Introducción al análisis estadístico

Objetivos:

1. Aprender lo que es estadística y el uso de algunos métodos estadísticos para el análisis de data.
2. Aprender a organizar y resumir datos en tablas y gráficas.

I. Estadística

La estadística es una ciencia que nos permite colectar data, clasificarla y presentarla de forma ordenada para su análisis. Los métodos estadísticos permiten descifrar de forma más precisa los resultados para poder tomar decisiones y hacer estimaciones. Además, el método estadístico ayuda a obtener información a través de la interpretación de unos datos mediante tablas, gráficas, medidas de resumen, comparaciones e interpretaciones.

Hay que tener en consideración unas etapas para obtener información de los datos:

- Diseñar el proceso de colección de datos.
- Resumir los datos
- Analizar la data
- Reportar las conclusiones obtenidas del análisis.

II. Recolección de datos

Al momento de realizar una investigación es importante hacer un diseño experimental que permite planificar bien cómo se va a tomar los datos. Se debe seleccionar una población de interés de la cual se quiere concluir o estudiar. De la muestra de la población seleccionada se va obtener unos datos a través de la investigación con la cual, se infiere sobre la población a partir de la muestra investigada.

Para realizar una investigación se debe tener unos objetivos o saber para que se recolecta los datos. Además, se debe hacer unas observaciones de las cuales uno puede diseñar y seleccionar como se va realizar la investigación.

Se debe pensar de la siguiente forma:

- ¿Para qué se realiza la investigación?
- ¿Qué provee la investigación?
- ¿Cómo se debe realizar la investigación?

Los datos pueden ser obtenidos a través de muestreos, experimentos y/o estudios observacionales. Algunos modos de recolección de datos son las entrevistas, cuestionarios, transeptos, observaciones directas y trampas.

III. Resumir los datos

Los datos de una investigación pueden organizarse y resumirse usando tablas, gráficas y medidas numéricas de resumen para un mejor análisis de resultados. Las tablas y las gráficas ayudan a analizar e interpretar los datos y permite presentar de forma clara los resultados.

A. Tablas

Durante la investigación se puede generar muchos datos que son colectados en tablas conocidos como datos crudos. Al finalizar la parte experimental de la investigación, se organiza de un modo mejor los datos crudos. Los datos crudos son organizados en tablas más simples y son utilizados para hacer diversas pruebas estadísticas.

En los reportes científicos, como son los informes de laboratorio, la tabla de datos se coloca resultados e información relevante para comunicar los hallazgos. Los datos se organizan en una tabla de modo que uniforme, proveyendo medidas de medición, tiempos, etc. Por

ejemplo, si se toma mediciones de temperatura, pH y conductividad en un cuerpo de agua, entonces cada parámetro sería una columna en la tabla y en el encabezado de cada columna indicaría el parámetro. Se debe colocar las unidades de medidas utilizadas en la columna correspondiente a la data medida, es decir, Temperatura (ºC). El título de la tabla se coloca en la parte superior de la tabla y debe ser claro con la información necesaria para entender la información dada en la tabla. Las tablas de datos crudos pueden tener líneas o bordes delineando toda la tabla. Las tablas para reportes científicos se hacen con son líneas horizontales al comienzo y al final.

Ejemplo de una tabla para reporte científico

Tabla 1. Parámetros físico-químicos monitoreados en la Quebrada de Oro durante el mes de enero de 2020

Fechas de muestreo	**Temperatura (ºC)**	**pH**	**D.O. (ppm)**
15/enero/202			

B. Gráficas

Las gráficas permiten presentar las relaciones entre las variable independiente y dependiente de la investigación. Al igual que las tablas, las gráficas presentan un resumen de los resultados más relevantes de la investigación. La interpretación de los resultados puede ser más fácil al ser presentados en gráficas.

En la gráfica se debe escoger cual es la **variable dependiente** que es la variable a la cual uno le mide algo durante el experimento, y la **variable independiente** que es la variable que uno manipula o cambia y afecta a la variable dependiente. Cuando se construye una gráfica, la variable independiente va en el eje de x y la variable dependiente en el eje de y. Se recomienda que se coloque ambos eje en cero y se escoja un intervalo apropiado para la data. Se debe rotular los ejes indicando la variable y las unidades de medidas. Se puede usar leyendas para indicar aspectos diferentes presentados en la misma gráfica. Para reportes científicos escritos se debe usar patrones y tonos grisáceos para designar los diferentes aspectos en la leyenda. Se debe escoger la gráfica apropiada para presentar los datos. Además, se debe colocar un título a las gráficas que sea corto pero que indique lo que se proyecta en la gráfica que va en la parte posterior de la gráfica. **El colocar temperatura versus tiempo no es un título apropiado**. Se debe rotular los ejes incluyendo las unidades de medidas de las variables. Ejemplos de gráficas son las gráfica de punto, barra, "pie chart", de línea, histograma entre otras.

Las gráficas se colocan como si fuera una figura en el reporte científico con una pequeña oración en la parte posterior de la gráfica con el título de la misma lo más completo posible para que el lector entienda lo que se está proyectando.

C. Medidas numéricas de resumen

Hay medidas que dan una idea de donde está el centro de la distribución como es la moda mediana y media. Estas son medidas de tendencia central que permite localizar el centro de un grupo de datos.

1. La **moda** mide el dato más frecuente, quiere decir, el que más se repite. Esta medida se puede usar para datos cuantitativos y cualitativos. A veces la moda puede estar lejos del centro de la distribución.

Ejemplos: Tiene una serie de medidas de número de frutos por plantas.

3,5,**7,**9,**7**,9,5,**7**,1 Moda: 7
3,2,3,5,4,3,2 Moda: _____
8,12,15,12,8,9 Moda: _____

2. La **mediana** es el valor central en unos datos organizados de menor a mayor.
n = el número de datos u observaciones
Si n es impar, la posición de la mediana es n+1/2.
Si n es par, la mediana es el promedio de los datos en posiciones n/2 y n/2 +1.
Ejemplos:
7,9,11,11,13 n=5 mediana = n+1/2 ⇒ 5+1/2 ⇒ 3
1,5,6,**7,8**,10,10,11 n= 8 mediana = n/2 y n/2 +1⇒ 8/2 y 8/2 +1 ⇒ 4 y 5
Mediana = 7+8/2 = **7.5**
3,4,5,6,7,8,12,14,21 n= ___ mediana =

4,6,8,10,12,15,15,17 n= ___ mediana =

3. La **media o promedio** se obtiene sumando todos los datos recopilados y dividiendo el resultado por el total de datos. Puede ser afectada por datos anormales.
Promedio = Σx/n Σ = sumatoria x = los datos n = total de datos
Ejemplos: 8,9,15,6,9,18,20,25 n=8
8+9+15+6+9+18+20+25/8⇒ 110/8⇒13.75

83,85,92,110,125,144 n=_____

D. Medidas de variabilidad o dispersión

Ayudan a saber el grado de concentración o dispersión de los datos con respecto al centro de la distribución. Incluye la descripción de la cantidad de dispersión, la varianza y la desviación estándar. La varianza y la desviación estándar mide la dispersión con referencia a la media o promedio. La desviación del dato es cero si el dato es igual a la media o promedio. Sin embargo, la desviación es positiva si el dato (X) es mayor que la media y la desviación es negativa si el dato (X) es menor que la media. Los datos que están agrupados más cercanos van a tener valores relativamente pequeños, y más dispersos los datos más grandes los valores en las medidas de dispersión.

Por ejemplo, estas midiendo la altura (cm) de 10 plántulas de una misma especie germinadas en un invernadero. ¿Cuál de los dos casos crees que va a tener una medida de dispersión más grande? Compare los datos, ¿Cuál caso tiene datos más dispersos?

Caso 1:
6.5, 7, 7.5, 7.2, 6.6, 7.8, 6, 6.2, 7.4, 6.8 promedio = 69/10 = 6.9cm

Caso 2:
6.5, 7, 7.5, 10, 5, 6.6, 7.8, 6, 6.2, 6.4 promedio= 69/10= 6.9cm

1. El **rango o amplitud** es la diferencia entre el dato mayor (H) y menor (L).
Rango = H-L
Ejemplo: 8,10,15,89,17
89-8 = 81

2. La **varianza** es la distancia a partir del centro. Mide la medida de la dispersión de la data sobre la media de los datos.
s^2 es la varianza muestra.
$s^2 = \Sigma(X-x)^2/n-1$ n= número de datos x= promedio o media

X	X-x	$(X-x)^2$
8	8-9	$(1)^2=1$
12	12-9	$(3)^2=9$
15	15-9	$(6)^2=36$
9	9-9	$(0)^2=0$
4	4-9	$(-5)^2=25$
6	6-9	$(-3)^2=9$
		80

$x = 54/9 = 9$ n=6

$\Sigma(X-x)^2/n-1 \Rightarrow 80/5$

16

3. La **desviación estándar** es la raíz cuadrada de la varianza. La unidad de medida de la desviación estándar es la misma que la unidad de medida de los datos. Se calcula
$s = \sqrt{s^2} \rightarrow \sqrt{\Sigma(X-x)^2/n-1}$
$\sqrt{80/5} \Rightarrow \sqrt{16} \Rightarrow 4$

E. Comparar los datos obtenidos experimentales y control

1. La **prueba estadística Chi Cuadrado** de Contingencia también conocida como Chi Cuadrado de Independencia ayuda a comparar los datos obtenidos de 2 grupos independientes (control y un experimental).

2. **Anova (Análisis de Varianza)** es usado para ver la variabilidad entre los grupos y dentro de los grupos es decir comparar los tratamientos y controles como los datos obtenidos dentro de cada tratamiento y controles. Así que se busca ver una variabilidad total del experimento desglosados en variabilidad entre grupos + variabilidad dentro de grupos. Esta prueba se usa cuando tiene más de un tratamiento o grupo experimental, en donde se prueba simultáneamente todas las medias.
Se usará las siguientes anotaciones para designar: t como tratamientos y n_i como repeticiones
Y_{ij} quiere decir j en una observación y i es un tratamiento

$Y_{i\bullet} = \sum_{i=1}^{n_i} Y_{ij}$ es la suma de todas las observaciones del tratamiento i

$Y_{\bullet\bullet} = \sum_{i=1}^{t} Y_{i\bullet}$, es la suma de todas las observaciones

$\bar{Y}_{i\bullet}$ es la media (promedio) de las observaciones del tratamiento i

$\bar{Y}_{\bullet\bullet}$ es la media (promedio) de todas las observaciones

$n_{\bullet} = \sum_i n_i$ es la cantidad total de observaciones

Las sumas de cuadrados se calculan de la siguiente manera:

- SCTotal es la suma de cuadrados total que es a base de las medidas y sus promedios y depende de la variabilidad dentro de los datos de los tratamientos y la variabilidad entre los tratamientos.
- SCEntre es la suma de cuadrado entre tratamientos (SCTratamientos) que está basado en la medida de la variabilidad del promedio de los tratamientos
- SCError es la variabilidad que tiene los datos que no son explicado por la diferencia entre tratamientos. SCError es lo mismo que la suma de cuadrados dentro del tratamiento también designada como la suma de cuadrado residual.

SCTotal $= \sum_{ij}(y_{ij} - \bar{y}..)^2$
SCEntre = SCTratamientos = $n\Sigma_i(\bar{y}_i - \bar{y}..)^2$

SCDentro = SCResidual = SCError = SCTotal – SC Tratamientos

Fuente de variación	Suma de cuadrados	grados de libertad	Cuadrado medio	F
Tratamientos	SCTratamiento	t-1	CMTratamiento	F = CMTrat./CMRes.
Residual	SCResidual	$n_{\bullet} - t$	CMResidual	
Total	SCTotal	$n_{\bullet} - 1$		

t = # tratamientos $n_{\bullet} = \sum_i n_i$ es la cantidad total de observaciones

Los cuadrados medios se calculan de la siguiente manera:
CMTratamiento = SCTratamiento/t-1
CMResidual = SCResidual/ $n_{\bullet} - t$

Las hipótesis que se analizan a través de esta prueba son:
H_0 = μtratamiento 1 = μtratamiento 2 = H_0 = hipótesis nula
H_a = al menos un μ es diferente H_a = hipótesis alterna

Estadístico de la prueba:
F = CMTratamiento/CMResidual
Región de rechazo:
$F > F_{\alpha}$ (grados de libertad: t-1, $n_{\bullet} - t$)

$df_1 = t-1$ $df_2 = n_\bullet - t$

Se rechaza la hipótesis nula cuando F excede a F_α. Se calcula F_α en la tabla de Porcientos de puntos de la distribución de F que se encuentra como apéndice en los libros de estadísticas.
Ejemplo:

Tratamiento	Especie 1	Especie 2	Especie3	Especie4	sumatoria	Promedio
Control	28	35	27	21	111	27.75
Suelo 1	21	36	25	18	100	25
Suelo 2	26	38	27	17	108	27
Suelo 3	16	25	22	18	81	20.25
					400	

$n_\bullet = 16$ $t=4$

Fuente de variación	Suma de cuadrados	grados de libertad	Cuadrado medio	F
Tratamientos	136.5	3 t-1	45.5	**0.98** F = CMTrat./CMRes.
Residual	555.5	12 $n_\bullet - t$	46.292	
Total	692	15 $n_\bullet - 1$		

$\bar{y}.. = 400/16 = 25$

$SCTotal = \sum_{ij}(y_{ij} - \bar{y}..)^2$
$(28-25)^2 + (35-25)^2 (27-25)^2 + (21-25)^2 + (21-25)^2 + (36-25)^2 + (25-25)^2 + (18-25)^2 + (26-25)^2 + (38-25)^2 + (27-25)^2 + (17-25)^2 + (16-25)^2 + (25-25)^2 + (22-25)^2 + (22-25)^2 + (18-25)^2 = 692$

$SCTratamiento = n\Sigma_i(\bar{y}_i - \bar{y}..)^2$
$4[(27.75-25)^2 + (25-25)^2 + (27-25)^2 + (20.25-25)^2] = 136.5$

SCResidual = SCTotal – SC Tratamientos
= 692-136.5 = 555.5

CMTratamiento = SCTratamiento/t-1 136.5/3 = 45.5

CMResidual = SCResidual/ $n_\bullet - t$ 555.5/12 = 46.292

F = CMTrat./CMRes. 45.5/46.292 =0.98

Región de rechazo para la hipótesis nula:
F > F_α (grados de libertad: t-1, $n_\bullet - t$)

0.98 _ 3.89 Se rechaza o se acepta?

Hay o no hay diferencia significativa en los promedios de los tratamientos de suelos

Tabla de porciento de distribución de F
http://www.socr.ucla.edu/Applets.dir/F_Table.html#FTable0.05

Ejercicio:
1. Calcula el promedio, la varianza y la desviación estándar a los siguientes datos:
23.6, 24.1, 24.2, 26.4, 25.3, 25.7, 25.6, 25.4, 24.1

2. Calcula el promedio, la varianza y la desviación estándar a los siguientes datos:
23.6, 24.1, 24.2, 26.4, 25.3, 25.7, 25.6, 25.4, 30

3. Realiza varias grafica para resumir los siguientes datos obtenidos a través de un monitoreo de un cuerpo de agua. Calcula la temperatura y el oxígeno disuelto promedio para el cuerpo de agua con su desviación estándar.

Día de muestreo	Temperatura (^oC)	D.O. (mg/L)	pH
24/02/2004	23.6	6.14	7.6
06/03/2004	24.1	6.22	6.9
22/03/2004	24.2	5.95	6.6
07/04/2004	26.4	6.01	7.6
20/04/2004	25.3	6.51	7.2
04/05/2004	25.7	5.62	7.3

4. Haz un análisis de varianza para los siguientes datos obtenidos a través de un experimento con plantas de tomate crecidas bajo distintos dos nutrientes. Doce plantas fueron usadas para el estudio, en donde 4 plantas fueron asignadas para cada tratamiento de forma aleatoria. Determine las hipótesis analizar.

Tratamientos	Grupo 1	Grupo 2	Grupo 3	Grupo 4		
Control	21	18	16	14		
Nutriente 1	12	14	15	10		
Nutriente 2	7	9	6	7		

Fuente de variación	Suma de cuadrados	grados de libertad	Cuadrado medio	F
Tratamientos	SCTratamiento	t-1	CMTratamiento	F = CMTrat./CMRes.
Residual	SCResidual	$n_\bullet - t$	CMResidual	
Total	SCTotal	$n_\bullet - 1$		

Referencia

Johnson R. 1995. Just the essentials of elementary statistics. Duxbury Press, CA, 478 pp.

Lyman Ott R. y M. Longnecker. 2001. An introduction to statistical methods and data analysis. 5ta ed. Duxbury, CA, 1,152 pp.

¿Cómo realizar un Anova de un factor en excel?