



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 <u>RESALTE</u> 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.

Untitl File Edi Show	led it Transform	Plot Univa	iate Multivariate de	Model Edit	Diversity Tir	neseries Geon	netry Stratigra View	— I phy Script	⊐ × Help
Row 🗹	attributes mn attributes	 Select Drag ro 	ows/columns	*	Cut 📔 Copy 🧿	Paste Select all	Bands Binary	Recover Decimals:	windows
	A	в	с	D	E	F	G	н	I.
Туре	-	-	-	-	-	-	-	-	-
Name	A	в	С	D	E	F	G	н	1
1	•								
2	•								
3	•								
4	•								
5	•								
6	•								
7	•								
8	•								
9	•								
10	•								
11	•								
12	•								
13	•								
14	•			_					
15	•			_					
16	•								
10									
19									
20	•								
				1				1	

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".

w attributes	۲	One-sample tes	ts (t, Wilcoxon, sing	le-case)	5	Bands	Recover	windows									
lumn attributes	0	Two-sample tes	ts		> :t all	🗌 Binary	Decimals:	- ~									
Peso (g	ramos B	ANOVA etc. (sev	eral samples)		· ·	G	н	1	J	к	L	м	N	0	Ρ	Q	R
-	-	correlation lateral and second				-	-	-	-		-	-	-	-	-	-	-
Peso (g	ramos) B	Intraciass correl	ation			G	н	1	J	К	L	M	N	0	P	Q	R
• 240		Normality tests															
• 369		Outlier tests															
• 372		Contingency tab	ole (chi^2 etc.)														
• 415		Mantel-Cochran	-Haenszel test														
• 422		Risk/odds															
• 454		Single proportion	on test														
• 476		Multiple propor	tion Cls														
• 503		Ratios of counts	CI														
• 507		Survival analysis															
• 515		Combine errors															
• 569																	
• 616																	
 620 																	
 637 																	
 689 																	
• 716																	
• 738																	
• 750																	
• 756																	
• 764																	
• 767																	
• 768																	
• 773																	
• 776																	
• 776																	
• 779																	
• 785																	
• 786																	
• 790																	
• 791																	
• 799																	
- 155																	

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.

Tests for nor	mal distribution	-		anes ocor	View	aprily series	i icip									
	All			Paste	Bands	Recover	windows									
N	75			Select all	Binary	Decimals:										
Shapiro-Wilk W	0.9705				onlary			1.	1.			1				1.
p(normal)	0.07758			F	G	н	1	J	ĸ	L	м	N	0	Р	Q	R
Anderson-Darli	ng A 0.8871			-	-				-		-	-	-	-	-	-
p(normal)	0.02219			F	G	н		,	K	L	M	N	0	P	Q	ĸ
p(Monte Carlo	0.0224			-			-									
Lillietors L	0.1303															
p(normal)	0.001425			-												
p(Monte Carlo	0.0034															
Jarque-Bera JB	1.839			-												
p(normal)	0.3987							_			_					
p(Monte Carlo	0.2913			-												
								-								
Copy	🚔 Print	Monte Carlo N:	9999													
	~		-													
Clos	e 🌾 Help		Recompute													
-																
• 637	_															
• 639																
• 7/6																
• 738																
• 750																
• 756																
• 764																
• 767																
• 765																
• 778																
• 776																
• 776																
• 779																
• 785																
• 786																
• 790																
• 791																
• 799																

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"

🥭 Histogram	ור	Esta es la	- п	
Histogram	Numbers	frecuencia		
Bin start	Bin end	Peso (gramos)	95% CI lo	95% Cl hi
201	381.83	2	0.24344	6.9774
381.83	562.67	5	1 6499	11.157
562.67	743.5	16	9.5353	24.239
743.5	924.33	24	16.269	32.834
924.33	1105.2	17	10.345	25.344
1105.2	1286	11	5.6657	18.547

Estos son los límites de cada categoría

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 <u>https://www.r-project.org/</u>. 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):

 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 "<u>peces.xlsx</u>" con modificaciones para su uso en R.

_				
E5	× 1	$\times \checkmark f_x$		
	Individuo_pob_1	Peso_gr_pob_1	Individuo_pob_2	Peso_gr_pob_2
	1	240	1	201
	2	369	2	227
	3	372	3	404
	4	415	4	450
	5	422	5	504
	6	454	6	516
	7	476	7	549
	8	503	8	599
	9	507	9	601
	10	515	10	618
	11	569	11	627
	12	616	12	627
	13	620	13	633
	14	637	14	636
	15	689	15	639
	16	716	16	690
	17	738	17	710
	18	750	18	712
	19	756	19	713
	20	764	20	722
	21	767	21	723
	22	768	22	723
	23	773	23	730
	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"

Name			Input File				
Encoding	Automatic	~	1;240;1;201 2;369;2;227 3;372;3;404	<pre>Peso_gr_pob_1;</pre>	Individuo_pob_2;	eso_gr_t ×	ty
Heading		-	5:422:5:504			_	
Row names	Automatic	~	7:476:7:549			_	
Separator	Semicolon	~	9;507;9;601			_	
Decimal	Period	~	11:569:11:627			_	
Quote	Double (")	~	13;620;13;633				w
Comment	None	~	15:689:15:639			-	
na stripos	NA		4				
Strings as	factore		Data Frame				
Sumgs as	actors		Individuo nob 1	Peso ar nob 1	Individuo nob 2	Peso ar +	
				240	1	201 227	
			3 4	372	3	404	nd Euclidean Me
			5	422	5	504	aarch
			7	476	7	549	aich
			9	507	9	601 618	
			11	569	11	627	level/Hierarchic
			13	620	13	633	icres includence
Background Jobs ×			15	689 716	15	639	
			1			•	
					C terrent) (Council	ed Since R-3.0.0
					Import	Cancel	
			BH	Boo	ost C++ Header Files		

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.

Flie Edit	Code view Plots Session Build L	Jebug Profile loois Help	10			
de muestre	eo y prueba de hipót 🐐 🛛 🔊 Untitle	ed16 × 👘 peces_peso >	🖻 extraer in » 👝 🗖	Environment Connections Tutorial		-0
(🚛 🍸 Filter		Q,	📹 🚍 🜃 Import Dataset 👻 Memory 👻 🞻		ist - C -
	Individuo_pob_1 [‡] Peso_gr_pob_	1 [‡] Individuo_pob_2	Peso_gr_pob_2	R • 💼 Global Environment +	Q	
1		240	1 201	Data		
2		369	2 227	peces_peso 75 obs. of 4 variables		
3			3 404			
4		415	4 450			
5		422	5 504			
6		454	6 516			
7		476	7 549			
8		503	8 599	File Diete Backanes Hele Viewer Dresontation		
9		507	9 601		0	
10		515 1	0 618	Name Description	Version	
11		569 1	1 627	liser library		0
12		616 1	2 627	abind Combine Multidimensional Arrays	14.5	
13		620 1	3 633	ade4 Analysis of Ecological Data: Exploratory and Euclidean Methods	1.7-22	
14	14	637 1	4 636	in Environmental Sciences		
15		689 1	5 639	agricolae Statistical Procedures for Agricultural Research	1.3-5	
16		716 1	6 690	AlgDesign Algorithmic Experimental Design	1.2.1	•
ch	a second are send as a set of second			ape Analyses of Phylogenetics and Evolution	5.7-1	•
snowing	i to 17 of 75 entries, 4 total columns			Data Analysis Using Regression and Multilevel/Hierarchical	1.13-1	•



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

Histograma y Curva de Distribución Normal



 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":

Console	Terminal ×	Background Jobs	×	
🤦 R4	.1.3 · ~/ 🗭			
T				εακο-ι
		Varl	Freq	
1		(240,389]	2	
2		(389,538]	7	
3		(538,687]	4	
4		(687,836]	24	
5		(836,985]	19	
6	(985,	1.13e+03]	12	
7 (1.	13e+03,	1.28e+03]	6	
>				

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)))

```
Console Terminal & Background Jobs *

    R4.1.3 ~/ 
    Var(peces_peso$Peso_gr_pob_1)
[1] 48002.84
    sd(peces_peso$Peso_gr_pob_1)
[1] 219.0955
    sd(peces_peso$Peso_gr_pob_1)/(sqrt(length(peces_peso$Peso_gr_pob_1)))
[1] 25.29897
```

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

 $p(\mu - \sigma < X \le \mu + \sigma) = 0.6826 = 68.26 \%$

 $p(\mu - 2\sigma < X \le \mu + 2\sigma) = 0.954 = 95.4 \%$

 $p(\mu - 3\sigma < X \le \mu + 3\sigma) = 0.997 = 99.7 \%$

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:

$$Z = \frac{X_i - \dots}{1 - 1}$$

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.

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



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:

<i>n</i> es el número de pruebas.	n= 65
<i>k</i> es el número de éxitos.	k= 15
<i>p</i> 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?

- B- 5	 C² 								Bino	omial y Poiss	on - Excel	<u>ji</u>				-1	🔰 yuliza	morales	39@gmail.o	:om 🕜		-	0
Archivo	Inicio	Insertar	Disposi	ición de página Fó	rmulas Datos Re	visar Vista	Ayuda	💡 įQue	é desea ha	icer?													24 Comp
Pegar %	Cortar Copiar Copiar fo papeles	ত ormato চ	Calibri N K	- 11 - § - 1⊞ - 1⊘ Fuente		- & - 1 1 1 Aline	환 Ajustar 1 문 Combin ación	texto ar y centrar	- \$	Imero - % 00 Número	- - 	Formato condicional	Dar form - como tab Estilos	nato Estilos ola - celda	i de li	Sertar El	minar Fo	imato	∑ Autos ↓ Reller Ø Borra	iuma ~ har ~ r ~ E	Ar Ordenar y filtrar → dición	Buscar	
DISTR.BIN	*	: ×	× 1	=DISTR.BINOP	M.N(820,3000,0.25,F	ALSO)																	
A A	8	с	D	E F	Argumentos de funció	n							?	×		L	м	N	0	р	Q	R	s
19			Paso 2	2. Aplique la f	DISTR.BINOM.N	Núm_éxito	820			1	= 820				LSE)							
21		Utilizand	la función	n BINOMDIST		Ensayos	3000			±	= 3000				INOM	(820.	3000.	0.25.	FALSO)				
22		O LA EC	UIVALEN	TE SEGÚN SU VERS		Prob_éxito	0.25			1	= 0.25					(,	,	,	,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,				
22						Acumulado	FALSO			1	= FALSO				%								
24 25 26 27 28 29 30 31 32 33 34 35					Devuelve la probabilidad Resultado de la fórmula <u>Ayuda sobre esta función</u>	de una variable N = 0.00023	e aleatoria disc lúm_éxito es	creta siguiend : el número d	do una distri de éxitos en	ibución binon los ensayos.	= 0.00022 nial.	Aceptar	c	ancelar									
36 37 38																							

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.

Anthivo	Inicio	Insertar I	Disposición de	página Fórm	ulas Dates	Revisar V	ista Ayuda	Ç ¿Qué di	tsea hacer?					and the second second				A Compar
Regar	K Cortar Copiar & Copiar	, Internation	K 5 -	- 11 - 1 (ii - [🌣 -	= A · A	= = = 3	君 Ajustai	Texto nary centrar -	General S = % co	- - % 23	Tormato condicional -	Dar tormatio - I como tabla -	istilos de orida -	tuertar Diminar D	antiato	Σ Autosuma Relietar = Borrar =	Art Codenary B filtur - sele	P uscary scionar -
Pot	rtapapeles	8	h	<i>sertie</i>			Mrescón		Nimero			Estilos		Celdes			Edición	
NSTR.N	0	X	/ fz =	DISTR.NORM.E	STAND.N(2.5	(3,falso)												
d.	A	В	C	D	E	F	G	Н	1	J	К	ι	M	N	0	Р	Q	R
			NUMERO	S FAVORITOS														
<u> </u>		6																
		3																
		7								que otros ele	ementos puede	en formar part	e de nues	tra poblecion. QUE	tan proba	ble es que en	este conjunto d	e numeros apar
		17																
		/																
							1. xas numer	os favoritos										
1		4					media-9.17											
		15					desv, estand	lar= 3.09										
							estadarizaci	00==										
1		13																
i –		10							(2.024data)									
1		12					0	Automentos de	harido						2	×		
		10						- 9										
		10						DISTRINGRMEST	AND N									
		11							z	2.93		1	= 2.98					
		9 17647059	martin						Acumulado	falso		•	= FAISO					
		3 69518765	des Estanda	ir.														
		14	amolitud										- 0.0054	154095				
1		14-Mar	rango					ververve la 050/0	ocum normial estanda	er (erene una me	cua ue cero y una	vesneu01 estan	uer ue unoj.					
3										z e	el valor pera el o	que se desea la dr	stribución.					
4																		
5																_		
5							1	tesultado de la fó	mula = 0.00545001	\$5								
1																		
8								toucla sobre estat	lanción					Aceptar	Cark	elar		
3																		
	Ho	oja1 💮										< l						
Indificar	GAcce	subditiant todo co	micho													III. (III.	四	
						-												11.50 444
C						Q E	E 🔒		. 🖪 🖻		φ <u>γ</u> ι	۹ 📲	C 4			~ EN	ic (\$ d) ■	3/30/2022

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 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):

$P \leq 0.05 (5\% \ o \ menor)$

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:

Ζ α=0.05 (2)=1,96, Ζ 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. $t = \underline{medial - media2}$ Error Estándar para la diferencia de medias



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 Oxigenació n	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:

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

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.

Β	∙ ج	¢- ∓						Tarea clas	e 7 t de es	tudent -	Excel						
Arc	nivo	Inicio Insertar	Diseño de página Fórm	nulas Dato	s Rev	visar	Vista										
Obte	ner dato ernos *	os Nueva consulta - Co Fuente Obtener y transf	r consultas una tabla s recientes formar Con	Conexiones Propiedades Editar vínculo exiones	2↓ ∡↓ 0	Z A Z Ordenar	Filtro Ordenar	Borrar Volver a ap Volver a ap Volver a ap Volver a ap	plicar 5 co He	exto en olumnas	in i	Análisis hipótesis Pr	de Previsión s * evisión	Agrupar Agrupar Desagrupa Esquema	• +]] ar • -]] i - 5i	Análisis de ?→ Solver Análisi	e datos
A1	D	* E × V	fx No dato														
	A	В	C	D	E		F		G		н			1	J	K L	м
1	rograma	institucional de maestría en	Ciencias Biológicas (PIMCB)														
2		Alumr	10			IBT. Carlo	os Alberto Z	ornoza Caro									
3		Maest	10			Dr. Ja Ricco	vier Ponce	Saavedra		_						9 8	
5				-		Biotec	scauscica - N spología de	alimentos	Análisis d	e datos							
7 8 9	1∪=Laoxig Ha=Sihay Nodato	enacion es igual tanto en hombr diferencia entre hombres y muje SpO2 Oxigenación Femenino	res como en mujeres independienter res en la oxigenación independienter SpO2 Oxigenación Masculino	mente de la edad (f	102 F≠M0 102 F≠M0	12 F)			Histog Media Genera Jerarqu	rama móvil ición de iía y pero	números ale centil	atorios			•	Cancelar	
11	1	89.06	96.87					Prueba t	Muorte	ion							
12	2	95.68	95.31					_	Prueba	t nara m	nedias de do	s muestras	s empareiadas		-		
13	3	94.57	92.09	_				A4.4.	Prueba	t para d	ios muestras	suponien	do varianzas i	quales	-		
14	4	96.54	93.61	-				Marianza	Prueba	t para d	los muestras	suponien	do varianzas d	lesiguales		1	5p02 0x
16	6	94.75	90.39	-				Observad	Prueba	z para n	nedias de do	s muestra:	s				N
17	7	95.27	93.81					Diferencia									Mean:
18	8	96.92	86.77					Grados de	Ilbertad	_	_	26			Variance:	5.0363	Variance
19	9	91.67	89.16	_				Estadísti	co t		2.1	07589108					
20	10	92.39	94.84	_				P(T<=t) un	a cola			0.022431509			Difference 95% conf	between means: interval (narametric):	2.0227
21		31.63	87.47					Talor crit	tico de t (u	na cola	1	.70561792	Menor a 0.05 rec pudiera aumenta	hazo H0 pero r si ampliamos el	95% conf	2.1279	(0.2175 3 p (same
22	12	32.67	92.08	-				r'[l<=t] t Valor críl	us colas tico de t íd	os cola	2.0	0.044053013 55529439	numero de mues	u as	Uneq. var Monte Ca	t: 2.1076	p (same
24	14	95.44	88.20	-				. alor crit		05 00la	2.0	00020400					p counter
25	15	92.79	93.41														
26	16	90.55															

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.

F	• ب	<i>Ĉ</i> - ∓				Tarea clase 7 t de	estudent - Ex	cel						⊠ – ∂
4	rchive	Inicio Insertar	Diseño de página Fórmu	ulas Datos Revisar	Vista								Iniciar sesio	in 🗣 Compai
OŁ	etener date	os Nueva Nueva consulta - Co Fuente Obtener y transf	ir consultas una tabla is recientes formar	Conexiones Propiedades Editar vínculos xiones	ar Filtro Ordenar y	To Borrar To Volver a aplicar To Avanzadas filtrar	Texto en columnas B Herramientas	I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I I	s de Previsión sis * Previsión	현문 Agrupar - 현문 Desagrupar 문문 Subtotal Esquema	*≣ A - 1∃ ?, S	nálisis de d olver Análisis	atos	
A	10	• : × 🗸	$f_{\!\! m x}$ No dato											
1 2 3 4 5	A Programa	B Institucional de maestría en Alumr Maest Grup Área	Ciencias Biológicas (PIMCB) no ro o	D E IBT.C Dr Bio Bio Bio	F Javier Ponce S Sestadística - M tecnología de a	G avedra iercoles imentos	Prueba t pa Entrada Rango pa	н ra dos muestras : ara la variable <u>1</u> :	suponiendo varia	inzas desiguales	Aceptar		MN	O P
6 7 8 9	1. Se pret H0=La orig Ha= Si hay	ende probar si hay diferencia genación es igual tanto en hombi diferencia entre hombres y muje	e entre hombres y mujeres tanto e res como en mujeres independientem res en la oxigenación independientem	en oxigenación como en tiempo d ente de la edad (M o2 F = Mo2 M) ente de la edad (MD2 F ≠ MO2 F)	e corrido, inde	ependientemente de la e	Rango pa Diferencia	ara la variable <u>2</u> : a hipotética entre	SCS10:SCS26 las medias:		Cancelar Ay <u>u</u> da			
10 11 12	No date	SpO2 Oxigenación Femenino 89.06 95.68	SpO2 Oxigenación Masculino 96.87 95.31			Prueba t para dos m	<mark>⊠</mark> <u>R</u> ótule <u>A</u> lfa: [0	0.05						
13 14 15 16 17	3 4 5 6 7	34.30 36.54 34.75 35.27	83,61 93,44 90,39 93,81			Media Varianza Observaciones Diferencia hipotética	Opciones Rango En un	de salida o de <u>s</u> alida: a <u>h</u> oja nueva:		1		nino 94.905)	SpO2 Oxigenación Mase N: 15 Mean: 91.687 95% cont.: (90.017	ulino 93.357)
18 19 20 21	8 9 10 11	96.92 9167 92.39 9163	86.77 89.16 34.84 87.47			Grados de libertad Estadístico t P(T<=t) una cola Valor crítico de t	© En un	libro nuevo	Menor a 0.05 recha	azo H0 pero	95% conf. leterval (ns: metric): bootstrap):	Variance: 9.0946 2.0227 (0.078576 3.9668) (0.2175 3.8583)	
22 23 24 25	12 13 14 15	92.67 95.12 95.44 92.79	91.85 92.08 88.20 93.41			P(T<=t) dos colas Valor crítico de t	; (dos cola	0.0448630 2.05552943	pudiera aumentar s 19 número de muestr- 9	si ampliamos el as	ti 2. Uneq. var. t: 2. Monte Carlo permu	1279 1076 itation:	<i>p</i> (same mean): 0.0419 <i>p</i> (same mean): 0.0449 <i>p</i> (same mean): 0.046	1 Critical rva 5
25 27 28 29 30 31	16	30.55												

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.

Untitled												
File Edit	Transform	Plot	Univaria	te Multivariate	Model	Diversity	Timeseries	Geometry	Stratigraphy	Script Help	0	
Show		-0	Su	mmary statistics					View			
Row at	tributes	۲	Or	ne-sample tests (t,	Wilcoxor	n, single-ca	se)		Bands	Recover wi	ndows	
Colum	n attributes	0	Ти	/o-sample tests			•	Two-s	ample tests (F, t	, Mann-Wh, K	olm-Sm etc.)	
	SpO2 Ox	igen; S	AN	IOVA etc. (several	samples))	•	Two-s	ample paired te	sts		
1	• 89.06	9	Co	orrelation				Fand	t tests from para	ameters		
2	• 95.68	9	Int	raclass correlation	n							
3	• 94.57		No	ormality tests								
4	• 94.30		Co	ontingency table (chi^2 etc	.)						
5	• 96.54		Ma	antel-Cochran-Ha	enszel te	st		L				
5	• 94.75		Ris	k/odds				L				
7	• 95.27		Sir	ngle proportion te	st							
в	• 96.92		M	ultiple proportion	CIs			L				
9	• 91.67		Ra	tios of counts CI				L				
11	• 92.39		Su	rvival analysis				L				
12	• 02.67		Co	mbine errors								
13	• 95.12		2.08		-			1				
14	• 95.44											
15	• 92.79											
16	• 90.55											
17	•											
18	•											
19	•											

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.

t test F tes	t Mann-Whitnev	Mood median	Kolm-Smirnov	Anderson	Epps-Sinaleton	Coeff of variation
tests for eq	ual means					
SpO2 Oxiger	nación Femenino	SpO2 Oxige	enación Masculin	10		
N:	16	N:	15			
Mean:	93.709	Mean:	91.687			
95% conf.:	(92.514 94.905)	95% conf.:	(90.017 93	357)		
Variance:	5.0363	Variance:	9.0946			
Difference b	etween means:	2.0227				
95% conf. in	terval (parametric):	(0.078576 3.9	9668)			
95% conf. in	terval (bootstrap):	(0.22404 3.84	44)			
t:	2.1279	p (same me	an): 0.041971	(Critical <i>t</i> value (p=0.	05): 2.0452
Uneq. var. t:	2.1076	p (same me	an): 0.044935			
Monte Carlo	permutation:	p (same me	an): 0.0441			
B	0000	Permutation N	0000	Rec	ompute	

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).

a = precision (Error maximo admisiole en terminos de proporcion=aira). 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:

$$\overline{X} \pm t_{\alpha(2)\nu} S_X$$

 $d = \overline{X} + t_{\alpha(2)\nu} S_X$

Si tomamos la parte positiva y hacemos que esto sea igual a la letra *d*:

(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 $\pm d$.)

Entonces:



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 S²)
- 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?

 $a = d = s^2 = n$ 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")</pre>

#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"])</pre>

#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.