FACULTAD DE BIOLOGÍA

UNIVERSIDAD MICHOACANA DE SAN NICOLÁS DE HIDALGO

MUESTREO Y PRUEBA DE HIPÓTESIS

ELABORACIÓN DE TABLAS DE FRECUENCIAS, HISTOGRAMAS Y PRUEBAS DE PARA DETERMINAR EL AJUSTE A LA DISTRIBUCIÓN NORMAL

Conceptos y aspectos básicos del tema:

De acuerdo con lo que se revisó en clase,

- Defina frecuencia en función de un conjunto de datos agrupados:

- Defina Clase e intervalo de clase:

- Defina Frecuencia Relativa:

- Si tenemos un conjunto de datos con n=40, de acuerdo con la regla de Sturges ¿Cuántas categorías o clases (k) deberemos hacer para tener una buena representación en un histograma?

- Si a regla empírica nos dice que 5 ó 6 clases son suficientes para esta *n*, ¿qué tan diferente es *k* con respecto a lo obtenido con Sturges?

El diagrama de tallo y hoja es una representación de las frecuencias en una población estadística con la diferencia con respecto a un histograma, de que no se pierde el valor absoluto de cada uno de los datos.

Con el siguiente conjunto de datos, obtenidos del diámetro del cráneo de bebés recién nacidos, elabore la tabla de frecuencias correspondiente, incluyendo la frecuencia relativa y la acumulada para cada una de las categorías que proponga. Elabore también un diagrama de tallo y hoja y un Histograma de Frecuencias para comparar las tendencias y la información que podemos obtener a partir de estos resultados.

Veamos cómo hacer esto en **Excel** siguiendo las instrucciones que están en el archivo.

Una vez terminada la parte de aplicar las técnicas y obtener la tabla y los gráficos, por favor incluya sus resultados y concluya sobre la información que puede obtener con este análisis.

LA DISTRIBUCIÓN NORMAL (DN) Y SUS PROPIEDADES: TENDENCIA CENTRAL, SIMETRÍA Y VARIACIÓN

Conceptos y aspectos básicos del tema:

- La agrupación de la mayor cantidad de los datos posibles hacia el centro de la distribución es lo que confiere forma de "campana" a la línea de tendencia resultante de un histograma.

- Si los datos se ajustan a una DN, entonces "emergen" los PARÁMETROS, que son los atributos o características de la Distribución Normal.

- Cuando se utiliza un modelo de DN cuya media es "cero" y la Desviación Estándar es igual a "uno", se le llama "Distribución Normal Estandarizada" y es el fundamento para las técnicas de la Estadística Paramétrica, bajo el supuesto de que todas las poblaciones estadísticas que se

ajusten a la DN podrán ser descritas en función de sus parámetros, utilizando para ello los "estimadores".

En el siguiente ejercicio, con los datos correspondientes a los pesos en gramos de peces de dos poblaciones ubicadas en localidades geográficamente distintas (llamémosles **Pob 1** y **Pob 2**) hagamos su análisis e interpretación.

Los datos se encuentran en el archivo **peces.xlsx:**

Efectúe el Análisis de Frecuencias y RESALTE las características de cada población.

Población 1.

Población 2.

TABLAS DE FRECUENCIAS E HISTOGRAMAS CON EL PROGRAMA PAST

PAST (Hammer et al. 2001) es un programa gratuito en el que continuamente se están actualizando sus contenidos y módulos. Originalmente orientado a Estadística y Paleontología, se ha enriquecido con técnicas útiles para la Bioestadística en varios campos de la Biología como la Taxonomía y la Ecología principalmente. Aceptado como herramienta en publicaciones internacionales, su facilidad de manejo y gratuidad, le hace una opción interesante para su uso en análisis de datos provenientes de trabajos de investigación en el área de las ciencias biológicas. Por esta razón se utiliza como ejercicio en este curso.

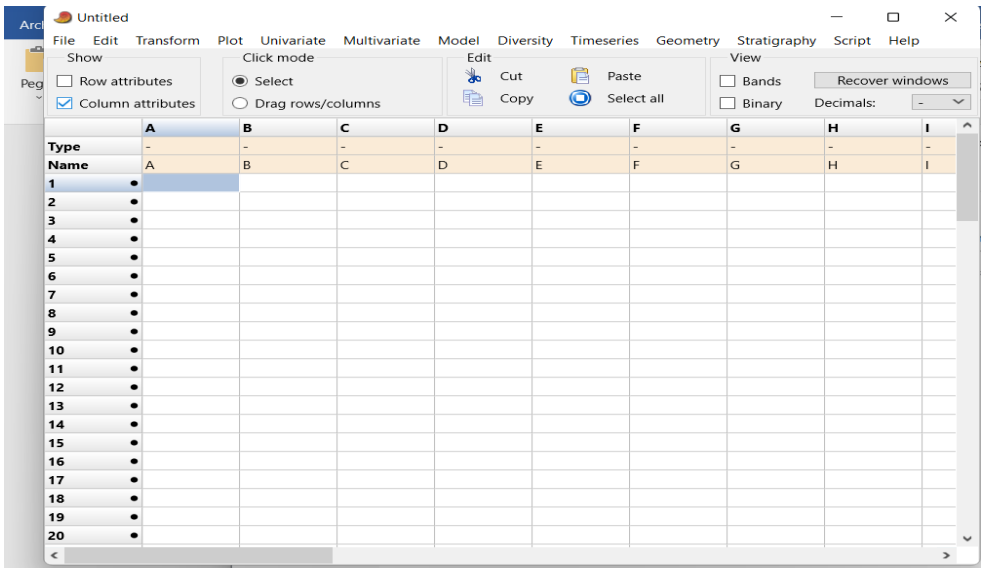
Enseguida le pedimos que revise el pequeño tutorial que se les anexa en una presentación en PowerPoint.

TUTORIAL PAST 4.0

Para determinar las pruebas estadísticas de normalidad en el programa PAST:

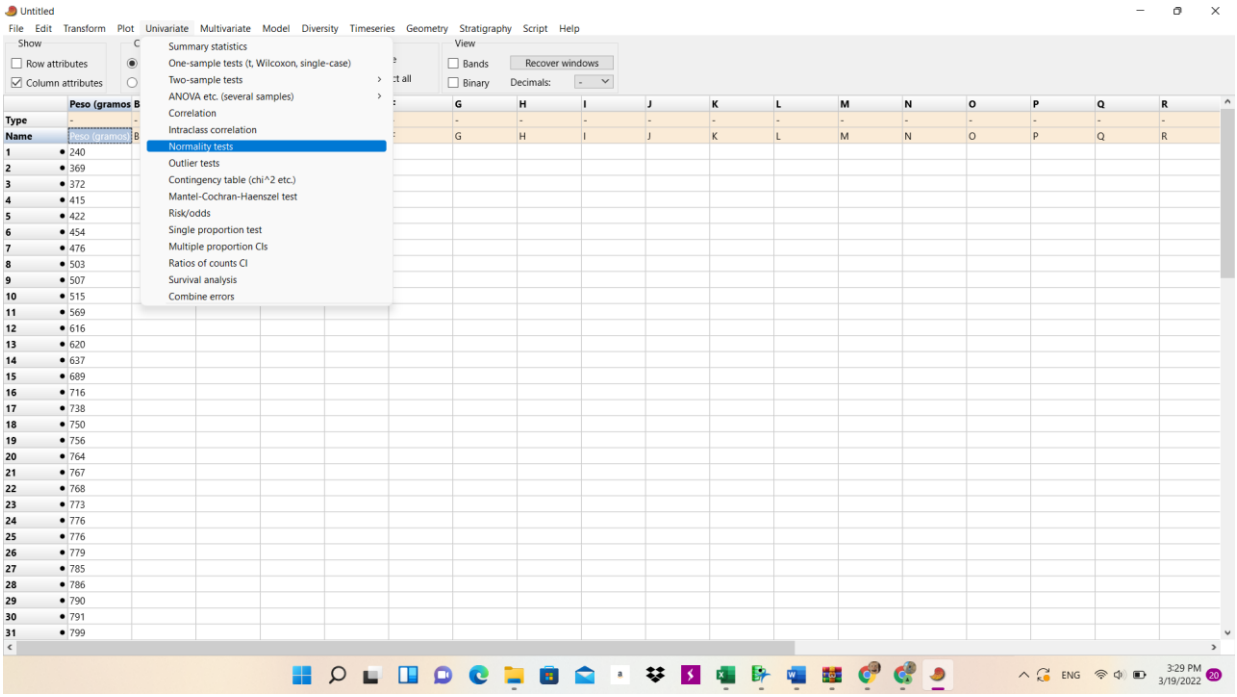
1. Primero se tiene que entrar al programa de Past; una vez que entras seleccionas el cuadro de ”column attributes” para poder pegar los datos con rótulos.

2. Selecciona la columna de todos los pesos (gramos) de la población 1, las copias para poder pegarla en Past. Usar Ctrl/C para copiar y Ctrl/V para pegar.

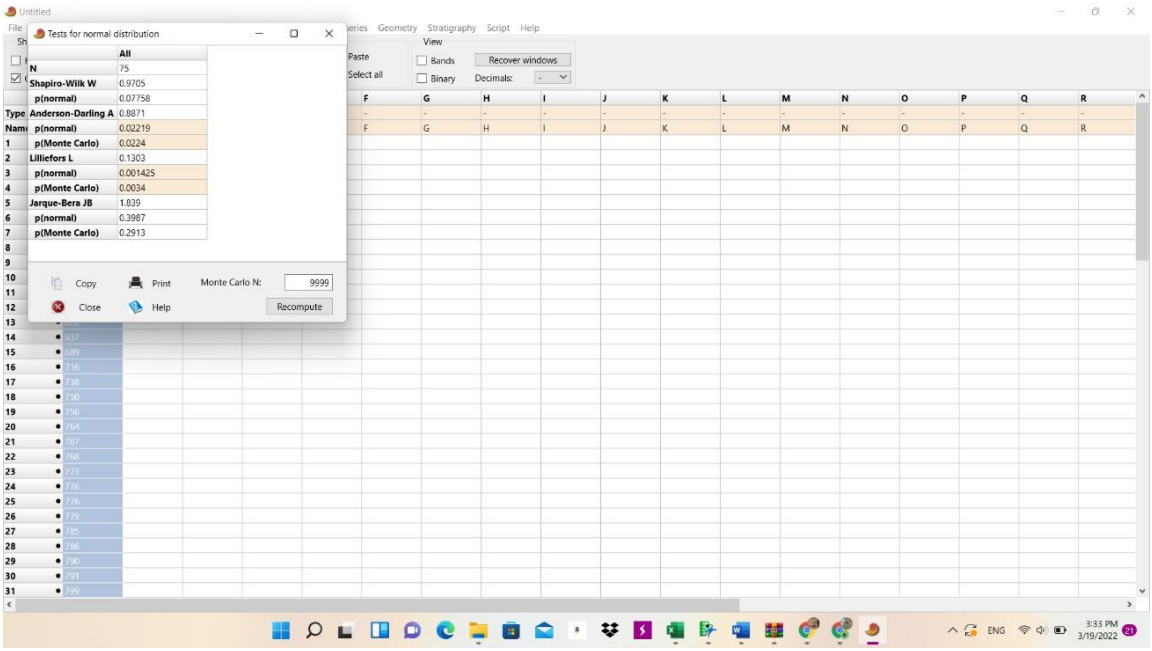


3. Los datos quedan pegados en Past.

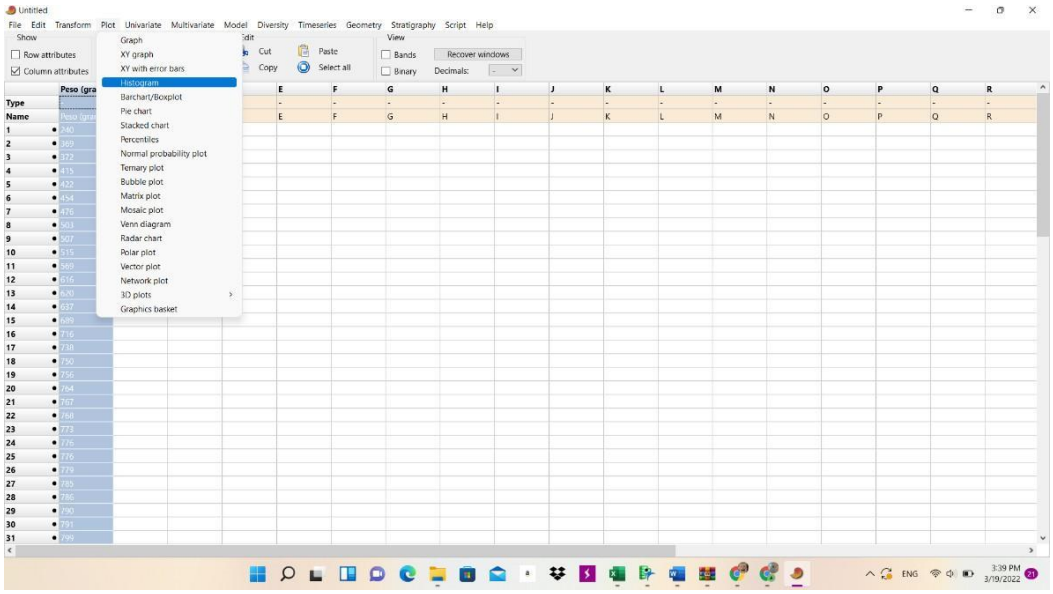
   Después seleccionas todos tus datos (Entrar en modo Ejecución) y das clic en la parte que dice
   “Univariate” y seleccionas **“Normality test”**.



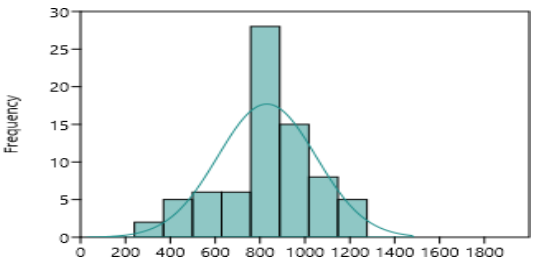
4. El programa dará información de las pruebas estadísticas de la normalidad de Shapiro-Wilk, Anderson- Darling y otras. La más aceptada es la de Shapiro-Wilk.

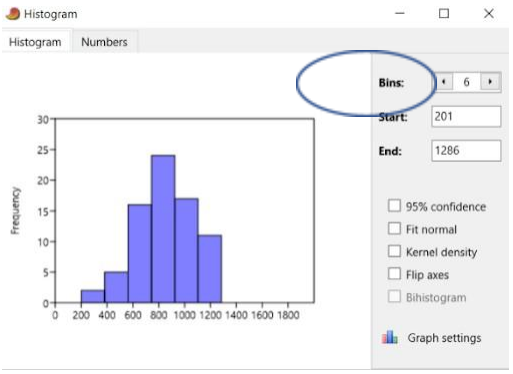


5. Ahora para generar el histograma en el programa Past, igualmente seleccionas tus datos, te diriges en la pestaña que dice “Plot” y luego buscas la opción “Histogram”.

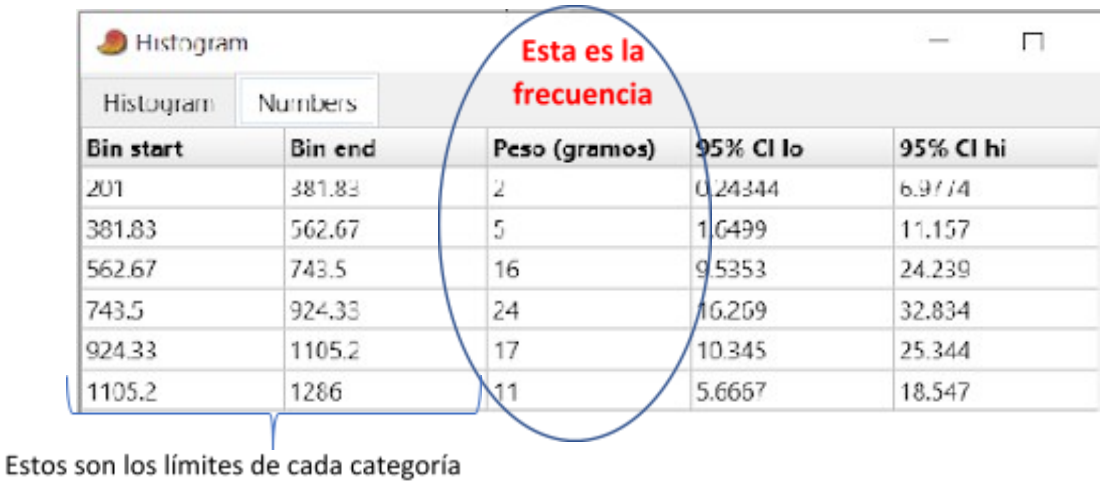


6. Se genera el histograma con la curva normal que se ajusta a sus datos.





7. Para obtener la tabla de Frecuencias solo debe elegir la pestaña “Numbers”



NOTA: Si se desea tener más o menos clases o categorías, en las opciones que aparecen a la derecha del histograma lo puede hacer. La palabra “bin” a eso se refiere.

TABLAS DE FRECUENCIAS E HISTOGRAMAS CON EL PROGRAMA R (Interface RStudio)

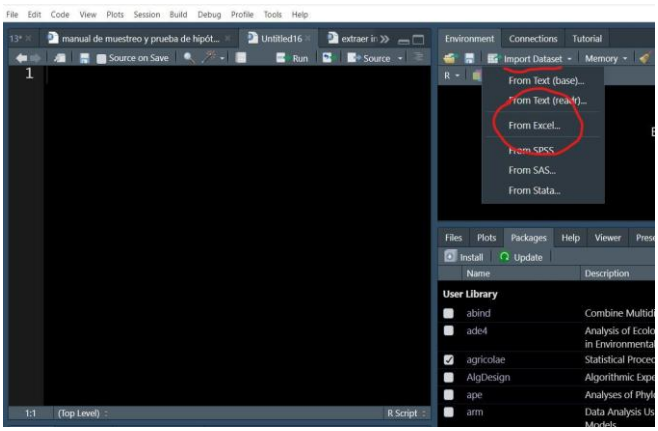
R es un entorno de software libre para computación estadística y gráficos. Se ejecuta desde diferentes plataformas como UNIX, Windows y MacOS. Es un programa gratuito con actualizaciones en comunidad, orientado a la generación de cualquier análisis y gráficos estadísticos y del cual se puede tener acceso desde https://www.r-project.org/. En publicaciones internacionales se recomienda utilizar este software dados los procesos visibles "paso a paso" y su fácil replicación. Por lo anterior, se sugiere el uso y/o conocimiento de este software en este curso.

Para determinar las pruebas estadísticas de normalidad en el programa R se utilizarán los mismos datos que para el ejemplo utilizado en el programa PAST (**peces.xlsx**):

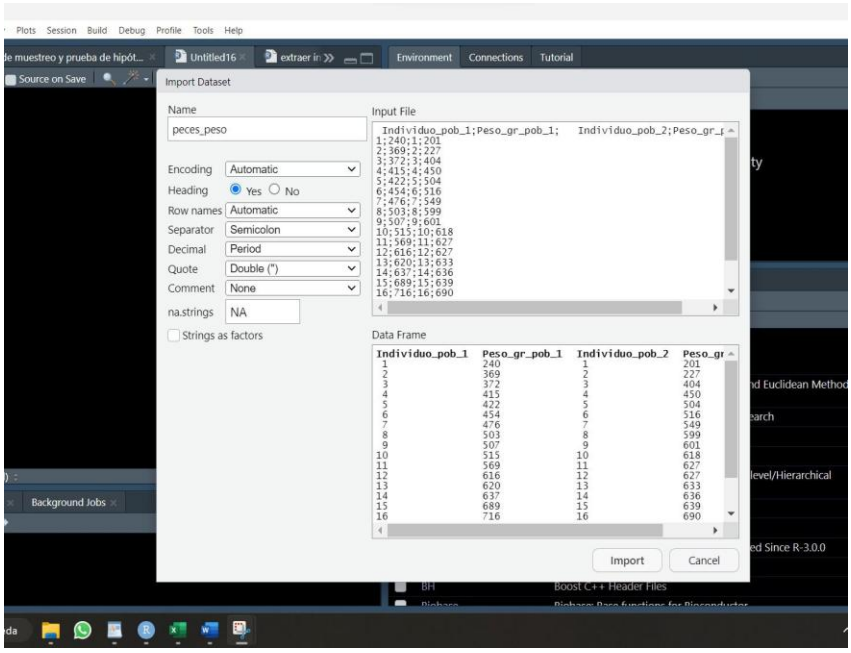
1. Se sugiere ampliamente generar bases de datos saturadas (sin celdas vacías) en formatos ASCII (e.g. extensión csv), evitar caracteres especiales como acentos, eñes (ñ), espacios, arroba (@), entre otros. A continuación, se muestra parte de la base de datos "**peces.xlsx**" con modificaciones para su uso en R.

	A	B	C	D
1	Individuo_pob_1	Peso_gr_pob_1	Individuo_pob_2	Peso_gr_pob_2
2	1	240	1	201
3	2	369	2	227
4	3	372	3	404
5	4	415	4	450
6	5	422	5	504
7	6	454	6	516
8	7	476	7	549
9	8	503	8	599
10	9	507	9	601
11	10	515	10	618
12	11	569	11	627
13	12	616	12	627
14	13	620	13	633
15	14	637	14	636
16	15	689	15	639
17	16	716	16	690
18	17	738	17	710
19	18	750	18	712
20	19	756	19	713
21	20	764	20	722
22	21	767	21	723
23	22	768	22	723
24	23	773	23	730
25	24	776	24	772
26	25	776	25	774

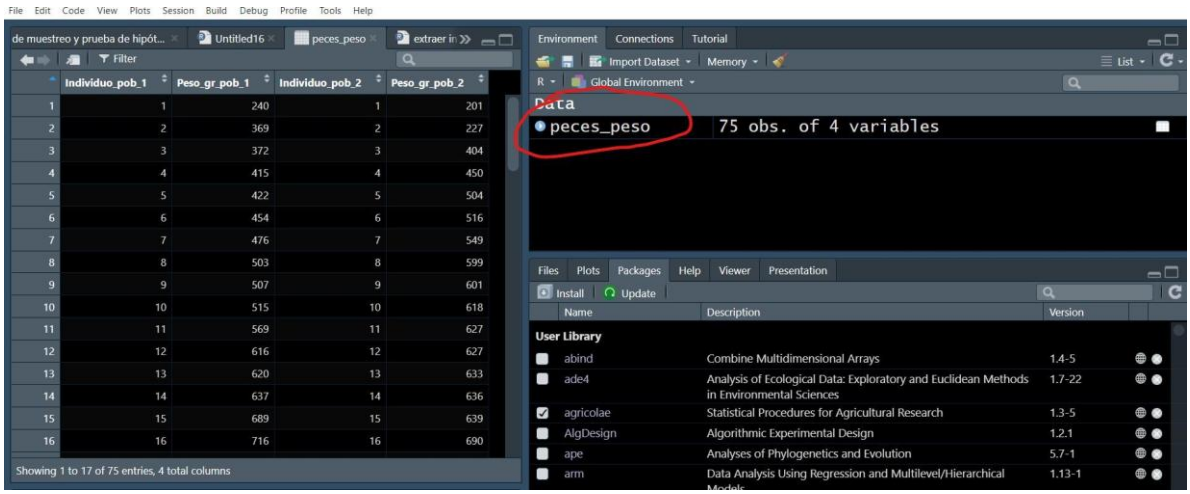
2. Una vez hecho lo anterior se abre el archivo desde RStudio y buscamos la secuencia "importar DataSet
   -> From Excel" (ver figura).

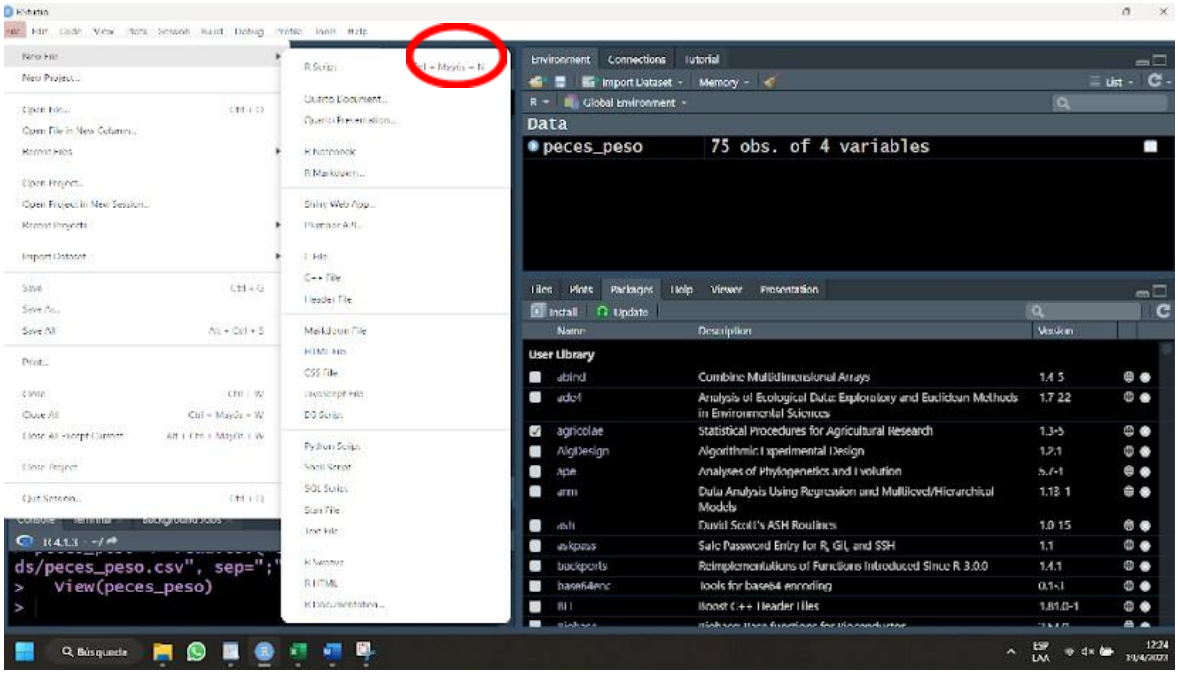


3. A continuación, nos pedirá configurar “opcionalmente” la base de datos. Se sugiere marcar las casillas “Yes” para “heading” (reconocer primera fila como título de columna). Posteriormente presionamos “import”



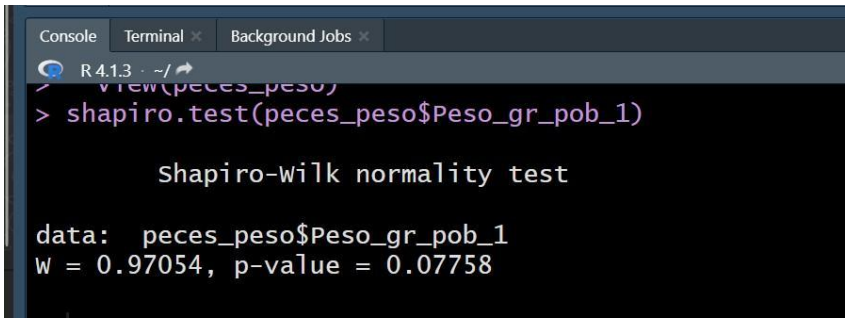
4. Una vez hecho lo anterior se generará una base de datos y se guardará de manera temporal en el ambiente de trabajo. Posteriormente hay que abrir un “scrip” para generar las líneas de comandos para análisis.





5. Una vez hecho lo anterior se selecciona la columna de la base de datos y se utiliza el comando "shapiro.test" y los datos a analizar "peces_peso$Peso_gr_pob_1" para generar el análisis de normalidad. Para ejecutar la orden se selecciona toda la línea de comandos:

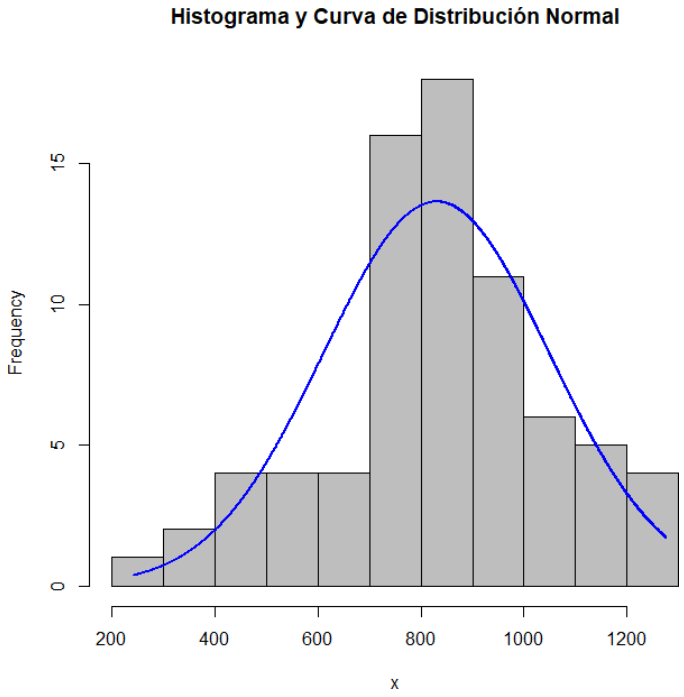
shapiro.test(peces_peso$Peso_gr_pob_1) La consola mostrará los resultados:



6. Para generar el histograma se seleccionan los mismos datos y se utiliza la siguiente línea de comandos:

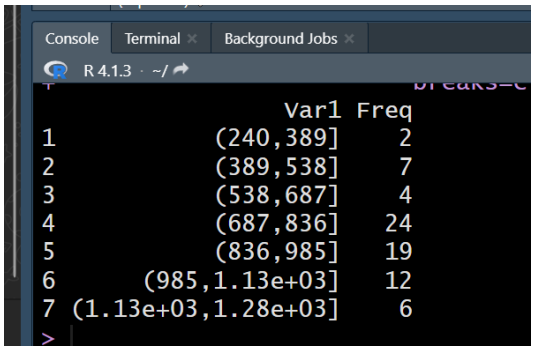
windo
ws()
library(
rcompa
nion)
plotNormalHistogram(peces_peso$Peso_gr_pob_1, prob = FALSE, main = "Histograma y Curva de Distribución Normal", length = 2000)

En la sección "Plots" se generará el histograma y la curva del modelo de normalidad



7. Para obtener la tabla de Frecuencias se utiliza la siguiente línea de comandos (nótese que los datos siempre se declaran de la misma manera "peces_peso$Peso_gr_pob_1":

```
library(agricolae)
clases<-with(peces_peso,sturges.freq(Peso_gr_pob_1))
as.data.frame(table(cut(as.numeric(peces_peso$Peso_gr_pob_1),
                        breaks=clases$breaks)))
```



Si requiere calcular medidas de dispersión y/o tendencia central (ver sección a continuación), usted puede utilizar las siguientes líneas de comandos:

Varianza

var(peces_peso$Pe

so_gr_pob_1)

Desviación

estándar

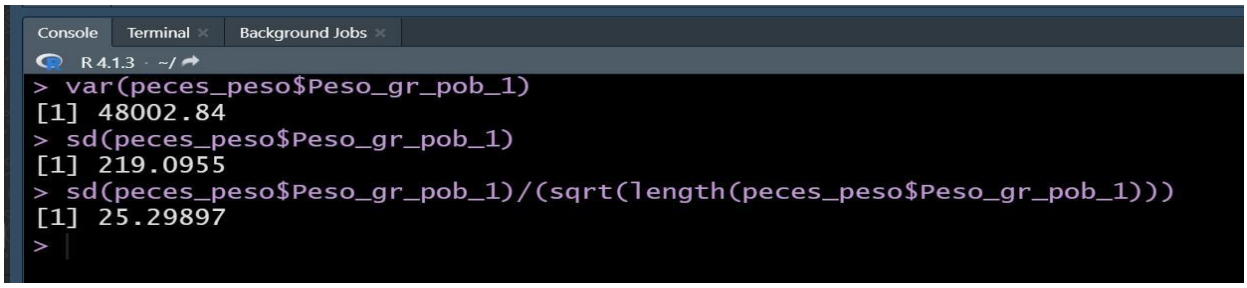
sd(peces_peso$Pes

o_gr_pob_1)

Error estándar

sd(peces_peso$Peso_gr_pob_1)/(sqrt(length(peces_peso

$Peso_gr_pob_1)))



ESTADÍSTICA DESCRIPTIVA

Recordemos:

a) Si un conjunto de datos (una muestra) se ajusta a una Distribución Normal, entonces se puede representar con la Distribución de frecuencias y la función de densidad de probabilidad correspondiente que tiene la forma de una "campana" perfectamente simétrica, con todos los datos posibles para una media y una varianza determinada distribuidos a ambos lados de la media.

b) La media y la varianza son los "parámetros" fundamentales y a partir de ellos se genera la Estadística Paramétrica.

c) Hay otros parámetros derivados de la varianza como son: La Desviación Estándar y la simetría por ejemplo.

Del conjunto de Datos que se presenta en el archivo: **EJERCICIO ESTADÍSTICA DESCRIPTIVA 1**

a) Represente la Curva Normal para esos datos indicando la posición y valor de la media y la Desviación Estándar, además de los atributos no paramétricos: rango (incluyendo su amplitud), la moda y la mediana.

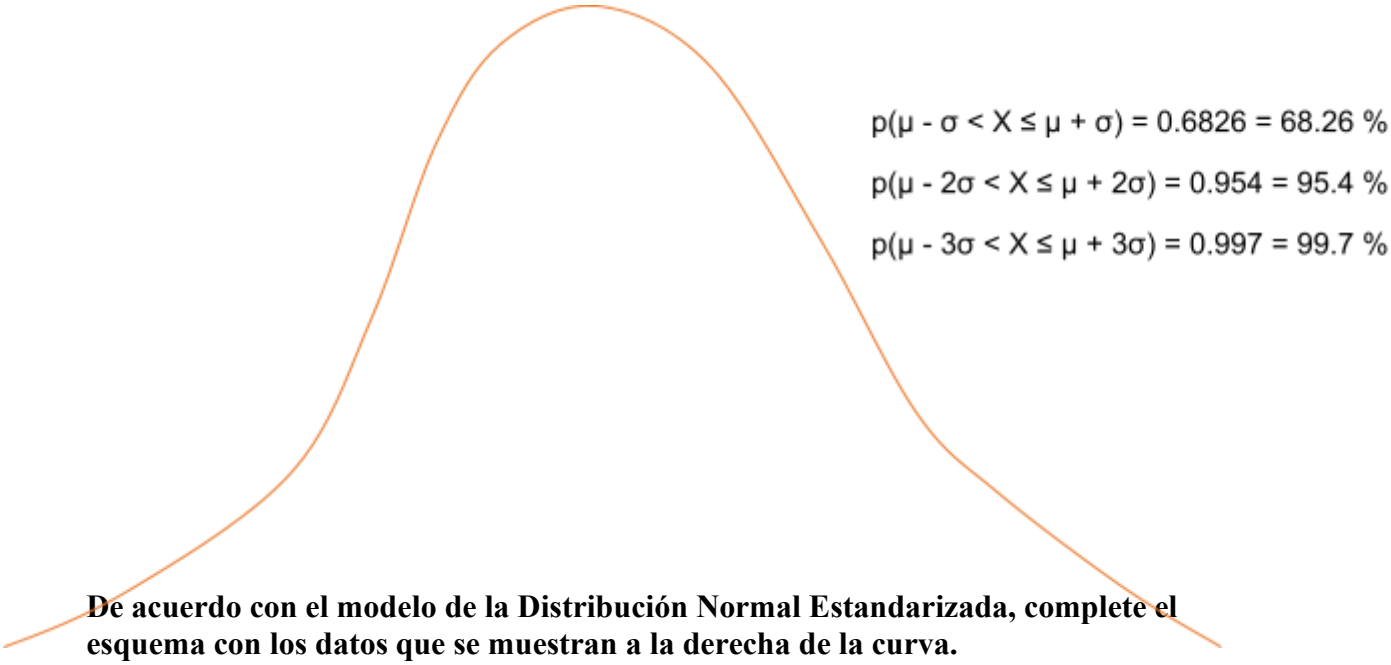
b) De acuerdo con sus datos ubique la posición de los valores: Media + 1 SD; Media + 2SD y Media + 3 SD.

Recordemos:

Cuando se hace que la media de una población sea igual a cero y una Desviación Estándar sea igual a uno, se define una CURVA NORMAL ESTANDARIZADA. Este es el modelo a partir del cual se define la función de Densidad de Probabilidad para la Distribución Normal.

De acuerdo con el gráfico que ya tiene, establezca una probabilidad aproximada de obtener al azar en esa población un valor = Media + 2SD +1 (Use la posición del valor obtenido con respecto a la media)

LOS VALORES Z Y SU UTILIDAD



De acuerdo con el modelo de la Distribución Normal Estandarizada, complete el esquema con los datos que se muestran a la derecha de la curva.

Recordemos:

a) Estandarizar implica que cualquier valor sea transformado en unidades desviación estándar.

Así: Para cualquier valor de *Xi* de una población normal con media μ y desviación estándar σ, se aplica:

$$Z = \frac{X_i -}{}$$

b) De tal forma que un valor Z es la distancia a la que se encuentra un valor con respecto a su media, medida en unidades desviaciones estándar (puede decirse que es la posición en la curva)

c) Al obtener un valor Z podemos conocer la probabilidad de que cualquier dato pueda ser parte de la población estadística definida por una media y una varianza determinada (las provenientes de nuestra muestra).

d) **Cuando se trata de muestras, los parámetros media y desviación estándar se sustituyen por las estimaciones obtenidas a partir de la muestra.**

Hagamos un ejercicio con valores Z

Suponga una población de peces en cautiverio (N=3,000) a los que se midió su longitud promedio = 63 mm y una desviación estándar = 5.4 mm.

Conteste las siguientes preguntas:

1. ¿Qué proporción de la población es mayor de 69 mm ?
2. ¿Qué probabilidad hay de que al azar se mida un pez menor de 71 mm ?
3. ¿Cuántos peces puede haber que midan más de 75 mm?
4. ¿Podremos saber cuantos peces habrá que midan entre 58 y 72 mm?
5. Dibuje la Curva Normal y ubique los valores Z de los que obtendrá respuestas en términos de proporción.

Utilice la función de Distribución de Probabilidades en EXCEL para una Distribución Normal Estandarizada para obtener la probabilidad de una cola y la probabilidad acumulada que requiera para dar respuesta a las preguntas.

Nota: Tenga cuidado al elegir la función, ya que esta puede variar con la versión de Excel que tenga instalada.

Ejercicio Evaluación: Abra el archivo **EJERCICIO ESTADÍSTICA DESCRIPTIVA 2** y conteste lo que se le pregunta. Entregue sus resultados con el formato de informe que su profesor le indique.

DISTRIBUCIÓN BINOMIAL

Recordemos:

a) "Dependiendo de la distribución que se tenga en los resultados de un experimento o muestreo, podremos estimar la probabilidad de que estos resultados se repitan y entonces hacerles predecibles con un cierto grado de precisión o certeza".

b) Si solo hay dos posibles respuestas para un evento, la distribución de los resultados no puede ser normal. Esto genera la Distribución Binomial cuya predictibilidad depende de los resultados que se obtengan en un n número de ensayos independientes.

c) A uno de los resultados posibles se denomina **éxito** y tiene una probabilidad de ocurrencia ***p*** y al otro, fracaso, con una probabilidad ***q = 1 - p***.

d) La función de probabilidad de la distribución binomial, también denominada función de la distribución de Bernoulli, es:

$$p(X=k)=\binom{n}{k}p^k \cdot q^{n-k}$$

$$\binom{n}{k}=\frac{n!}{k!(n-k)!}$$

El número combinatorio

Ejercitemos el uso de esta función para predecir la ocurrencia de un evento determinado.

Enunciado:

Se sabe que en un sistema acuático de los insectos que se atrapan por trampa el 65% de ellos son diferentes

especies de Chironomidae. En una muestra con 8 trampas para insectos en los que se capturaron 85 especies diferentes de insectos ¿Cuál es la probabilidad de que se capturen 15 especies que no sean Chironomidae?

Veamos como resolverlo:

Paso 1. Identificar nuestros datos:

n es el número de pruebas. n= 65

k es el número de éxitos. k= 15

p es la probabilidad de éxito. P= 0.35

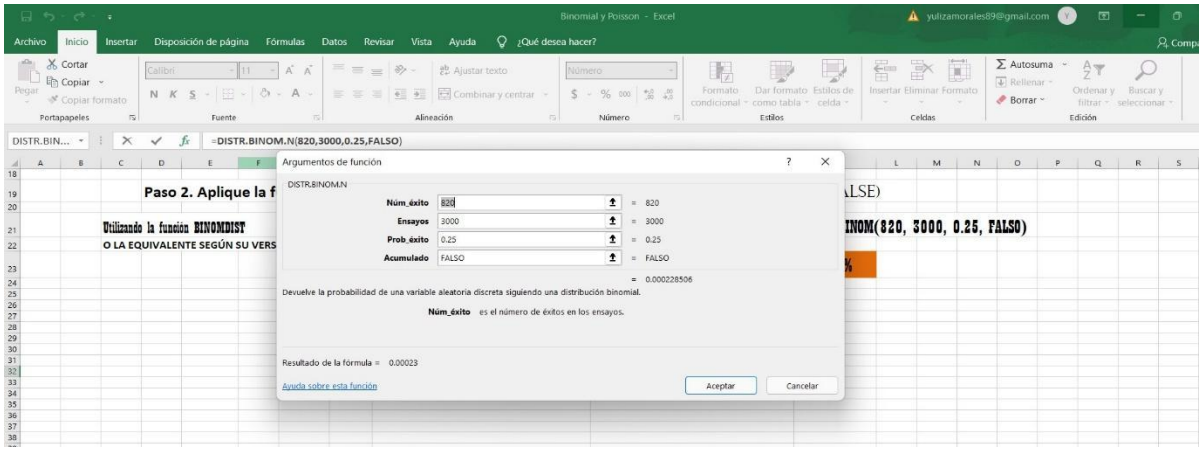
q es la probabilidad de fracaso. q= 0.65

¿Por qué el número de éxitos es 15?

Excel tiene la función para obtener la probabilidad que buscamos calcular, solo recuerden que hay que revisar cuál es la función apropiada en la versión de Excel de que ustedes disponen; para ello debemos buscar que contenga los elementos de la que en este manual se muestra a continuación:

DISTR.BINOM(15,85,0.35, FALSO).

¿Qué ocurrirá con la probabilidad que obtengamos al utilizar la palabra FALSO? ¿Y si usamos VERDADERO?



2. Ahora apliquemos la función:

(La imagen es ilustrativa)

¿Cuál es el resultado?

Interprete acorde con el enunciado del problema planteado.

¿Y si establecemos la restricción de encontrar un máximo de 10 especies que no sean Chironomidae?
Resuélvalo por favor

LA DISTRIBUCIÓN DE POISSON

Recordemos:

a) La distribución de Poisson expresa la probabilidad de un número ***k*** de eventos ocurriendo en un tiempo fijo, **si estos eventos ocurren con una frecuencia media conocida y son independientes del tiempo transcurrido desde el último evento**

b) La función de densidad de probabilidad para la distribución es: $$POISSON = \frac{e^{-\lambda}\ \lambda^{x}}{x!}$$

X= No. de eventos

λ= Media o número esperado
0, 1= define si obtenenos la probabilidad acumulada o no

donde:

Apliquemos la función en Excel:

Enunciado:

Supongamos que se trata del número de incendios que se dan en los bosques de un estado de la República durante el mes más caluroso del año. Por término medio, en este estado se producen 13 incendios al mes.

Interesa saber cuál es la probabilidad de que en un año en particular, el número de incendios en un mes sea exactamente 10.

La función en Excel es: **POISSON DIST=Poisson(10,13,0).**

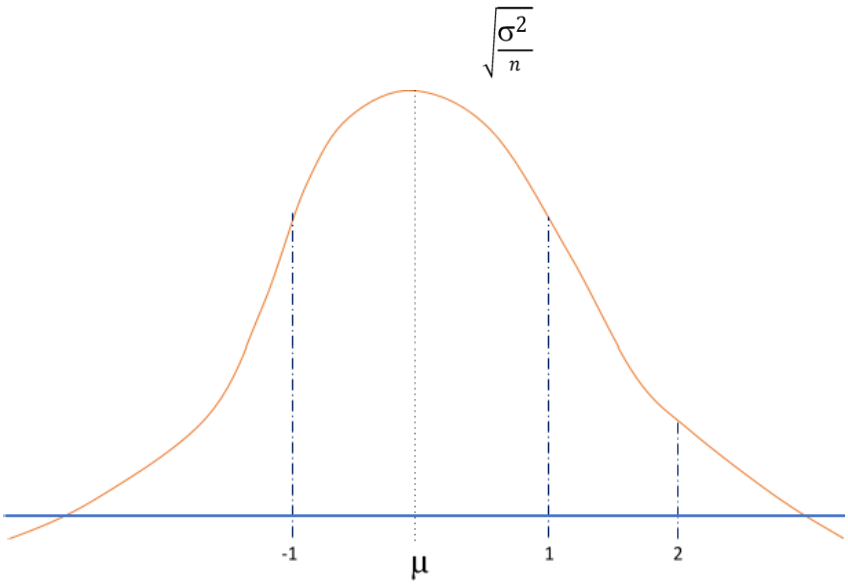
¿Cuál valor debemos usar para obtener la probabilidad acumulada? ¿Cómo se interpreta esto?

EL CONCEPTO DE ERROR ESTÁNDAR Y LOS LÍMITES DE CONFIANZA PARA UNA MEDIA

Recordemos:

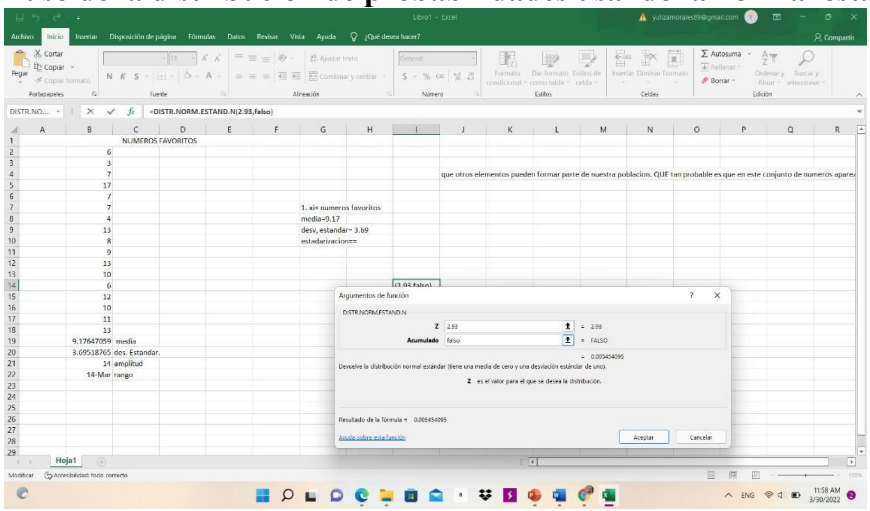
El Error Estándar es el equivalente a la Desviación Estándar en una muestra, SOLO QUE CUANDO HABLAMOS DE ERROR ESTÁNDAR SE TRATA DE LA DISTANCIA A LA QUE SE ENCUENTRA UNA MEDIA CON RESPECTO AL PARÁMETRO (μ) EN EL SUPUESTO DE QUE NUESTRA MEDIA ES SOLO UNA DE TODAS LAS MEDIAS POSIBLES QUE FORMAN PARTE DE ESA POBLACIÓN DE MEDIAS Y NUESTRA VARIANZA ES UN ESTIMADOR DEL PARÁMETRO (σ^2) DE ESA MISMA POBLACIÓN.

Por esa razón con nuestros estimadores podemos estimar qué tanto está alejada nuestra media del parámetro ("media verdadera") y con ello concluir sobre qué tan "buena" es nuestra muestra.



Con el archivo **ESTADISTICA DESCRIPTIVA 2** establezca si la muestra de 30 plantas es
Recuerden que el Error Estándar se calcula:
Ubique en la curva el valor obtenido
suficientemente buena la estimación del tamaño promedio de las plantas, de acuerdo con el valor del Error Estándar obtenido.

El uso de la distribución de probabilidades usando la normal estandarizada en excel



La imagen es ilustrativa para orientar en la búsqueda y uso de la función en Excel.

LA PRUEBA ESTADÍSTICA DE HIPÓTESIS

Recordemos :

Una meta importante del trabajo estadístico es el poder llegar a conclusiones sobre una o más poblaciones (generalmente usando sus medias). Las pruebas más sencillas permiten comparar DOS MUESTRAS.

Se puede iniciar mediante el establecimiento de un enunciado concreto que se desea probar (este enunciado se llama **hipótesis nula (H0:).** Una hipótesis nula debe ser enunciada antes de examinar los datos e incluso más recomendablemente **ANTES de que estos sean colectados.**

Hipótesis nula:

Una hipótesis nula común es que la media de una población no sea diferente de algún valor específico, lo cual se escribiría:

H0: μ=0.75 g por ejemplo. Una respuesta diferente a la H0:, será una respuesta alternativa y por eso se llama **Hipótesis alternativa (HA:)**

ESTE CASO SE CONOCE COMO PRUEBA DE HIPÓTESIS PARA UNA MEDIA HIPOTÉTICA o HIPÓTESIS
PARA UNA MUESTRA (El valor específico contra el que comparamos nuestra muestra)

Para llegar a una conclusión con respecto a nuestra H0:, se requiere de un criterio objetivo para **RECHAZARLA o NO**. Para esto se usa una **PRUEBA ESTADÍSTICA** asociada con alguna distribución de probabilidades (en nuestro curso usamos Z y T student) y el valor de **alfa**, también llamada **SIGNIFICANCIA ESTADÍSTICA**, que en términos de probabilidad se interpreta como: "la probabilidad de que **AL RECHAZAR LA H0: COMETAMOS UN ERROR**". Esto se comoce como la **probabilidad de cometer ERROR TIPO I** (Rechazar H0: cuando **NO DEBIMOS HACERLO**, es decir que en realidad sea acertada.

Por convención (acuerdo):

***P≤ 0.05 (5% o menor*) Se declara**

que H0: se Rechaza. (Preferentemente menor e idealmente claramente

menor)

La variación de nuestra muestra será la que se use para la prueba, bajo el supuesto de que las muestras a comparar pertenecen a la misma población estadística.

Cuando al investigador le interesa solo una dirección de la prueba, esto es probar que la media es *"mayor que"* o "menor que" **y no solo "diferente de", se trata de una PRUEBA DE HIPÓTESIS DE UNA COLA**.

Los *valores críticos* se modifican un poco, así por ejemplo:

Z α=0.05 (2)=1,96, Z α=0.05 (1)=1.645

Ejemplo utilizando la Distribución Z:

Supongamos que usted es responsable de un área silvestre con una importante población de

iguanas cuyo peso promedio, a partir de una muestra de 25 animales fue de 1.78 kg, con una desviación estándar de 0.32 kg. Se requiere vender iguanas que pesen 2.0 kg o más.

¿Cuál sería la probabilidad de que una iguana pesara más de 2.0 kilos? ¿Qué proporción de la población tendría estas características?

Si usa la probabilidad acumulada, recuerde que equivale a usar las dos colas de la curva de DN, con esto en mente, diga ¿cuál es la proporción de iguanas que son menores de 2.0 kg de peso?

El problema de la distribución Z es que requiere de tamaños de muestra "grandes" (30 ó más datos por muestra), por lo que con muestras menores, la confiabilidad disminuye.

La Distribución t de Student

El estadístico William Sealy Gosset(1876-1937), publicó con el seudónimo ''Student'' y generó una prueba basada en una distribución leptokúrtica (a la que llamó T), que se modifica de acuerdo con los grados de libertad, o sea el tamaño de la muestra, lo cual permite comparar muestras relativamente "pequeñas". Cuando las muestras son "grandes", esta distribución es igual que la de Z. A su prueba se le conoce como la **Comparación de Medias de t de Student,** o

simplemente la PRUEBA DE T. $t = \frac{media1 - media2}{Error\ Estándar\ para\ la\ diferencia\ de\ medias}$

El modelo: $\boldsymbol{t = \frac{\bar{X}1 - \bar{X}2}{S_{\bar{X}1-\bar{X}2}}}$

Recuerden que se obtiene a partir de las Sumas de Cuadrados, para conservar la variación de cada una de las muestras

EJERCICIO PRUEBA T PARA DOS MUESTRAS

Ejercicio

Como parte de una evaluación post COVID se hizo un estudio sobre el desempeño de 31 personas midiendo el tiempo que duraban corriendo y el nivel de oxigenación al final del ejercicio.

Se pretende probar si hay diferencia entre hombres y mujeres tanto en el nivel de oxigenación al final del ejercicio, como en tiempo de corrido, independientemente de la

edad.

Datos:

Población Femenina:

Sexo	Edad (años)	Peso (kg)	Nivel Oxigenación	Tiempo de corrido (min)
F	38	81.87	89.06	8.63
F	40	75.98	95.68	11.95
F	42	68.15	94.57	8.17
F	43	85.84	94.30	8.65
F	44	73.03	96.54	10.13
F	45	66.45	94.75	11.12
F	47	79.15	95.27	10.60
F	48	61.24	96.92	11.50
F	49	76.32	91.67	9.40
F	49	73.37	92.39	10.08
F	50	70.87	91.63	8.92
F	51	77.91	92.67	10.00
F	51	67.25	95.12	11.08
F	52	76.32	95.44	9.63
F	52	73.71	92.79	10.47
F	57	59.08	90.55	9.93

N=16

Población Masculina:

Sexo	Edad (años)	Peso (kg)	Nivel Oxigenación	Tiempo de corrido (min)
M	38	89.02	96.87	9.22
M	40	75.07	95.31	10.07
M	43	81.19	92.09	10.85
M	44	89.47	89.61	11.37
M	44	81.42	93.44	13.08
M	45	87.66	90.39	14.03
M	47	77.45	93.81	11.63
M	48	91.63	86.77	10.25
M	49	81.42	89.16	8.95
M	51	69.63	94.84	10.95
M	52	82.78	87.47	10.50
M	54	83.12	91.85	10.33
M	54	79.38	92.08	11.17
M	54	91.63	88.20	12.88
M	57	73.37	93.41	12.63

N= 15

Se requiere:

a) PLANTEAR LAS HIPÓTESIS CORRESPONDIENTES

b) HACER LA PRUEBA DE T (2C)

c) CONCLUIR ESTADÍSTICAMENTE USANDO EL MEJOR RESULTADO EN FUNCIÓN DE LA PRUEBA DE F SOBRE LA IGUALDAD DE VARIANZAS. USAR α=0.05 Y α=0.01.

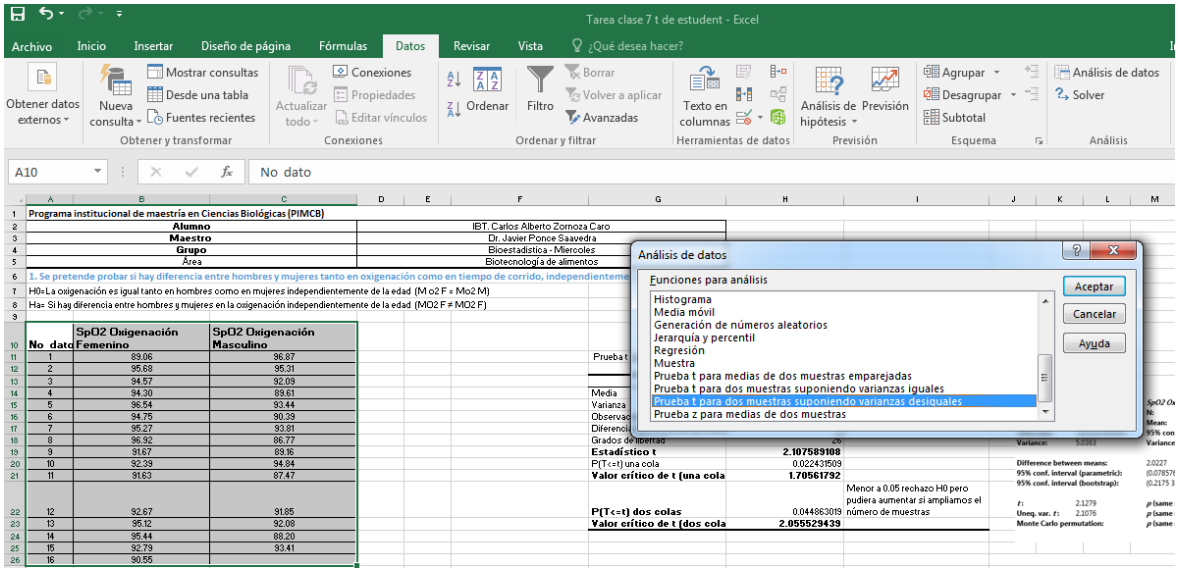
d) CONCLUIR EN FUNCIÓN DEL ENUNCIADO Y EL INTERÉS DEL ANÁLISIS DE LAS MUESTRAS PRESENTADAS, INCLUYENDO LOS LÍMITES DE CONFIANZA PARA LAS MEDIAS

e) INCLUIR GRÁFICOS QUE ILUSTREN LOS RESULTADOS

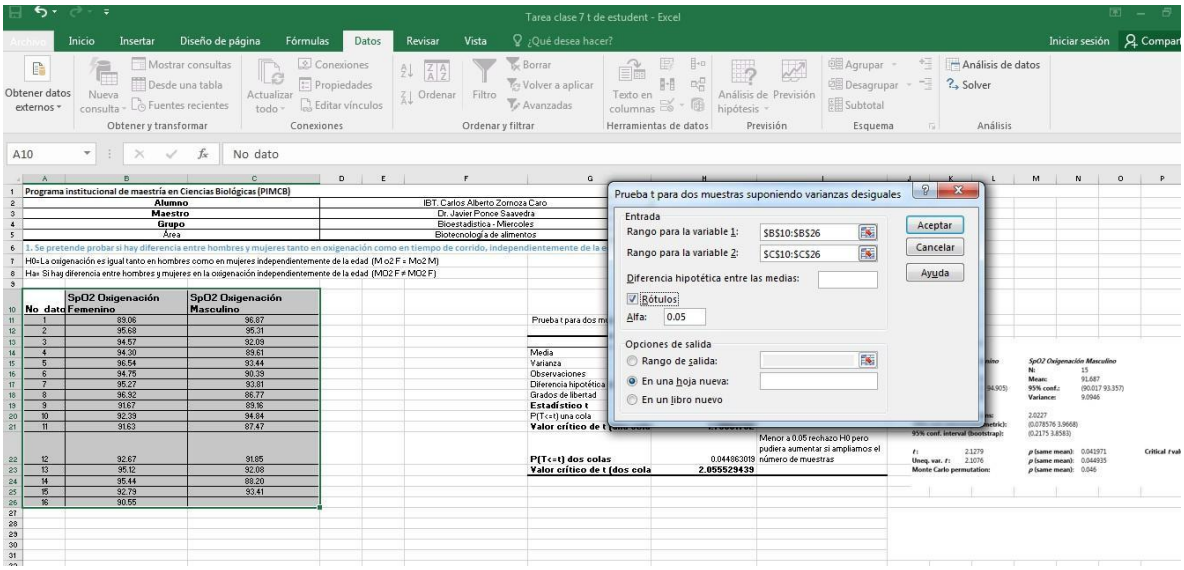
Recordemos:

Primero hay que hacer la prueba de proporción de varianzas de Fischer (F).

En Excel se selecciona la pestaña de **Datos** y seleccionamos la opción **Análisis de Datos**. Seleccionamos prueba de t para varianzas iguales o desiguales, DEPENDIENDO DE LA PRUEBA DE F anterior.



En el cuadro de diálogo seleccionas los datos de cada muestra para cada variable en análisis incluyendo los encabezados (rótulos). La diferencia hipotética entre medias es cero PARA UNA HIPÓTESIS DE DOS COLAS (no colocar nada está bien ya que te lo pondrá por default). Finalmente elegir el rango de salida que puede ser allí en esa hoja o en otra, siempre y cuando evitar encimar datos.



Ahora lleva a cabo el mismo análisis utilizando el Programa **PAST** (Para ver una guía rápida haga **clic aquí**)

Recordemos la "receta":

1. Copiar los datos desde Excel y pegarlos en **PAST** en el modo de edición (con los atributos de columnas y filas activados).

2. Desactive el modo edición y estálisto el modo de ejecución.

3. Seleccione las dos columnas a comparar (misma variable, diferente tratamiento) y en la pestaña UNIVARIATE elija la prueba a realizar.

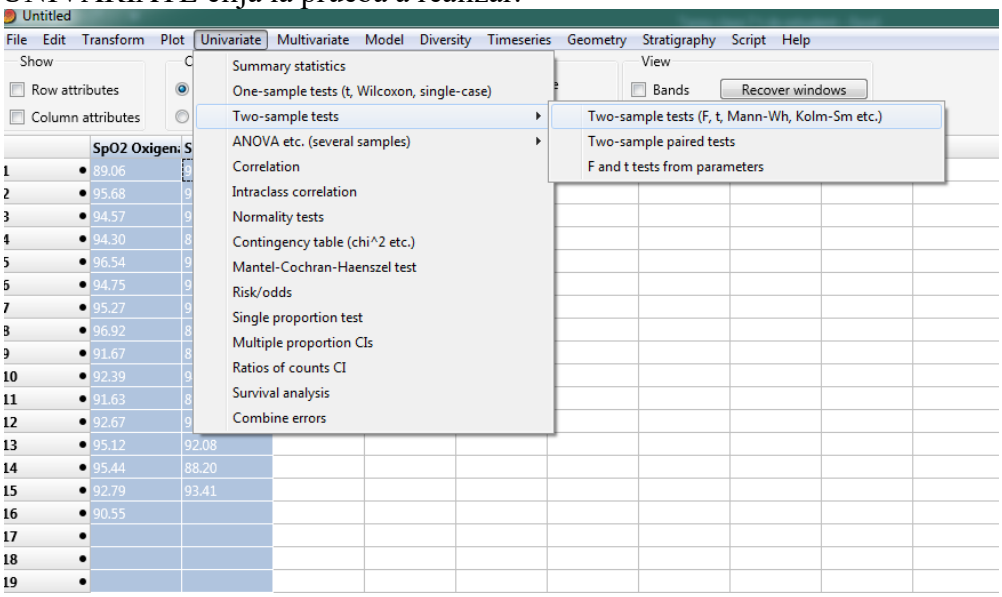
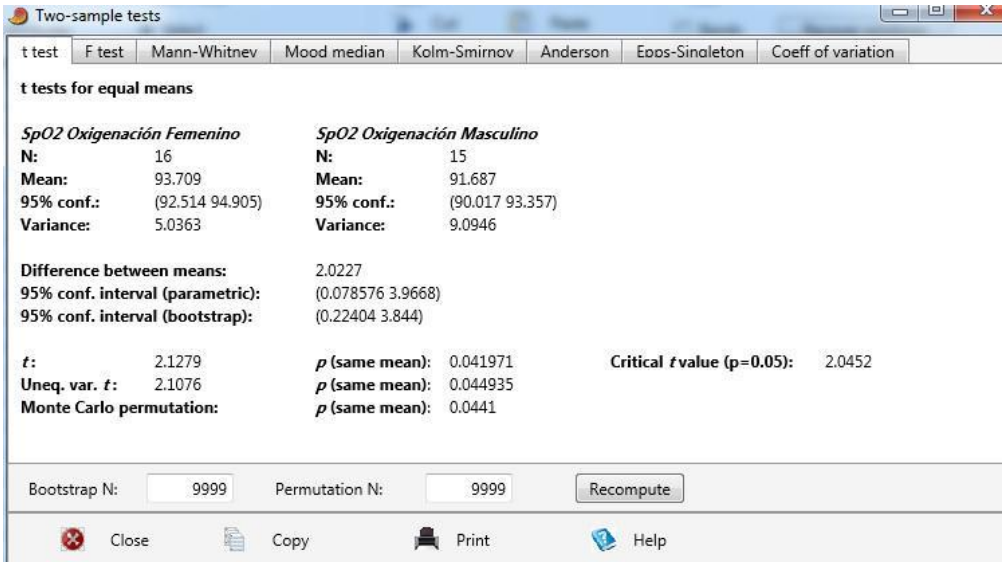


Imagen ilustrativa

En este programa tiene la opción de obtener la prueba de F y así elegir el mejor resultado de su prueba de t.



LA DETERMINACIÓN DEL TAMAÑO DE MUESTRA

Recordemos:

El muestreo es una herramienta para determinar qué parte de una población debemos analizar cuando no es posible realizar un censo. Depende de los objetivos del estudio el elegir una muestra probabilística o no probabilística.

Determinar **el tamaño de la muestra** que se va a seleccionar es un paso importante en cualquier estudio de investigación. Se debe justificar convenientemente de acuerdo con el planteamiento del problema, la población a muestrear, los objetivos y el propósito de la investigación.

MUESTREO PROBABILÍSTICO

Se basa en el principio de **equiprobabilidad**, esto quiere decir que todos los individuos de la muestra seleccionada, tendrán las mismas probabilidades de ser elegidos. Lo anterior ayuda para que la muestra extraída cuente con representatividad.

El muestreo para puede variar dependiendo de cómo se obtengan las muestras:

Muestreo Aleatorio Simple (La selección de la muestra es independiente del investigador.

Muestreo Sistemático (La selección incluye alguna intervención del investigador mediante algún tipo de regla para seleccionar la muestra)

Muestreo Estratificado, combinado con aleatorización por estratos o conglomerados (El investigador decide "dividir" el espacio muestreal de acuerdo con su conveniencia y en cada parte aplicará un muestreo que incluya aleatorización).

El **ERROR** en la muestra siempre existirá y será representado por la **VARIANZA;** mientras que el **NIVEL DE CONFIANZA** es la probabilidad de que la estimación se ajuste a la realidad, es decir que caiga en el intervalo de confianza esperado.

Si se desea estimar en términos de ALPHA (la literatura lo denomina tamaño de muestra para proporciones), habrá dos posibilidades:

a) **QUE SE CONOZCA EL TAMAÑO DE LA POBLACIÓN y**

b) **QUE SE DESCONOZCA EL TAMAÑO DE LA POBLACIÓN**

Ejercitemos ambas posibilidades:

a) CÁLCULO DEL TAMAÑO DE LA MUESTRA CONOCIENDO EL TAMAÑO DE LA POBLACIÓN

La fórmula para calcular el tamaño de muestra cuando se conoce el tamaño de la población es la s

En donde, N = tamaño de la población
Z = nivel de confianza, expresado en valor Z (1-95 en este caso) = Z para un alpha de 0.05 = 1.962. Si la confiabilidad fuera 99%, entonces Z = 2.58)
P = probabilidad de éxito, o proporción esperada
q = probabilidad de fracaso (1-p)
d = precisión (Error máximo admisible en términos de proporción=alfa). Por ejemplo si podemos permitirnos un máximo de 10%, se usa 0.1.

$$n = \frac{N \times Z_a^{\,2} \times p \times q}{d^2 \times (N-1) + Z_a^{\,2} \times p \times q}$$

de la madurez sexual, si sabemos que el

Supongamos la siguiente pregunta:

¿Cuántos peces deberemos revisar para conocer la talla al momento número de adultos de la población en el estanque es de 15,000?

Se establecen como restricciones una confiablidad de 95% y una precisión (alpha) MENOR DE 5%.

La proporción esperada refiere a la probabilidad de éxito, que significa LA PROPORCIÓN DE LA POBLACIÓN DE INTERÉS QUE ESPERARÍAMOS ESTUVIERA REPRESENTADA EN LA MUESTRA !!!!

Recordemos: Si se desconoce la proporción, SE DEBE USAR 0.5 (50%) que maximiza la estimación !!!

Resuelva y de respuesta a la pregunta.

b) CÁLCULO DEL TAMAÑO DE LA MUESTRA DESCONOCIENDO EL TAMAÑO DE LA POBLACIÓN, PERO CONOCEMOS LA PROPORCIÓN

La fórmula para calcular el tamaño de muestra cuando se desconoce el tamaño de la población es la siguiente:

En donde

Z = nivel de confianza

p = probabilidad de éxito, o proporción esperada

q = probabilidad de fracaso

d = precisión (error máximo admisible en términos de proporción)

$$n = \frac{Z_a^{\,2} \times p \times q}{d^2}$$

Hagamos el siguiente ejercicio:

Determinar el tamaño de muestra necesario para estimar la altura promedio de una población de venados silvestres adultos de los que se desconoce el tamaño de la población, pero se sabe que el 16% de la población son adultos. Se quiere 95% de confiabilidad y una precisión para la estimación menor a 10%.

ESTIMACIÓN DEL TAMAÑO DE MUESTRA MEDIANTE LOS LÍMITES DE CONFIANZA

Si nos hiciéramos la siguiente pregunta:

¿Qué tan grande debe ser una muestra para alcanzar determinada precisión al estimar la media de una población?

La respuesta obviamente está relacionada con los intervalos de confianza para la media, por lo que esto se debe considerar en la estimación.

Recordemos la ecuación para el intervalo de confianza: $\bar{X} \pm t_{\alpha(2)\nu} S_X$

Si tomamos la parte positiva y hacemos que esto sea igual a la letra *d*: $d = \bar{X} + t_{\alpha(2)\nu} S_X$

(De ahí surge la fórmula en la que *d* = La mitad del intervalo de confianza (en términos de precisión), lo que significa que μ es estimada para caer entre ± *d*.)

Entonces:

$$n = \frac{S^2 t^2_{\alpha(2),(n-1)}}{d^2}$$

Ahora, el tamaño de la muestra necesario para obtener la precisión deseada depende de:

1. La anchura del intervalo de confianza (representada por *d*)

2. La variabilidad en la población (estimada por S^2)

3. La distribución de probabilidad de t de Student

El problema con la ecuación es que n aparece en ambos términos y para resolverla se utiliza un **método ITERATIVO**.

(Iteración: Proceso de ensayo y error que proporciona cada vez resultados más precisos)

Recordemos que se inicia con un primer número arbitrario* (*Se recomienda sea un número relativamente grande para obtener una primera aproximación*)** y ***con el número obtenido, se hace una nueva estimación para obtener una segunda estimación y así sucesivamente hasta

que la ecuación muestra que la igualdad se cumple ya que la estimación es muy próxima al número utilizado. Entonces ya se puede tomar una decisión.

Hagamos el ejercicio correspondiente:

En un experimento con suplementos alimenticios que ayudan a disminuir peso, se tomo una muestra aleatoria de personas que los consumieron durante 60 días. Se midió la diferencia en peso al inicio y final del tratamiento y los resultados son los siguientes:

Reducción de peso: **0.26, -0.54, -1.33, -1.62, -0.77, 0.42, -0.18, 0.02, -0.61, -1.18, -1.29, -0.89, -0.52, -0.79,**
-0.52, -0.21, -0.32, -0.39, -0.41, -0.87, 0.21, 0.17, 0.09, 0.23, -0.07

Se requiere un tamaño de muestra que garantice una *n* con un intervalo de confianza de 95% y que el error con respecto al parámetro no sea mayor de 0.35 kg a ambos lados de la media.

¿Qué datos necesitamos?

$\alpha =$ $d =$ $s^2 =$ ***n* inicial =**

Aplique el método iterativo y obtenga la respuesta sobre el tamaño mínimo de muestra necesario para obtener las características deseadas en nuestro muestreo.

ANEXO: Usando R

Para usar la paquetería R para elaborar histogramas y obtener la estadística descriptiva, se puede usar la siguiente información.

Se debe tener instalado el programa Rstudio.

Les anexamos algunas instrucciones básicas para usar este software con este fin, elaboradas por el M.C. Carlos Alberto Hernández Luna.

#En cada línea, se mostrará cuál

es la sintaxis del comando,

#Cómo, para qué y como se

utiliza.

#Para empezar, se menciona la función del símbolo

"#". Este sirve para inhabilitar #una línea de

Este es el caso. Podemos leer, pero para el programa no es ninguna instrucción ejecutable.

comando, lo que es muy útil para hacer alguna

anotación en algún #comando y no olvidar para que

usa.

#Para poder cargar un archivo en R, se pueden utilizar dos comandos:

#**setwd()** o **getwd(),** y entre ambos paréntesis

y entre comillas ("") se pone #la ubicación de

la carpeta que contenga el archivo. Ej.:

setwd("*C:/Users/dexte.LAPTOP-0JLG8NFL.000/OneDrive/Escritorio/Estadística Multivarada/Estadística en R/R Studio/Para_compartir***")**

#Pueden copiar la dirección de la ubicación.

#Generalmente, cuando uno le da "copiar-pegar" a la

ubicación de la carpeta, el #separador que aparece es

"\", y uno debe cambiarlo por "/". Ej.:

#C:\Users\dexte.LAPTOP-0JLG8NFL.000\OneDrive\Escritorio\Estadística Multivarada\Estadística en R\R Studio\Para_compartir

#cambia por C:/Users/dexte.LAPTOP-0JLG8NFL.000/OneDrive/Escritorio/Estadística Multivarada/Estadística en R/R Studio/Para_compartir

#Ya con la carpeta especificada, sigue cargar el

archivo. En general, R puede trabajar #con archivos

.xlsx, .xls, .txt o .csv, y cada extensión tiene un

comando específico. #Para abrir un archivo **.txt**

separado por tabulaciones se usa **"read.delim".**

#Veamos la siguiente instrucción:

Gorgosaurus_txt <- read.delim("C:/Users/dexte.LAPTOP-0JLG8NFL.000/OneDrive/Escritorio/Estadística Multivarada/Estadística en R/R Studio/Para_compartir/Gorgosaurus_dataset_complete.txt")

#Para abrir un archivo en excel se puede hacer

de varias formas: la primera es: #**la pestaña en la**

parte derecha de la pantalla llamada

"Enviroment",

#**en el apartado de "Import Dataset", selecciona la opción "From Excel"**

#**La segunda forma es con el comando**

read_excel("Nombre del archivo.xlsx") #La

primera palabra (en estos ejemplos "Gorgosaurus" o

"Gorgosaurus_txt) seguida #de la **<-** es un nombre

que se le da a la base de datos, para poder usarla en

los análisis #Entonces quedaría así:

Gorgosaurus <- read_excel("Regresión_Gorgosaurus.xlsx")

#Con el comando **View(Nombre del objeto)** se pueden

ver la tabla que se cargó #Ejemplos:

View(Gorgosauru

s)

View(Gorgosauru

s_txt)

#Note que las instrucciones que implican una acción no

llevan ningún símbolo ni antes #ni después.

#Ya con las bases de datos cargadas, se pueden comenzar los análisis.

#La mayoría de los análisis de estadística básica no

requieren de ninguna paquetería #especial, por lo que se

pueden hacer fácilmente.

#Veamos como:

#Para calcular promedio se usa el comando **mean**, y la sintaxis es

#**mean(nombre de la base de datos$nombre de la columna donde están los datos de los que se obtendrá la media)**

#Así, si queremos la media de la variable CH en la base de datos Gorgosaurus escribimos:

Media_general_CH<-mean(Gorgosaurus$CH)

#donde el símbolo **$** sirve para indicar que columna es

de la que se quiere obtener el valor #**Para seleccionar**

una serie de datos específicos en una base de datos

#**se usan corchetes**, añadiendo nuevamente el nombre de la base de datos

#, seguido del símbolo de dinero con la columna de interés, **un doble** signo de igual ==

#y entre comillas van los datos

que son de interés. #Ejemplos:

Media_adultos_CH<-mean(Gorgosaurus_txt[Gorgosaurus_txt$Edad=="Adulto", "CH"])

Media_adultos_CH<-mean(Gorgosaurus_txt[Gorgosaurus_txt$Edad=="Juvenil", "CH"])

#Recordemos que lo que está en verde es solo el nombre con el que se identificará ese resultado

#o conjunto de resultados.

#Para el Histograma de frecuencias se usa el comando

hist(nombre de la base$columna)

Histograma_general_CH <-hist(Gorgosaurus$CH)

#para personalizar la gráfica se agrega al código los comandos **main = "con el**

título que se quiere en la gráfica"; **#col = "color en el que se quiere que**

salgan las columnas",

#ylab = "Nombre del eje y", #xlab = "nombre del eje x",

#xlim = c(Escala de X), xlim = c(Escala de Y)

#Ejemplo: Histograma_general_CH <- hist(Gorgosaurus$CH,

main = "Altura de la corona", col = "blue",

ylab = "Frecuencia", xlab = "Altura (mm)",

xlim = c(0, 100), ylim = c(0, 20))

windows()

Esta última instrucción hace que se genere la gráfica.