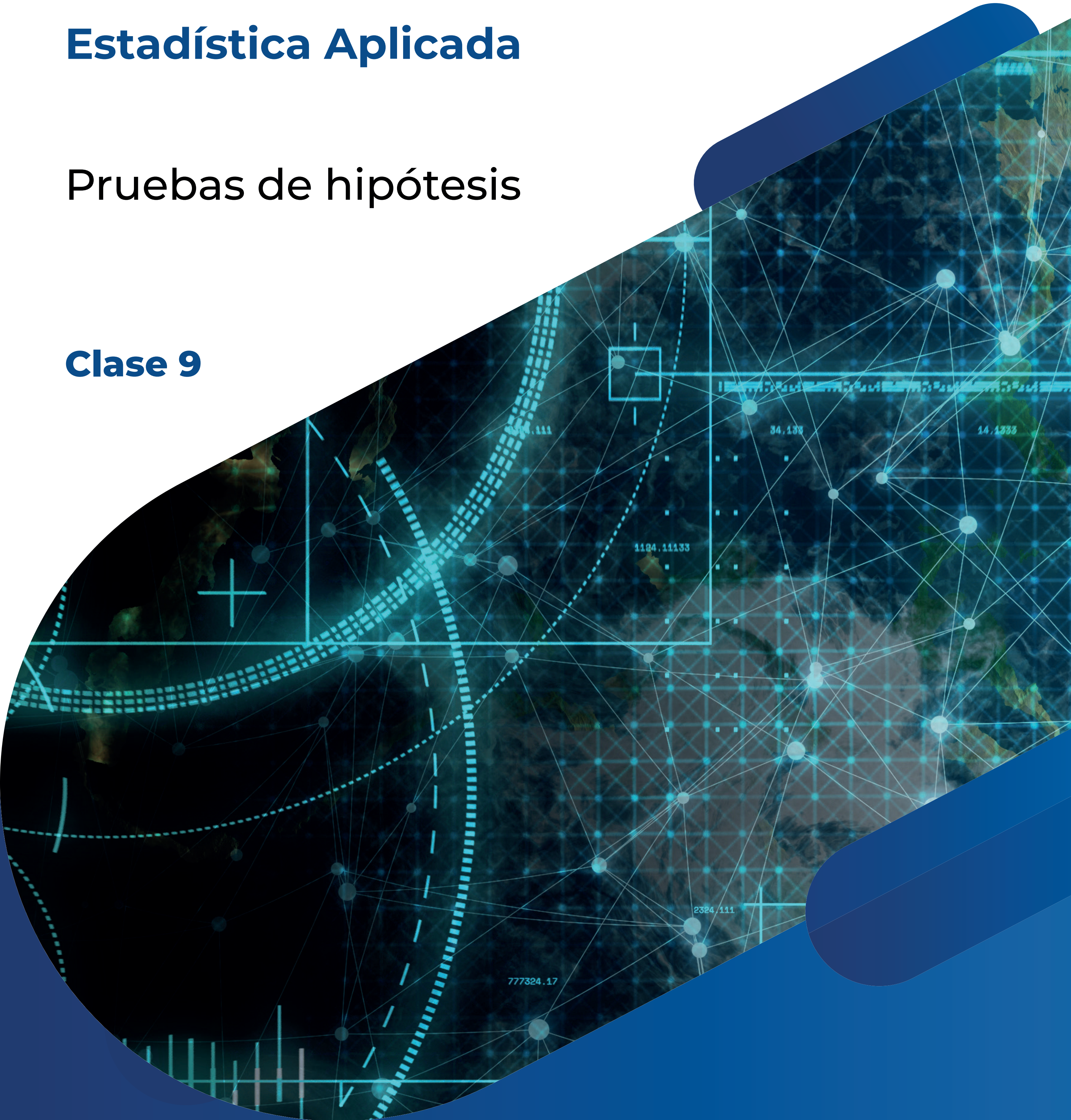
Estadística Aplicada

Pruebas de hipótesis

Clase 9



Ingeniería en ciberseguridad

PUCE Virtual

La excelencia no se improvisa

1. INTRODUCCIÓN DE LA CLASE

En esta clase, vamos a revisar dos temas que son fundamentales en el análisis estadístico. El primero se refiere al concepto de la inferencia. Una vez que hemos comprendido los conceptos de parámetros, estadísticos y distribuciones, nos preguntamos qué inferencias podemos deducir respecto a parámetros de una población en particular o, visto de otra manera, deducir si una observación pertenece o no a nuestra población.

El segundo tema que revisaremos está relacionado con el diseño de experimentos. Como veremos, las inferencias se basarán en las muestras recogidas; sin embargo, si estas no se llevaron a cabo mediante un procedimiento estructurado y metodológico, podrían llevarnos a conclusiones que son erróneas. Por lo tanto, es fundamental para el investigador conocer estas técnicas y aplicarlas correctamente.

Clase 9:

Analizar información contextual sobre hábitat, infraestructura y movilidad, mediante técnicas estadísticas descriptivas e inferenciales, para la adecuada toma de decisiones

Reto # 4

9.) Pruebas de hipótesis

Una hipótesis es una conjetura sobre la forma en que funciona un proceso. Es una explicación tentativa de algún proceso. Antes de estudiar y medir a los individuos en una muestra, un investigador formula hipótesis que predicen cómo deberían verse los datos. En general, una hipótesis predice que los datos no mostrarán nada nuevo o interesante. La llamada hipótesis nula (abreviada H0) sostiene que, si los datos se desvían de la conjetura de alguna manera, esa desviación se debe estrictamente al azar.

De acuerdo con Schmuller (2022) “La otra hipótesis, la hipótesis alternativa (abreviada H1 o Ha), explica las cosas de manera diferente. Según la hipótesis alternativa, los datos muestran algo diferente”.

Para hacer una prueba de hipótesis trabajaremos en base a una o más muestras de las cuales obtendremos las evidencias para indicar si H0 está en lo correcto o no.

Diferencia estadísticamente Insignificante

Es la diferencia entre el valor de la media poblacional bajo hipótesis y el valor de la media muestral que es lo suficientemente pequeña para atribuirla a un error de muestreo, y, por lo tanto, no cuenta como evidencia en contra de H0.

9.1) Estimación para muestras independientes

Hay cinco pasos que debemos completar para probar una hipótesis:

a) Enunciar la hipótesis nula y la hipótesis alternativa.
b) Determinar el nivel de significancia.
c) Calcular el estadístico de la prueba.
d) Determinar el valor o los valores críticos.

e) Enunciar la decisión o el hallazgo.

Hipótesis nula e hipótesis alternativa

Toda prueba de hipótesis debe incluir una hipótesis nula y una hipótesis alternativa. La hipótesis nula, denotada por H0, representa el *statu quo* e implica afirmar la creencia de que la media de la población es ≤, = o ≥ un valor específico. Se cree que la hipótesis nula es verdadera a menos que exista evidencia abrumadora de lo contrario. La hipótesis nula es la que debe rechazarse o no rechazarse.

La hipótesis alternativa, denotada por H1, representa lo opuesto a la hipótesis nula y es verdadera si se determina que la hipótesis nula es falsa. La hipótesis alternativa siempre establece que la media de la población es <, ≠ o > un valor específico.

Errores tipo I y II

Recuerde que el propósito de la prueba de hipótesis es verificar la validez de una afirmación sobre una población, basada en una sola muestra. Como nos basamos en una muestra, nos exponemos al riesgo de que nuestras conclusiones sobre la población puedan ser erróneas debido a un error de muestreo. Aquí se nos pueden presentar dos tipos de errores:

Si rechazamos H0 cuando en realidad es cierto, se conoce como error de tipo I. La probabilidad de cometer un error de tipo I se conoce como α, el nivel de significancia.

Por el contrario, cuando no rechazamos H0 cuando en realidad es falsa, se conoce como error de tipo II. La probabilidad de cometer un error de tipo II se conoce como β.

Calcular el estadístico de la prueba

La fórmula para la prueba de hipótesis está muy relacionada con el estadístico t. En la medida que el valor t calculado aumente, mayor será la evidencia en contra de H0. Sin embargo, si el valor t es menor que el valor del intervalo de confianza utilizado solo indicará que no hay evidencias en contra de H0 no necesariamente que H0 está en lo correcto.

Figura 1

Fórmula para la prueba de hipótesis

$$Z = \frac{\bar{X} - U}{\left(\frac{SD}{\sqrt{n}}\right)}$$

Calcular el valor crítico

El valor crítico divide el área bajo la curva de distribución normal en dos regiones: el área donde no rechazamos H0 y el área o áreas donde rechazamos H0. Obtenemos el valor crítico mediante la función qt(), y esta va a diferir si estamos haciendo una prueba de dos colas o de una cola.

Prueba de dos colas

Utilizaremos esta prueba cuando queremos saber si nuestra muestra proviene de una población en particular dado por una media μ y desviación σ.

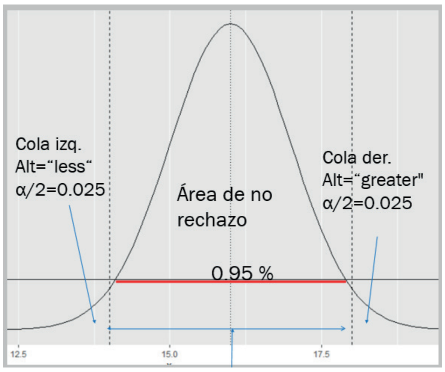
Por ejemplo, tengo una muestra y asumo que esta viene de una población con media de 7,5. Se desea saber si es verdad o no.

Asumimos que si $\overline{X}$ (la media de la muestra) se halla dentro del 95% (1 - α), HO es verdadero, porque solo hay un 2.5% que puede ser atribuido a diferencia insignificante. Por lo tanto, si el estadístico calculado se halla dentro del intervalo de -1.96 y +1.96, HO sigue manteniéndose. Caso contrario, si el estadístico es mayor que 1.96 o menor que -1.96, parece poco probable que la media esté centrada en el valor μ establecido. Por tanto, H1 podría ser verdadero.

Figura 2

Áreas de rechazo para prueba de dos colas

Creación del autor: Alfonso Prado



Es importante mencionar que si el estadístico calculado se hallara dentro de la región de no rechazo, debe interpretarse como que no existe evidencia (datos) que confirmen que H0 está equivocado; es decir, podría ser que otra muestra presente un resultado distinto.

Veamos un ejemplo:

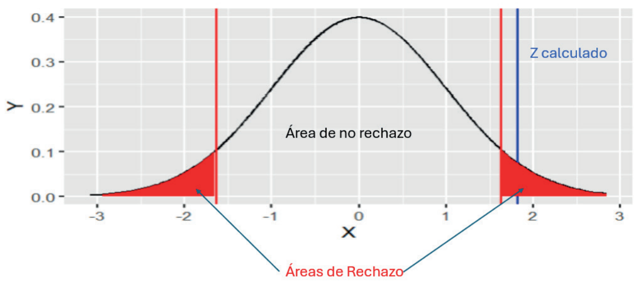
#Usted está interesado en cambiar su carro, y considera que el cambio le costará unos 25.000 US$, realiza una pequeña investigación de mercado en 40 distribuidores y obtiene una media de 27312, con una desviación de 8012 US$. A un nivel de significancia del 10% pruebe si su hipótesis es verdadera.

```
mu=25000
xmedio=27312
n=40
sd=8012
alfa=.1
#Calculamos el estadistico
Z calc<- (27312-25000)/(8012/sqrt(40))
Zcalc
Zcritico=qnorm(0.9 + 0.1/2 )
Zcritico
df4 <- data.frame(X=serie4 , Y=dnorm(serie4, mean=0 ,sd=1))
prob_valor <- dnorm(x, mean=50, sd=10)
```

Figura 3

Gráfico obtenido del problema

Creación de autor Alfonso Prado



En este caso rechazamos la hipótesis H0 debido a que el Zcalculado se halla más distante que el Zcritico

Prueba de una cola

No siempre el problema trata de encontrar si el estadístico está entre las dos colas. Hay veces que el enunciado indica interés en validar solo uno de los lados. Por ejemplo, una empresa puede estar interesada en validar si sus ventas pueden caer por debajo de cierto valor; en otras palabras, las ventas altas no son de interés.

Si tenemos una prueba de una cola, tendremos solo un área de rechazo, no dos. Si es una prueba de cola derecha, entonces el área de rechazo estará en la cola derecha; y si tenemos una prueba de cola izquierda, el área de rechazo estará en la cola izquierda. Veamos cómo obtener el valor crítico para cada una:

Si elegimos α = 0,01 y utilizamos una prueba de cola derecha, entonces necesitaremos determinar el valor Z crítico correspondiente. Debido a que se trata de una prueba de una cola, toda esta área debe estar en una región de rechazo en el lado derecho de la distribución.

Veamos un ejemplo:

```
#El gerente del hotel Embassy Suites Atlanta, reportó que el número promedio

#de habitaciones alquiladas por noche es de por lo menos 212.

#Es decir, mu>= 212. Sin embargo se cree que esta cifra puede estar algo sobrestimada.

#Una muestra de 150 noches produce una media de 201.3 habitaciones y una desvia-
ción estándar de 45.5 habitaciones.

#Compruebe si a un nivel del 1% la hipótesis es correcta

mu=212

n=150

xm <- 201.3

alfa=0.01

sd=45.5

#Calculamos el valor Z al que se encuentra

Zcalc= (xm - mu)/(sd/(sqrt(n)))

Zcalc

#Calculamos el Zcritico

Zcritico=qnorm(0.01 )

Zcritico

#visualizamos

serie4 <-rnorm(1000 , mean=0, sd=1)

df4 <- data.frame(X=serie4 , Y=dnorm(serie4, mean=0 ,sd=1))

ggplot(data=df4 , aes(x=X, y=Y))+

  geom_line()+

  geom_vline(xintercept=Zcalc, color=”BLUE”)   +

  geom_ribbon(data=subset(df4,X>-3  & X< Zcritico ),aes(ymax=Y),ymin=0,

            fill=”RED”)
```

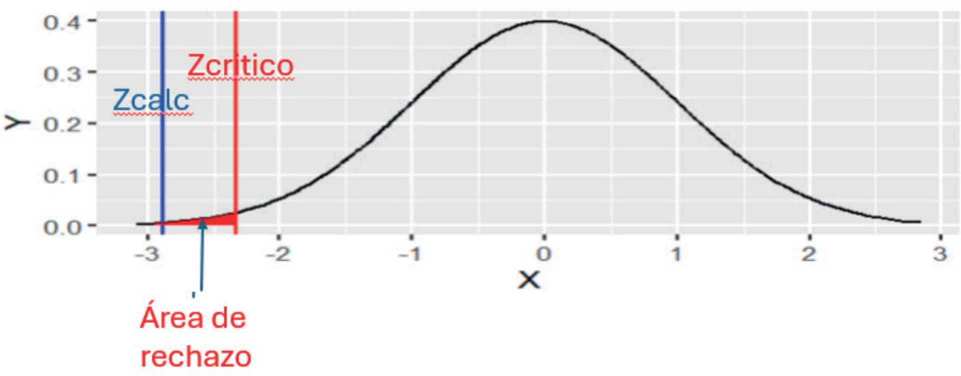
Como muestra la siguiente figura, necesitamos encontrar el valor Z que corresponde al área 1 – α.

Figura 4

Solución al problema

Con un 99% de confianza podemos estar seguros de que la estimación esta sobrestimada

Creación de autor Alfonso Prado



Por otro lado, si el enunciado del problema mencionara que es una prueba de cola derecha, el área de rechazo estará a la derecha y el valor crítico será 2,33 en lugar de -2,33.

Funciones Programáticas t.test

La solución de las pruebas de hipótesis vistas anteriormente podríamos considerarlas como soluciones analíticas; es decir, el analista debe calcular correctamente los valores críticos y tomar una decisión. Las funciones programáticas nos permiten hacer básicamente lo mismo, aunque con ciertas particularidades.

Existen 3 diferentes casos de uso de la función t.test.

Caso 1: Cuando queremos saber si nuestra muestra proviene de una población en particular. En este caso, indicaremos el vector x que será comparado contra el valor mu.

Caso 2: Un dataset con datos dependientes. Por ejemplo, cuando tengo un dataset con valores antes y después de un tratamiento. Requiere el argumento “paired= TRUE”

Caso 3: Dos dataset con datos independientes. Por ejemplo, tengo grupos de hombres y mujeres y queremos determinar si provienen de la misma población. Requiere el argumento “paired= FALSE”

La función t.test devuelve un objeto con 2 propiedades: t-value y p-value. Tvalue es la medida de la evidencia en contra de H0, mientras más grande es el valor desecharemos la hipótesis H0. P-value es simplemente una medida de la probabilidad de que los datos hayan ocurrido por casualidad, suponiendo que la hipótesis H0 sea cierta. Su cálculo depende del escenario planteado. Para el caso de que H0 haya sido rechazado, 1-pvalue será la probabilidad que la hipótesis H1 sea cierta.

Figura 5

Sintaxis de la función t.test

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, …)

# S3 method for formula
t.test(formula, data, subset, na.action, …)
```

Donde:

x, y son las muestras que considerar

mu es la media poblacional estimada

alternative indica cuál es la hipótesis alternativa

conf.level indica el intervalo de confianza para el cálculo

Función t.test

Describe en detalle el uso de t.test

Enlace

Para que t.test funcione correctamente, es necesario que se cumplan 2 condiciones: La distribución debe ser normal y la varianza en el caso de comparación de 2 dataset debe ser similar. Validación de estas premisas veremos en la siguiente clase

9.2) Estimación para muestras pareadas

Se trata de una prueba de un dataset con datos dependientes. Por ejemplo, cuando tengo una muestra que indica el efecto de un tratamiento antes y después. En este caso, una columna contiene la data del paciente antes y otra columna después del tratamiento. Esto lo llamaremos muestras apareadas, y usaremos el argumento “paired = TRUE.” En otras palabras, tratamos de encontrar si existe una diferencia entre dichas columnas.

Veamos un ejemplo:

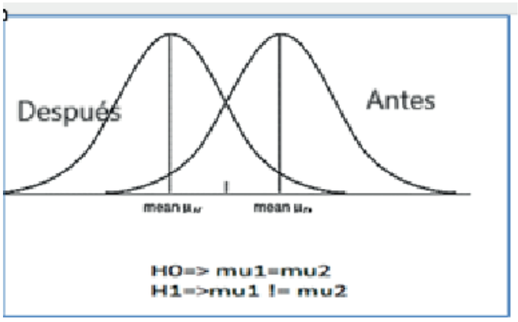
```
#Diez individuos participaron de programa para perder peso corporal por medio de una dieta. Los voluntarios fueron pesados antes y después de haber participado del programa y los datos en libras aparecen abajo. ¿Hay evidencia que soporte la afirmación de la dieta disminuye el peso medio de los participantes? Usar nivel de #significancia del 5%.


antes <- c(195, 213, 247 ,201, 187, 210, 215, 246, 294, 310)

despues <- c(187, 195, 221, 190, 175, 197, 199, 221, 278, 285)


t.test(x=antes, y=despues, alternative=”greater”, mu=0,

      paired=TRUE, conf. level=0.95)


Paired t-test

data:  antes and despues

t = 8.3843, df = 9, p-value = 7.593e-06

alternative hypothesis: true mean difference is greater than 0

95 percent confidence interval:

 13.2832     Inf

sample estimates:

mean difference

           17


#Obtenemos un tvalue de 8.3, es decir el tcalculado está 8 desviaciones

#separado de la media, por lo tanto, hay diferencias indicando que a dieta si

#tuvo efecto
```

El resultado se puede ver en la siguiente figura

Figura 6

Efecto de la dieta en participantes

Creación de autor Alfonso Prado



9.3) Estimación de la potencia de la prueba

Una vez realizado el experimento recuperamos la data y empezamos a inferir si existen diferencias entre los grupos. Sin embargo, vamos a entender que no todas las pruebas tienen igual potencia para inferir los resultados. Más aún, determinados tipos de experimentos van a requerir que se vea una clara distinción sobre la afectación del tratamiento lo cual podría requerir un mayor número de observaciones para lograr el mismo nivel de potencia. Para esto recurrimos a la función de potencia pwr.t.test() del paquete pwr.

La prueba pwr.t.test sirve para establecer los valores de n,d, sig.level, o pwr , esto se obtiene asignando en NULL el argumento que se desea encontrar de acuerdo con los otros argumentos que deben ser no nulos.

Figura 7

Uso de la función pwr.t.test.

Fuente: https://www.rdocumentation.org/packages/pwr/versions/1.3-0/topics/pwr.t.test

```
pwr.t.test(n = NULL, d = NULL, sig.level = 0.05, power = NULL,
    type = c("two.sample", "one.sample", "paired"),
    alternative = c("two.sided", "less", "greater"))
```

Donde:
n: Número de observaciones (por muestra)
d =Tamaño del efecto (d de Cohen):
sig.level: Nivel de significancia (probabilidad de error tipo l)
power: Potencia de la prueba (1 - probabilidad de error tipo ll)
Type: Tipo de prueba t: una, dos muestras o muestras pareadas
Alternative: una cadena de caracteres que especifica la hipótesis alternativa debe ser una de “mayor” o “menor”

Paquete pwr

Describe en detalle las funciones

Enlace

De todos los argumentos, el tamaño del efecto requiere una explicación adicional: El tamaño del efecto puede decirnos qué tan grande es realmente esta diferencia ente los grupos y se calcula mediante la d de Cohen que retorna la diferencia en términos de desviaciones estándar.

De acuerdo con Teck K, (2022), dado que la potencia se define como (1 - probabilidad de error tipo ll) nos indicaría la probabilidad de que la prueba t rechace la hipótesis nula de igualdad de dos medias, asumiendo que la hipótesis nula es falsa y su valor típico es 0.8.

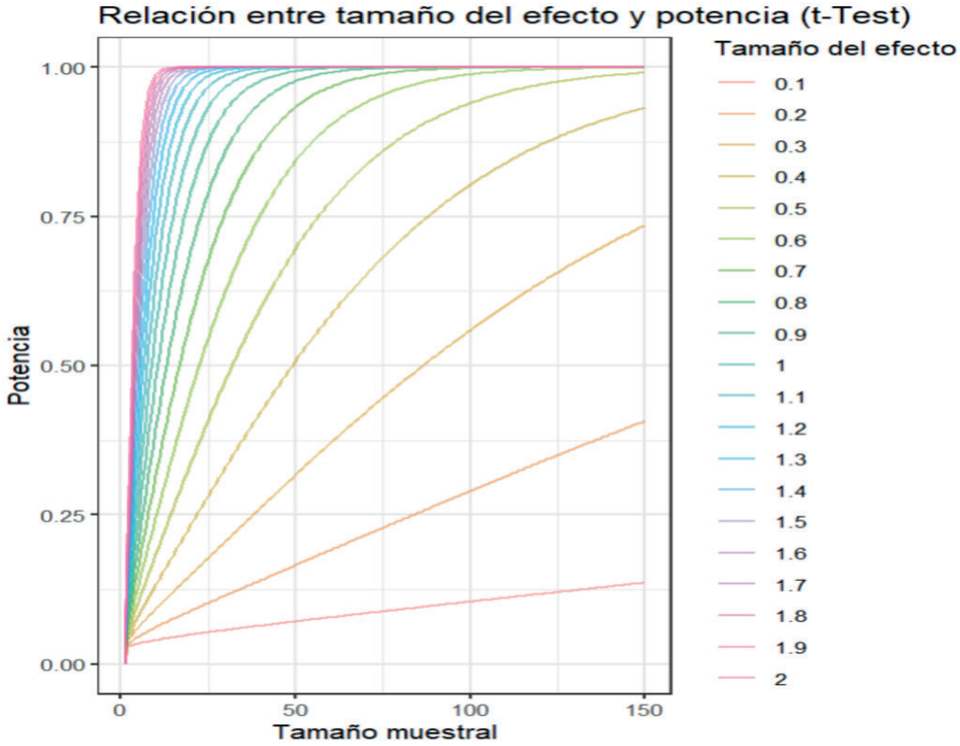
Note que la función no contiene ningún argumento sobre un dataset en especial, la prueba nos da la potencia que se obtendría bajo determinadas condiciones de número de observaciones, tipo de test y tipo de alternativa.

Veamos algunos ejemplos

```
require(pwr)

#Potencia para una sola muestra de 60 observaciones, doble lado ,

pwr.t.test(d=0.2,n=60,sig.level=0.10,type=”one.sample”,alternative=”two.sided”, power=NULL)

power = 0.456


## Muestras pareadas

d<-8/(16*sqrt(2*(1-0.6)))

pwr.t.test(d=d,n=40,sig.level=0.05,type=”paired”,alternative=”two.sided”)

power = 0.932


## Muestras independientes

d<-2/2.8

pwr.t.test(d=d,n=30,sig.level=0.05,type=”two.sample”,alternative=”two.sided”)

power = 0.567

## Cálculo de n para muestras independientes

pwr.t.test(d=0.3, power=0.75, sig.level=0.05,

type=”two.sample”,alternative=”greater”)

n = 120


## Igual que el anterior, pero ahora queremos que pueda detectar diferencias de d=0.1

pwr.t.test(d=0.1, power=0.75, sig.level=0.05,

type=”two.sample”,alternative=”greater”)

n = 1077
```

Figura 8

Explica la relación entre la potencia, el tamaño del efecto (d) y el tamaño de la muestra requerida



Referencias citadas en la Clase 9.
Schmuller J.(2022), Statistical Analysis with Excel For Dummies, Wiley

Teck K, (2022), Practical t-test Power Analysis with R, Practical Assessment, Research & Evaluation, Volumen 27, No 18

Definición de los términos citados en la Clase 9.

<table>
<tr><td>ERRORES TIPO I Y TIPO II</td><td>En la teoría de decisiones, es el error TIPO i es el que se comete al rechazar la hipótesis nula H_0, cuando es verdadera.

El error tipo II, es el error que se comete al aceptar la hipótesis nula H 0 cuando es falsa</td></tr>
<tr><td></td><td>
<table>
<tr><td>DECISIONES POSIBLES</td><td>HIPÓTESIS NULA Ho VERDADERA</td><td>HIPÓTESIS NULA Ho FALSA</td></tr>
<tr><td>Se acepta la Ho</td><td>Correctamente aceptada</td><td>Error de tipo II</td></tr>
<tr><td>Se rechaza Ho</td><td>Error de tipo I</td><td>Correctamente rechazada</td></tr>
</table>
</td></tr>
</table>

REGIÓN DE ACEPTACIÓN y RECHAZO	Es la región formada por el conjunto de valores con los cuales decidimos aceptar la hipótesis nula, el área de rechazo conocida también como región crítica, está formada por el conjunto de valores con los cuales se rechaza la hipótesis nula

Profundización Clase 9.

Pruebas de Hipótesis

Describe funciones programáticas

PRUEBAS DE HIPÓTESIS-V12

PUCE Virtual

La excelencia no se improvisa

síguenos

CV_ICYS_M