Consiste en obtener un cierto intervalo aleatorio a partir de la estimación puntal considerando un cierto error en la estimación y un determinado grado de confianza de que el intervalo construido contiene el parámetro que deseamos estimar. Los elementos que intervienen para este tipo de estimación son:

Ø	Parámetro a estimar
Øs	Estimador
K(Ø, Øs)	Distribución de probabilidades tabulada y conocida
n	Observaciones muéstrales
K1,k2	Coeficientes de confianza o puntos críticos
1 - α	Nivel de confianza
L1, L2	Limites del intervalo de confianza (L1 < Ø < L2)

Intervalo de confianza para estimar (μ)

Se hace uso de la distribución normal y la de "t" de studen de la siguiente manera

Poblaciones normales		**Poblaciones no normales**	
σ^2 conocida con un "n" cualquiera	Distribución normal	σ^2 conocida con un n ≥ 30	Por TCL Normal
σ^2 Desconocida		σ^2 Desconocida	
n < 30	"t" de student	n < 30	
n ≥ 30	Por TCL Normal	n ≥ 30	Por TCL Normal

Caso 1: σ^2 conocida con un "n" cualquiera o n ≥ 30

Ø = μ

Øs= xr

K(Øs, Ø) = $\frac{xr - \mu}{\frac{\sigma}{\sqrt{n}}} \sim$ N(0,1)

Con lo anterior procedemos a armar el intervalo de confianza

$$1 - \alpha = \Pr\left[z_{1} \leq \frac{xr - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{2}\right]$$

Despeje μ el nivel de confianza va a quedar como

$$1 - \alpha = \Pr\left[xr - z_{2*}\frac{\sigma}{\sqrt{n}} \leq \mu \leq xr + z_{1*}\frac{\sigma}{\sqrt{n}}\right]$$

Ese enunciado asegura que si se toman muchas muestras de tamaño n, (1 – α) % de ellas serán correctas.

En caso de ser el muestreo realizado sin reposición se tendrá que aplicar el factor de corrección de población finita en cada uno de los errores de estimación "e = $z_{2*}\frac{\sigma}{\sqrt{n}}$"

Caso2: σ^2 desconocida con un n < 30

Ø = μ

Øs = xr

K (Øs, Ø) = $\frac{xr - \mu}{\frac{\sigma}{\sqrt{n}}} \sim t_{n-1}$

Con lo anterior procedemos a armar el nivel de confianza

$$1 - \alpha = \Pr(t1 \leq \frac{xr - \mu}{\frac{\sigma}{\sqrt{n}}} \leq t2)$$

Despeje μ el nivel de confianza va a quedar como

$$1 - \alpha = Pr(xr - t2 * \frac{\sigma}{\sqrt{n}} \leq \mu \leq xr + t1 * \frac{\sigma}{\sqrt{n}})$$

En caso de ser el muestreo realizado sin reposición se tendrá que aplicar el factor de corrección de población finita en cada uno de los errores de estimación "e = $z_{2*} \frac{\sigma}{\sqrt{n}}$"

Determinación del tamaño de la muestra en la estimación de μ

Se parte de la normalización de la variable aleatoria "xr", es decir:

$$Z = \frac{xr - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Despejando n obtenemos que

$$n \geq \frac{z^2 * \sigma^2}{e^2}$$

Como se ve se necesita conocer de ante mano la desviación estándar, el error de muestreo permitid, y el nivel de confianza que se determina a partir de z.

Intervalo de confianza para estimar proporción poblacional "P"

Se usa la distribución normal en donde

Ø = P

Øs = p

K(Ø, Øs) ~ N(0,1)

Establecemos el nivel de confianza

$$1 - \alpha = Pr\left[z1 \leq \frac{p-P}{\sqrt{\frac{p(1-p)}{n}}} \leq z2\right]$$

Despejando P obtenemos que

$$1 \quad - \alpha = Pr\left[p - z2 * \sqrt{\frac{p(1-p)}{n}} \leq P \leq p + z1 * \sqrt{\frac{p(1-p)}{n}}\right]$$

Si multiplicásemos los limites por n obtendríamos convertiríamos la ecuación del nivel de confianza es la estimación del nivel de confianza para μ.

La determinación para el nivel de confianza de la proporción poblacional, requiere que se tome una gran cantidad de elementos para la muestra.

Determinación del tamaño de la muestra en la estimación de P

Partiendo de la normalización de la variable aleatoria p

$$Z = \frac{p - P}{\sqrt{\frac{P(1-P)}{n}}}$$

Despejamos n y obtenemos que:

$$n = \frac{z^2 * P(1-P)}{(p-P)^2}$$

Se suele utilizar a P = 0,50 para así poder dar un tamaño lo mayor posible para la muestra.

Invervalo de confianza para estimar la varianza poblacional

De una poblacion normal con un n < 30 se usa “Chi cuadrado”(x^2) en donde

$\emptyset = \sigma^2$

$\emptyset s = Ssomb^2(Ss^2)$

K(Øs, Ø) = $\frac{Ss^2(n-1)}{\sigma^2}$ ~ Chi cuadrado (n-1)

Armamos el intervalo de confianza

$$1\text{-}\alpha = Pr[x1^2 \leq \frac{Ss^2(n-1)}{\sigma^2} \leq x2^2]$$

Despejando la varianza poblacional obtendremos

$$1 - \alpha = Pr[\frac{Ss^2(n-1)}{x2^2} \leq \sigma^2 \leq \frac{Ss^2(n-1)}{x1^2}]$$

Unidad 10: Contraste, Docima o Verificación de hipótesis.

También llamada test de hipótesis, es un procedimiento de la inferencia estadística para la toma de decisiones.

Decisión Estadística

Es una decisión que se toma respecto a la población en base a evidencias proporcionadas por la muestra. Para ello se forman conjeturas sobre algunas características de una distribución población llamada *Hipótesis Estadística.*

Hipótesis Estadística

Suposiciones que se establecen con respecto al valor de un parámetro en base a valores dados por la muestra, estas pueden ser:

- Hipótesis Nula (H_0): Indica que no se presentan cambios en la población.

- Hipótesis Alternativa (H_1): Indica cambios o efectos que se presentaron en la población.

Concepto de Docima

Experimento aleatorio que se realizar para decidir sobra la veracidad sobre la hipótesis nula, es decir decidir si lo que establece la hipótesis nula es cierto o falso dando paso al rechazo o no de la misma. Se maneja de la siguiente manera

$H_0 : \emptyset = \emptyset_0$

$H_1 : \emptyset \neq \emptyset_0$

Esto va de la mano con una probabilidad "α" llamado nivel de significancia que representa la probabilidad de rechazar la hipótesis nula cuando esta es verdadera, es decir α = Pr [$H_0\ sea\ rechazada/\ H_0\ sea\ cierta$]. Todo esto se realiza mediante estimaciones puntuales dentro de un intervalo de confianza con sus respectivos puntos críticos.

Errores y sus probabilidades

A la hora de rechazar o aceptar una hipótesis nula se puede cometer los siguientes errores:

- Error tipo 1: $H_0\ sea\ rechazada/\ H_0\ sea\ cierta$ y la probabilidad de este es igual a α.
- Error tipo 2: $H_0\ no\ se\ rechaze/\ H_0\ sea\ falsa$ y la probabilidad de este es igual β.

Tanto α y β varían en sentido inverso y no se puede disminuir una sin aumentar la otra. La única manera de hacer disminuir las probabilidades de ambos a la vez es aumentar el tamaño de la muestra.

Análogamente "1 – α" representaría la zona de aceptación para la hipótesis nula y"1 – β" representaría, por así decirlo, la decisión correcta en donde se rechaza la hipótesis nula siendo esta falsa.

Tipos de docimas

Dado un parámetro población X y un valor particular X_0 en el cual se verifica la hipótesis nula, es decir H_0 : X = X_0, se pueden plantear

- Docimas de hipótesis compuestas o inexactas: Es cuando se contrasta un valor contra un conjunto de valores, esta se divide en Docimas Bilaterales y Docimas laterales izquierdas o derecha.

- Docimas de hipótesis simples o exactas: Cuando se contrasta un valor para la hipótesis nula contra un valor para la hipótesis alterna, esta se divida en Docimas laterales izquierdas o derechas.

Docimas bilaterales (≠): Las hipótesis son puestas de la siguiente manera

H_0 : Ø = $Ø_0$

H_1 : Ø ≠ $Ø_0$

Considerando un estimador insesgado con distribución normal, podemos decir que:

1 – α = Pr[Øs1 ≤ Øs ≤ Øs2/H0 sea cierta]

α = Pr[Øs < Øs1] + Pr[Øs > Øs2]

β = Pr[Øs1≤ Øs≤ Øs2/H0 sea falsa]

1 – β = Pr[Øs < Øs1]+ Pr[Øs > Øs2]

Docimas compuetas Laterales (<,>):

- Derecha (>): Las hipótesis son puestas de la siguiente manera

  H_0 : Ø = $Ø_0$

  H_1 : Ø > $Ø_0$

  Considerando ahora solo un punto crítico (Øs*]) podemos decir que:
  1-α = Pr[Øs ≤ Øs*/H0 es cierta]

α = Pr[Øs > Øs*]
β = Pr[Øs ≤ Øs*/H0 sea falsa]
1-β = Pr[Øs > Øs*]

- Izquierda:
  H_0 : Ø = $Ø_0$
  H_1 : Ø < $Ø_0$

  Considerado ahora un solo punto crítico (Øs*) podemos decir que:

  1-α = Pr[Øs ≥ Øs*/ H0 sea cierta]
  α = Pr[Øs < Øs*]
  β = Pr[Øs ≥ Øs* /H0 sea falsa]
  1-β= Pr[Øs < Øs*]

Docimas simples laterales: Opera de la misma manera que las Docimas laterales compuestas solo que la forma de expresar las hipótesis es de la siguiente manera

H_0 : Ø = $Ø_0$
H_1 : Ø = $Ø_1$

Siendo $Ø_1$ un valor que se encuentre a la izquierda o derecha de el valor de la hipótesis nula

Docimas para la media poblacional (μ)

Se trabaja con poblaciones normales y no normales en las cuales podremos usar solamente la distribución de t de "student" y la distribución normal.

Pasos para docimar: Los pasos son los mismos que la de la unidad anterior solo que ahora se agregan las hipótesis, es decir

1. Identificar el parámetro a docimar (Ø)
2. Seleccionar un estimador (Øs)

3. Determinar el estadístico K(Øs, Ø)
4. Calcular los puntos críticos K*
5. Plantear Hipótesis
6. Calcular las probabilidades involucradas

Se realiza el desarrollo para todos los tipos de docima.

Caso1: Varianza poblacional conocida con muestra de cualquier tamaño o con n ≥ 30 (Para poblaciones no normales)

D. Simple	Izquierda	Derecha
Parámetro poblacional, su estadístico y si distribución	Ø = μ Øs = xr K(xr,μ)~ N(0,1)	Ø = μ Øs = xr K(xr,μ)~ N(0,1)
Planteo las hipótesis	H_0: μ = μ_0 H_1: μ = μ_1 En donde $\mu_1 < \mu_0$	H_0: μ = μ_0 H_1: μ = μ_1 En donde $\mu_1 > \mu_0$
Calculo de puntos críticos	xr* = μ_0 - $Z_{1-\propto} * \frac{\sigma}{\sqrt{n}}$	xr* = μ_0 + $Z_{1-\propto} * \frac{\sigma}{\sqrt{n}}$
Toma de decisión	xr ≥ xr* Acepto H_0 xr < xr* Rechazo H_0	xr ≤ xr* Acepto H_0 xr > xr* Rechazo H_0

Como se dijo anteriormente hay dos tipos de errores cada uno con su cálculo de probabilidad, para la media poblacional es de la siguiente manera

- Derecha:

$$1-\alpha = Pr\ [xr \leq xr^* / H_0 \text{ sea cierta}]$$
$$\alpha = Pr\ [xr > xr^*]$$

$$\beta = Pr\ [xr \leq xr\# / H_0 \text{ sea falsa}]$$
$$1-\beta = Pr\ [xr > xr\#]$$

- Izquierda:

$$1-\alpha = Pr\ [xr \geq xr^* / H_0 \text{ sea cierta}]$$

$$\alpha = Pr\ [xr < xr^*]$$

$$\beta = Pr\ [xr \geq xr\#/H_0 \text{ sea falsa}]$$

$$1-\beta = Pr\ [xr < xr\#]$$

Siendo xr# = $\frac{xr* - \mu_1}{\sigma/\sqrt{n}}$ para ambos casos.

D. Compuesta	Izquierda	Derecha
Parámetro poblacional, su estadístico y si distribución	Ø = μ Øs = xr K(xr,μ)~ N(0,1)	Ø = μ Øs = xr K(xr,μ)~ N(0,1)
Planteo las hipótesis	H_0: μ = μ_0 H_1: μ < μ_0	H_0: μ = μ_0 H_1: μ > μ_0
Calculo de puntos críticos	xr* = μ_0 - $Z_{1-\propto} * \frac{\sigma}{\sqrt{n}}$	xr* = μ_0 + $Z_{1-\propto} * \frac{\sigma}{\sqrt{n}}$
Toma de decisión	xr ≥ xr* Acepto H_0 xr < xr* Rechazo H_0	xr ≤ xr* Acepto H_0 xr > xr* Rechazo H_0

El cálculo se sus probabilidades es igual al de la Docima simple

Docima compuesta bilateral

1. Parámetro poblacional, estadístico y distribución:
   Ø = μ
   Øs = xr
   K (xr,μ)~ N(0,1)

2. Planteo de hipótesis
   H_0: μ = μ_0
   H_1: μ ≠ μ_0

3. Calculo de puntos críticos

$$xr_1^* = \mu_0 - \ z_{1-\frac{\propto}{2}} * \frac{\sigma}{\sqrt{n}}$$

$$xr_2^* = \mu_0 + \ z_{1-\frac{\propto}{2}} * \frac{\sigma}{\sqrt{n}}$$

Recuerde que al ser bilateral hay un intervalo definido por dos puntos a diferencia de los laterales que se encuentran definido con un punto y ± ∞.

4. Toma de decisiones

$xr_1^* \leq xr \leq xr_2^* \longrightarrow Acepto \ \ H_0$

$xr < xr_1^* \wedge \ xr > xr_2^* \longrightarrow Rechazo \ \ H_0$

5. Cálculo de probabilidades

$$1-\propto = \Pr[xr_1^* \leq xr \leq xr_2^* \ / \ H_0 \ sea \ cierta \ ]$$

$$\propto = \Pr[xr \ < xr_1^*] + \ \Pr[xr > \ xr_2^*]$$

$$\beta = \Pr[[xr_1^* \leq xr \leq xr_2^* \ / \ H_0 \ sea \ falsa]$$

$$1 - \beta = \Pr[xr \ < xr_1^*] + \Pr[xr > \ xr_2^*]$$

Caso2: Varianza poblacional desconocida con un n < 30

Seguimos los mismos pasos anteriores pero la distribución del estadístico

K (xr,μ) = $\frac{xr-\mu}{\frac{\sigma}{\sqrt{n}}} \sim t_{n-1}$

Docima para la proporción poblacional (P)

Se procederá a dar la proporción poblacional en cada uno de los tipos de docima, en donde cada uno de ellos hace uso de la distribución normal y se siguen los pasos que se describieron en la media poblacional.

D. simple	Izquierda	Derecha
Parámetro poblacional, su estadístico y si distribución	Ø = P Øs = p K(p,P)~ N(0,1)	Ø = P Øs = p K(p,P)~ N(0,1)
Planteo las hipótesis	H_0: P = P_0 H_1: P = P_1 en donde $P_1 < P_0$	H_0: P = P_0 H_1: P = P_1 en donde $P_1 > P_0$
Calculo de puntos críticos	p* = p_0 - $Z_{1-\propto} * \sqrt{\frac{P_0(1-P_0)}{n}}$	p* = p_0 + $Z_{1-\propto} * \sqrt{\frac{P_0(1-P_0)}{n}}$
Toma de decisión	p ≥ p* Acepto H_0 p < p* Rechazo H_0	p ≤ p* Acepto H_0 p > p* Rechazo H_0

Cálculo de probabilidades

- <u>Derecha</u>:

$$1 - \alpha = \Pr[p \leq p^* / H_0 \ sea \ cierta]$$

$$\alpha = \Pr[p > p^*]$$

$$\beta = \Pr[p \leq p^{\#} / H_0 \ sea \ falsa]$$

$$1 - \beta = \Pr[p > p^{\#}]$$

- <u>Izquierda</u>:

$$1 - \alpha = \Pr[p \geq p^* / H_0 \ sea \ cierta]$$

$$\alpha = \Pr[p < p^*]$$

$$\beta = \Pr[p \geq p^{\#} / H_0 \; sea \; falsa]$$

$$1 - \beta = \Pr[p < p^{\#}]$$

Siendo $p^{\#} = \frac{p^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}$

D. simple	Izquierda	Derecha
Parámetro poblacional, su estadístico y si distribución	Ø = P Øs = p K(p,P)~ N(0,1)	Ø = P Øs = p K(p,P)~ N(0,1)
Planteo las hipótesis	H_0: P = P_0 H_1: P < P_0	H_0: P = P_0 H_1: P > P_0
Calculo de puntos críticos	p* = p_0 - $Z_{1-\propto} * \sqrt{\frac{P_0(1-P_0)}{n}}$	p* = p_0 + $Z_{1-\propto} * \sqrt{\frac{P_0(1-P_0)}{n}}$
Toma de decisión	p ≥ p* Acepto H_0 p < p* Rechazo H_0	p ≤ p* Acepto H_0 p > p* Rechazo H_0

El cálculo de probabilidades es exactamente igual que el de las Docimas simples

Docima compuesta bilateral

1. Parámetro, estimador y estadístico
   Ø = P
   Øs = p
   K (Øs,Ø) = $\frac{p-P}{\sqrt{\frac{P(1-P)}{n}}}$ ~ N(0,1)

2. Planteo de hipótesis
   H_0: P = P_0
   H_1: P ≠ P_0

3. Calculo de puntos críticos

$$p_1^* = p_0 - \ Z_{1-\frac{\alpha}{2}} * \sqrt{\frac{P_0(1-P_0)}{n}}$$

$$p_1^* = p_0 + \ Z_{1-\frac{\alpha}{2}} * \sqrt{\frac{P_0(1-P_0)}{n}}$$

4. Toma de decisión

   $p_1^* \leq p \leq p_2^* \longrightarrow Acepto \ \ H_0$

   $p < p_1^* \wedge \ xr > p_2^* \longrightarrow Rechazo \ \ H_0$

5. Cálculo de probabilidades

$$1 - \alpha = \Pr[p_1^* \leq p \leq p_2^* \ / \ H_0 \ sea \ cierta]$$

$$\alpha = \Pr[p < p_1^*] + \ \Pr[p > \ p_2^*]$$

$$\beta = \Pr[p_1^{\#} \leq p \leq p_2^{\#} \ / \ H_0 \ sea \ falsa]$$

$$1 - \beta = \Pr\left[p < p_1^{\#}\right] + \ \Pr[p > \ p_2^{\#}]$$

Siendo

$$p_1^{\#} = \frac{p_1^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}$$

$$p_2^{\#} = \frac{p_2^* - P_1}{\sqrt{\frac{P_1(1-P_1)}{n}}}$$

Curva operatoria (OC) y Curva de potencia

Esto va ligado con el concepto de α y β, solo aplicable a Docimas compuestas que como ya se dijo exponen al parámetro a un conjunto de valores teniendo en mente esto definiremos entonces

- Curva potencia: Es la que relaciona todos los valores posibles del parámetro con la probabilidad de rechazar H_0 para un determinado nivel de significación α
- Curva operatoria (OC): Es la que relaciona todos los valores posibles del parámetro con la probabilidad de aceptar H_0 para un determinado nivel de significación α