**PREGUNTAS:**

**UNIDAD 1**

**Los métodos estadísticos son clasificados en cinco pasos básicos:**

**Recopilación**

**Organización y corrección de datos**

**Presentación**

**Análisis o métodos de cálculo**

**Interpretación**

Definiciones importantes:

Tamaño de muestra (n): es la cantidad de elementos en una serie estadística

Máximo (xmáx): se llama máximo de una variable estadística al mayor valor que toma la variable en toda la serie estadística.

Mínimo (xmín): se llama mínimo de una variable estadística al menor valor que toma la variable en toda la serie estadística.

Recorrido (R): es la diferencia entre el máximo y el mínimo en una serie estadística. R = xmáx - xmín

Clase: se llama clase a cada uno de los intervalos en que podemos dividir el recorrido de la variable estadística. Los intervalos pueden o no ser de la misma amplitud.

Límite superior de la clase (Ls): es el máximo valor del intervalo.

Límite inferior de la clase (Li): es el mínimo valor del intervalo.

Marca de clase (xi): es el punto medio de cada clase y es el promedio entre los extremos del intervalo.

Cantidad de intervalos (k): se obtiene a partir de la **fórmula de Sturges**: **K=** **1 + 3,3.log n**. Para tamaños de muestra pequeños también es útil utilizar (n)1/2 (raíz cuadrada de n), aunque la fórmula de Sturges es válida para todos los casos.

Longitud de intervalos (l): es la diferencia entre el límite superior y el límite inferior de la clase. L = R / k

**¿De qué manera se pueden presentar los datos para realizar un estudio estadístico? ¿Qué ventajas y desventajas ofrece cada una de las formas de presentación de datos?**

Buscamos que la audiencia sienta que debe interesarse e informarse, por lo tanto queda a nuestro criterio la manera en que suministramos la información.

Se pueden presentar de 3 maneras:

***Texto:*** Esta forma de presentación permite llamar la atención sobre las comparaciones de importancia y destacar ciertas cifras. **Solo puede utilizarse cuando los datos por presentar son pocos.**

***Cuadros o Tablas y Graficas:*** Este tipo de presentación permite volcar un gran número de datos en forma resumida, lo que hace fácil y clara su lectura.

**¿Cómo se describe gráficamente un conjunto de datos?**

Las representaciones gráficas pueden hacerse utilizando un sistema geométrico de representación, en cuyo caso gozan de rigurosidad y precisión, o bien pueden utilizarse símbolos alusivos al tema en estudio. Existe una gran variedad de gráficos. Su elección depende de las variables en estudio y de las características que se quieren destacar.

Hay que advertir, que la representación gráfica no es más que un medio auxiliar de la investigación estadística, que es fundamentalmente numérica.

**¿Se describen de igual manera los conjuntos de datos cualitativos que cuantitativos?**

**Datos cualitativos:**

Los gráficos más usuales para **variables cualitativas** son los gráficos de barras, que pueden ser verticales u horizontales y los gráficos de sectores.

Gráfico de Pareto: Una variante importante de los diagramas de barras es el diagrama de Pareto. Las categorías están ordenadas de modo tal que en la parte izquierda aparezca la categoría con mayor frecuencia, seguida por la segunda mayor frecuencia y así, sucesivamente.

Gráfico de sectores: Con su respectiva frecuencia

Histograma y polígono de frecuencias

**Datos cuantitativos:** Hay distintas maneras de presentar los datos cuando no han sido agrupados en intervalos:

Gráfico de tronco y hojas: Observamos que la parte entera de los números son: 0, 1, 2 y 3. Esto nos permite dividir cada número en tronco (la parte entera) y hojas (la parte decimal).

Gráficos de puntos o puntigramas: Datos con sus respectivas frecuencias.

Histograma y polígono de frecuencias: Un histograma se obtiene construyendo sobre unos ejes cartesianos rectángulos cuyas áreas son proporcionales a las frecuencias de cada intervalo. Sobre el eje de abscisas se representan dos intervalos, el anterior al primero y el posterior al último, que no tienen valores en ellos.

Polígono de frecuencias

Ojiva: Llamamos ojiva al polígono de frecuencias acumuladas

**¿A qué llamamos patrón de comportamiento de un conjunto de datos?**

Una tabla de frecuencias, un histograma o un polígono de frecuencias describen una distribución de frecuencias, es decir, muestran el ***patrón*** de distribución de las frecuencias. En general las descripciones se refieren a aspectos de forma del histograma o el polígono de frecuencias.

Un aspecto a destacar es que la figura puede presentar un punto máximo principal. Este punto máximo se refiere al intervalo o clase con mayor frecuencia, a estos valores máximos se los denomina modos, modas o valores modales.

Otro aspecto a destacar es la simetría o asimetría de la distribución, las distribuciones pueden ser simétricas, asimétricas por derecha o asimétricas por izquierda.

**¿Qué pasa si todos los valores tienen la misma frecuencia?**

No hay moda.

**¿Cómo se describe numéricamente un conjunto de datos?**

La descripción de un conjunto de datos se realiza mediante **Métodos numéricos** debido a que **la representación gráfica de los datos es una primera incursión en el análisis de datos, pero tiene sus limitaciones.**

Necesitamos **medidas descriptivas** en forma de números que pueden concentrar mejor la atención en varias propiedades de un conjunto de datos que se investiga**.**

Las características muéstrales permiten caracterizar a una muestra con unos pocos valores, se llaman **estadísticos**.

**¿Qué medidas caracterizan a un conjunto de datos?**

Si bien cualquiera función de n observaciones de una muestra es una estadística, hay algunas que son especialmente interesantes. En términos del análisis de datos, nos interesaremos por cuatro propiedades básicas:

Medidas de tendencia central. Son medidas numéricas que **indican la localización del centro de la distribución**.

Medidas de dispersión. Son medidas numéricas que **indican el grado de variación de valores individuales alrededor del punto central.**

Medidas de asimetría: Indican el grado de asimetría, es decir, la **falta de simetría de ambos lados del valor modal** de una distribución

Medidas de apuntamiento: Es el grado de variación, o la velocidad con que sube y baja la distribución de izquierda a derecha,

**¿Qué característica tienen las medidas de tendencia central?**

Las medidas de tendencia central suelen llamarse promedios, y son el ‘valor típico’ o **REPRESENTATIVO DE UN CONJUNTO DE DATOS** en el sentido de que se emplea a veces para representar todos los valores individuales de un conjunto de datos.  
Las más frecuentemente utilizadas son la media aritmética, la mediana y la moda.

**¿Cuáles son las ventajas y desventajas de cada una de las medidas de tendencia central?**

**MEDIA ARITMETICA:**

La **media aritmética es el valor que tomaría la variable si estuviera uniformemente repartida entre todos los individuos** que forman la muestra. Tambien se la considera el **valor esperado**.

Interpretación: La estatura promedio o esperada de los estudiantes es de 166,55 cm

**La media aritmética considera todos los datos. Sin embargo, debido a que todas las observaciones se emplean para el cálculo, el valor de la media puede afectarse de manera desproporcionada por la existencia de valores extremos.**

**Ventajas de la media aritmética:**

Se trata de un concepto familiar para la mayoría de las personas y es intuitivamente claro.  
Cada **conjunto de datos posee una y sólo una media**.   
La media es útil para llevar a cabo procedimientos estadísticos como la comparación de medias de varios conjuntos de datos.

**Desventajas de la media aritmética**:

Aunque la media es confiable en el sentido de que toma en cuenta todos los valores del conjunto de datos, **puede verse afectada por valores extremos que no son representativos del resto de los datos**.   
El cálculo se hace tedioso cuando trabajamos con una gran cantidad de valores diferentes.   
Se presentan dudas al calcular la media para clases de extremo abierto, tales como, “mayor que 14” o “menor que 6”.

**MEDIANA**

La **mediana** es el valor medio o **valor central de un conjunto de observaciones.**

Interpretación: Sea la mediana 167,50. El 50% de los estudiantes observados miden 167,50 cm o menos y el otro 50% miden 167,50 cm o más

Cuando todas las observaciones se ordenan en forma creciente, la mitad de estas es menor que este valor y la otra mitad son mayor.

Si **n es impar la mediana es el valor de la observación que ocupa el lugar (n+1)/2.** Si **n es par la mediana es el promedio aritmético, de los valores que ocupan los lugares n/2 y (n+2)/2** del conjunto ordenado.  
  
**Ventajas de la mediana**

**Los valores extremos** **no afectan a la mediana tan intensamente** como a la media.   
La mediana es fácil de entender y se puede calcular a partir de cualquier tipo de datos (excepto datos cualitativos nominales) incluso a partir de datos agrupados con clases de extremo abierto, a menos que la clase mediana sea justamente una de las de extremo abierto. **[?]**

**Desventajas de la mediana**

Ciertos procedimientos estadísticos que utilizan la mediana son más complejos que aquellos que utilizan la media.

Debido a que la mediana es una posición promedio, **debemos ordenar los datos antes de llevar a cabo cualquier cálculo,** lo cual consume mucho tiempo si el conjunto de datos es muy grande.

**Modo, Moda o Valor Modal**

La **moda**, modo o valor modal de un conjunto de observaciones **es el valor de las observaciones que ocurre con mayor frecuencia en el conjunto**.

Interpretación: La estatura de los estudiantes universitarios observados que se presenta con mayor frecuencia es 161,88 cm.

***El modo es la única medida de tendencia central que puede ser calculada para variables cualitativas nominales.***

**Ventajas de la moda**

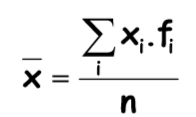
La moda, al igual que la mediana, se puede utilizar como una posición central para datos tanto cualitativos como cuantitativos.   
La moda **no se ve mayormente afectada por los valores extremos**. Podemos utilizar la moda sin importar qué tan grandes o qué tan pequeños sean los valores del conjunto de datos, e independientemente de cuál sea su dispersión.   
Podemos calcular la moda aun cuando una o más clases sean de extremo abierto**.**

**Desventajas de la moda*:***

A menudo, **no existe un valor modal** debido a que el conjunto de datos no contiene valores que se presenten más de una vez.

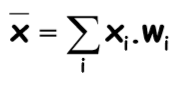
Cuando los conjuntos de datos contienen **muchas modas**, resultan difíciles de interpretar y comparar*.*

**¿A qué llamamos media pesada o media ponderada?**



**La media se define como:**

Al dividir fi por n, obtenemos **wi= fi/n**, que es la frecuencia relativa correspondiente a cada valor xi. Esta frecuencia relativa es usualmente llamada peso de cada valor xi de la variable estudiada.



**¿Qué característica tienen las medidas de dispersión?**

Las medidas de tendencia central nos indican los valores alrededor de los cuales se distribuyen los datos.   
Todas estas son valores iguales o mayores a cero, donde cero es la ausencia de dispersión.  
**Rango**: es la diferencia entre el mayor y el menor valor observados de una variable. (R=xmax – xmin)  
Interpretación: **La amplitud** de la muestra es de 35 cm.  
**Recorrido intercuartil**: Es la diferencia entre los percentiles 75 y 25, es decir, entre el tercer y primer cuartil **incluye el 50% central de la distribución.**

**Varianza: (𝜎2 𝑜 𝑠2): Es el promedio del cuadrado de las distancias entre cada observación y la media aritmética del conjunto.**

**Desviación Estándar (𝜎 𝑜 𝑠):**Interpretación: Las estaturas **se desvían, en promedio, respecto a la media aritmética**, en x cm.

**Coeficiente de Variación**: Nos indica qué **proporción de la media representa la desviación estándar.** Por esto, **suele expresarse en forma porcentual. 𝑐𝑣 = 𝑠 𝑥**

**El teorema de Chebyshev:** No importa qué forma tenga la distribución, al menos el 75% de los valores caen dentro de 2 desviaciones estándar a partir de la media de la distribución. Es decir 𝑃 (𝜇 − 𝑘 𝜎 < 𝑥 < 𝜇 + 𝑘 𝜎) = 1 − Siendo K la cantidad de desviaciones

**¿Qué es una puntuación Z? ¿Cuál es su utilidad?**

Una puntuación Z indica a qué cantidad de desviaciones estándar por encima o por debajo de la media se encuentra dicha observación.

Puntuación bruta: Es el valor observado, antes de ser convertido en una puntuación Z Para calcular una puntuación Z, se resta la media a la puntuación bruta, obteniendo el desvío. Luego se divide este desvío por la desviación estándar.

Nos permite relacionar variables con rangos muy variados entre sí, ya que la puntuación Z tiene una media 0 y desviación estándar 1.

**¿Qué característica tienen las medidas de posición no centradas?**

**Medidas de Posición No Centradas**

**Cuartiles:** La fórmula para obtener el lugar del k-ésimo cuartil, siendo n el número de observaciones, es: º𝑄𝑘 = 𝑘.(𝑛 + 1)/4   
Interpretación: (Primer cuartil) Significa que el 25% de las estaturas de los estudiantes universitarios observados son inferiores o iguales a 160 cm y el 75% restante son mayores o iguales a 160 cm.

**Deciles:** La fórmula para obtener el lugar del k-ésimo decil, siendo n el número de observaciones, es: º𝐷𝑘 = 𝑘.(𝑛 + 1)/10

**Percentiles**: La fórmula para obtener el lugar del k-ésimo percentil, siendo n el número de observaciones, es: º𝑃𝑘 = 𝑘.(𝑛 + 1)/100

**¿Qué aporta el gráfico de caja y extensiones al análisis gráfico de datos?**

Gráfico de caja y extensiones

Valores que serán necesarios para realizar el gráfico:

Valor mínimo: xmín

Valor máximo: xmáx

Mediana: x

Media aritmética: x

Primer cuartil: Q1

Tercer cuartil: Q3

Rango intercuartílico: RI = Q3 - Q1

REF1 = Q1 - 3.RI

REF2 = Q1 - 1,5.RI

REF3 = Q3 + 1,5.RI

REF4 = Q3 + 3.RI

Esta caja de ancho Rango intercuartílico (RI = Q3 - Q1) indica gráficamente el intervalo de variación de al menos el 50% de los valores centrales de la distribución.

Se añaden dos guías paralelas al eje, que llamaremos tensiones o bigotes, una de cada lado, de la siguiente forma: El primero de estos segmentos se prolonga, hacia la izquierda, desde el primer cuartil (o sea, desde la caja) hasta el valor (observado en la muestra) igual o inmediato superior a la REF2. El segundo de estos segmentos se prolonga, hacia la derecha, desde el tercer cuartil (o sea, desde la caja) hasta el valor (observado en la muestra) igual o inmediato inferior a la REF3.

**¿Qué medidas descriptivas se pueden leer en un gráfico de caja y extensiones? ¿Cuáles no se pueden leer?**

**Se pueden leer (CREO YO)**

Mediana, Media aritmética, Primer Cuartil, Tercer Cuartil, Valores atípicos, Valores anómalos y extremos de una distribución. Tambien nos informa de la simetría o aasimetria de la distribución.

**No se puede leer**

Varianza, Desviación estándar.

**¿Cuándo un dato es atípico y cuándo es anómalo?**

Si algún dato es menor a REF2 y mayor a REF1 o si es menor a REF4 y mayor a REF3 entonces dicho dato se considera **atipico**

Si algún dato es menor a REF1 o mayor a REF4 respectivamente lo indicamos con un asterisco y será denominado dato **anómalo.**

**¿Cómo se puede presentar una distribución bidimensional de variables?**

Si las variables son X e Y toman valores x1, x2, etc. e y1, y2, etc.   
Podemos escribir los datos como un ***listado*** si los datos son apareados y toman igual cantidad de valores ambas variables.

Y como ***tabla de doble entrada*** cuando los pares de datos se repiten  
Los valores fij indican frecuencias absolutas con que aparece el par xi e yj. Estas frecuencias se llaman **frecuencias conjuntas.**

**¿Qué es una distribución marginal?**

Son aquellas que resultan de sumar todas las frecuencias de la fila fi o de sumar todas las de la comluna fj

Al analizar las frecuencias marginales obtendremos distribuciones unidimensionales de X e Y, estas distribuciones de denominan distribuciones marginales.

**¿Cómo se define la independencia estadística?**

Si se cumple que f**ij**= (f**i** . f**j**)/n **para cada par de datos** entonces **la variable X e Y son independientes entre sí.** ( Donde **n** es la suma de todas las frecuencias marginales y se denomina **GRAN TOTAL**) Dependiendo el caso podemos definir una dependencia funcional o aleatoria. En los fenómenos deterministas podemos expresar una variable en función de otra, en este caso hablamos de dependencia funcional entre dos variables, en estos casos los valores de una variable quedan determinados precisamente por los valores de la otra. En cambio, tendremos casos donde podremos ver una cierta correlación entre los valores, pero no podremos definir una función que nos determine exactamente los valores de una variable a partir de la otra. En estos casos generalmente podemos, al verlos representados en un gráfico de dispersión, que las variables describen aproximadamente una curva plana. Estas correlaciones pueden ser estudiadas por medio del Calculo Numérico.

**¿Es lo mismo dependencia funcional que dependencia estocástica? ¿Por qué?**

Para algunos fenómenos, es posible encontrar una fórmula que exprese exactamente los valores de una variable en función de la otra: son los llamados **fenómenos deterministas.** Éste es el caso de **dependencia funcional** entre dos variables. En este tipo de relación, los valores que toma una de las variables que dan determinados, de un modo preciso, por los valores que toma la otra variable, que se considera como independiente.

Por ejemplo, al estudiar la caída libre de un cuerpo, donde Y representa la distancia recorrida y siendo X el tiempo transcurrido desde su lanzamiento, para una constante g = 9,8 m/s2, se sabe que Y = (1/2).g.Xˆ2

***Dependencia estocástica = dependencia aleatoria?????***

***En el caso donde no es posible encontrar una relación entre las variables mediante una curva plana se expresan mediante una representación denominada nube de puntos o dispersogramas. Aunque puede apreciarse que en ninguno de los casos es posible encontrar una relación funcional entre las dos variables, sin embrago, observamos una variación conjunta de las variables. Es decir, observar una relación, aunque no precisa, entre X e Y.***

***Al buscar dicha relación vamos obteniendo diferentes cocientes que nos ayudan a conocer la asociación que existe entre las variables estudiadas.***

**¿Qué es una distribución condicionada?**

Un tipo de distribución para la variable X es la que puede obtenerse **fijando un valor Y = yj**, que se conoce como **distribución de X condicionada** para Y = y j . Análogamente, se puede obtener la distribución de Y condicionada para X = x

**Defina el concepto de asociación.**

Al tratar de estudiar si existe o no una relación entre dos variables estadísticas, tratamos de analizar si hay algún tipo de relación entre ellas, si se puede medir la intensidad de esta relación mediante algún coeficiente, si sirve este coeficiente para comparar la intensidad de la relación de diferentes variables, etcétera. Todos estos interrogantes nos llevan a la necesidad de conocer la **asociación** que existe entre las variables estudiadas.

**¿Cuándo y para qué se utiliza un análisis de regresión?**

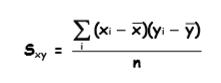
La teoría de la regresión o ajuste de curvas trata de dar medios de obtener, aproximadamente, el valor de una de las variables, cuando se da el valor de la otra.

**¿Cuándo y para qué se utiliza un análisis de correlación?**

La teoría de la correlación se ocupa de dar medidas de la dependencia entre las variables unidimensionales que entran en la variable bidimensional.

**¿Qué diferencia hay entre regresión y correlación?**

**¿Qué es la covarianza?**

Covarianza: Es un coeficiente cuyo valor nos indica el tipo de relación entre las variables.

La covarianza tiene las siguientes características:

**Es igual a cero si las variables son independientes.**

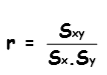
**Es positiva si las variables tienen dependencia directa.**

**Es negativa en caso de dependencia inversa.**

**Un problema de la covarianza es que no hay un máximo para el valor que puede tomar, por lo cual no nos sirve para comparar la mayor o menor intensidad de la relación entre las variables.**

**¿Cuáles son las características del coeficiente de correlación?**

El coeficiente de correlación lineal o coeficiente de Pearson es un coeficiente que permite estudiar la intensidad de la relación entre las variables



Siendo Sx y Sy, las desviaciones estándar de las variables X e Y

Puesto que las desviaciones estándar son positivas, r tiene el signo de la covarianza y, por tanto:

Si r > 0, la relación entre las variables es directa.

Si r < 0, la relación entre las variables es inversa.

Si r = 0, no existe relación lineal entre las variables.

Si r = +1, la correlación lineal es perfecta y positiva.

Si r = -1, la correlación lineal es perfecta y negativa.

**¿Cómo se debe interpretar el coeficiente de correlación?**

La interpretación del coeficiente de correlación como medida del grado de relación lineal entre dos variables es una **interpretación matemática pura** y **está completamente desprovista de implicaciones de causa y efecto.** El hecho de **que dos variables tiendan a aumentar o disminuir al mismo tiempo no implica que una tenga algún efecto directo o indirecto en la otra.** Ambas pueden estar sometidas a la influencia de otras variables, de manera que resulten con una estrecha relación matemática

Por ejemplo, en un período de varios años el coeficiente de correlación entre los sueldos de maestros y el consumo de licor ha resultado ser de 0,98.

Al valor de r² se lo denomina **coeficiente de determinación.** Expresa la proporción de la variación total en los valores de Y que se pueden explicar mediante una relación lineal.

Es r² % de la variación total de los valores de Y en nuestra muestra

**¿Qué es interpolar? ¿Qué es extrapolar?**

Si se tiene interés en un valor intermedio de X, que no ha sido analizado en la muestra, se emplea la línea de regresión para encontrar el valor predicho para Y , correspondiente al valor de x . Diremos, en este caso, que hacemos **interpolación.**

Hacer predicción para valores de X fuera del rango de la muestra se llama **extrapolación** y, en general, si se supone que la relación es lineal sólo en esta región de valores de X, **no es legítimo usar la línea recta para predecir valores de Y fuera de este intervalo** de valores de X.

**¿Son ambas acciones válidas para un análisis estadístico?**

No, Solo la interpolacion