UNIVERSIDAD DE LOS ANDES
FACULTAD DE CIENCIAS FORESTALES Y AMBIENTALES
ESCUELA DE GEOGRAFÍA

ESTADISTICA 21

Unidad I. Inferencia con muestras pequeñas

Distribución t de Student Definición. Características. Uso de la tabla t de Student.

En el curso anterior se discutió el uso de la distribución z, la cual se podía utilizar siempre y cuando los tamaños de las muestras fueran mayores o iguales a 30 ó en muestras más pequeñas si la distribución o las distribuciones de donde proviene la muestra o las muestras son normales. En esta unidad se podrán utilizar muestras pequeñas siempre y cuando la distribución de donde proviene la muestra tenga un comportamiento normal. Esta es una condición para utilizar las tres distribuciones que se manejarán en esta unidad; t de student, $\chi 2$ ji-cuadrada y F de Snedecor.

A la teoría de pequeñas muestras también se le llama teoría exacta del muestreo, ya que también la podemos utilizar con muestras aleatorias de tamaño grande.

En esta unidad se verá un nuevo concepto necesario para poder utilizar a las tres distribuciones mencionadas. Este concepto es "*grados de libertad*".

Para definir grados de libertad se hará referencia a la varianza muestral:

$$s^2 = \frac{\sum_{i-1}^{n} (x_i - \bar{x})^2}{n-1}$$

Esta fórmula está basada en *n-1 grados de libertad* (degrees of freedom).

Esta terminología resulta del hecho de que si bien "S^2" está basada en *n* cantidades éstas suman cero, así que especificar los valores de cualquier *n-1* de las cantidades determina el valor restante.

Por ejemplo, si *n=4*

$$(x_1 - \bar{x}),(x_2 - \bar{x}),.....,(x_n - \bar{x})$$

$$(x_1 - \bar{x}) = 8,(x_2 - \bar{x}) = -6,(x_4 - \bar{x}) = -2$$

entonces automáticamente tenemos ,

$$(x_3 - \bar{x}) = 0$$

así que sólo tres de los cuatro valores de estas diferencias están libremente determinados, llamado así , 3 grados de libertad. Entonces, en esta unidad la fórmula de grados de libertad será *(n – 1)* y su simbología ν (letra griega llamada “nu”). También, los grados de libertad se pueden deducir de acuerdo al “numero de parámetros que se desea estudiar”, es decir, si se trata de una media (μ), entonces los grados de libertad sera ν = (n – 1). Si queremos estudiar la diferencia entre dos medias ($\mu_1 - \mu_2$) los grados de libertad se calcularan como ν = (n – 2). Pero si el interés es conocer el promedio de la diferencia entre cambios en el tiempo o espacio (μd) los grados de libertad vendrían dado por ν = (n – 1)

DISTRIBUCION "t DE STUDENT"

La distribución de probabilidad de t se publicó por primera vez en 1908 en un artículo de W. S. Gosset.

En esa época, Gosset era empleado de una cervecería irlandesa que desaprobaba la publicación de investigaciones de sus empleados. Para evadir esta prohibición, publicó su trabajo en secreto bajo el nombre de "Student".

La distribución t normalmente se llama distribución *t de Student*, o simplemente distribución t.

Para derivar la ecuación de esta distribución, Gosset supone que las muestras se seleccionan de una población normal.

La distribución t difiere de la de Z en que la varianza de t depende del tamaño de la muestra y siempre es mayor a uno.

Únicamente cuando el tamaño de la muestra tiende a infinito las dos distribuciones serán las mismas.

Si estamos interesados en estudiar solo el parámetro media (μ)

Supóngase que se toma una muestra de una población que tiene una distribución normal con media **μ** y varianza **σ**.

Si $\boldsymbol{X}$ es el promedio de las *n* observaciones que contiene la muestra aleatoria, entonces la siguiente expresión

$$Z = \frac{X - \mu}{\sigma / \sqrt{n}}$$

es una distribución normal estándar. **μ= 0 y σ^2 = 1**

Supóngase que la varianza de la población σ^2 es desconocida.

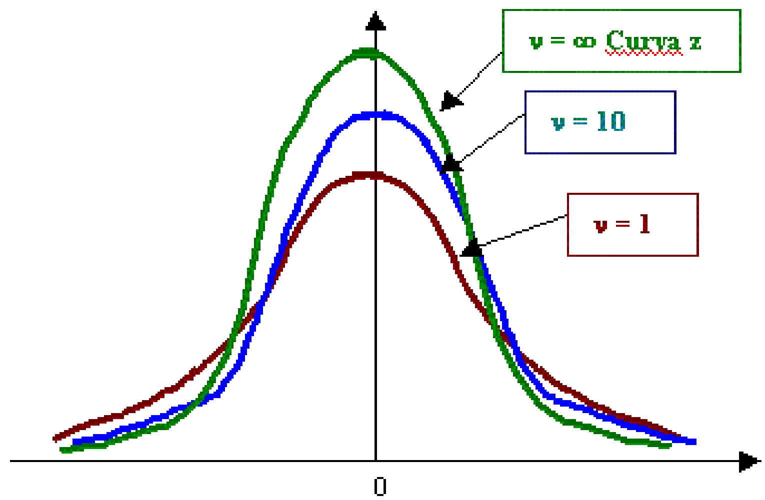
¿Qué sucede con la distribución de esta estadística si se reemplaza σ por ***S***?

$$t = \frac{X - \mu}{S / \sqrt{n}}$$

La media y la varianza de la distribución *t* son

μ= 0 y σ2 = ν/(ν – 2), para ν > 2

La siguiente figura presenta la gráfica de varias distribuciones *t.* La apariencia general de la distribución *t* es similar a la de la distribución normal estándar: ambas son simétricas y unimodales, y el valor máximo de la ordenada se alcanza en la media μ = 0. Sin embargo, la distribución *t* tiene colas más amplias que la normal; esto es, la probabilidad de las colas es mayor que en la distribución normal. A medida que el número de grados de libertad tiende a infinito, la forma límite de la distribución *t* es la distribución normal estándar.



Propiedades de las distribuciones t

1. Cada curva t tiene forma de campana con centro en 0.

2. Cada curva t, está más dispersa que la curva normal estándar z.

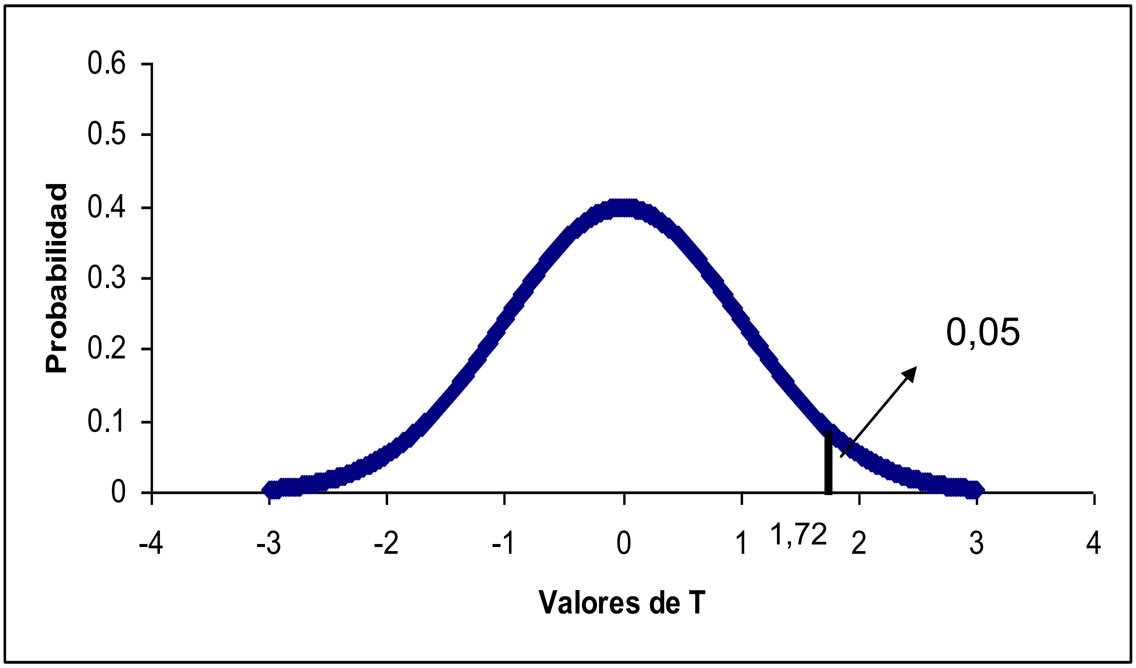
3. A medida que ν aumenta, la dispersión de la curva t correspondiente disminuye.

3. A medida que $\nu \rightarrow \infty$, la secuencia de curvas t se aproxima a la curva normal estándar, por lo que la curva z recibe a veces el nombre de curva t con $gl = \nu = \infty$

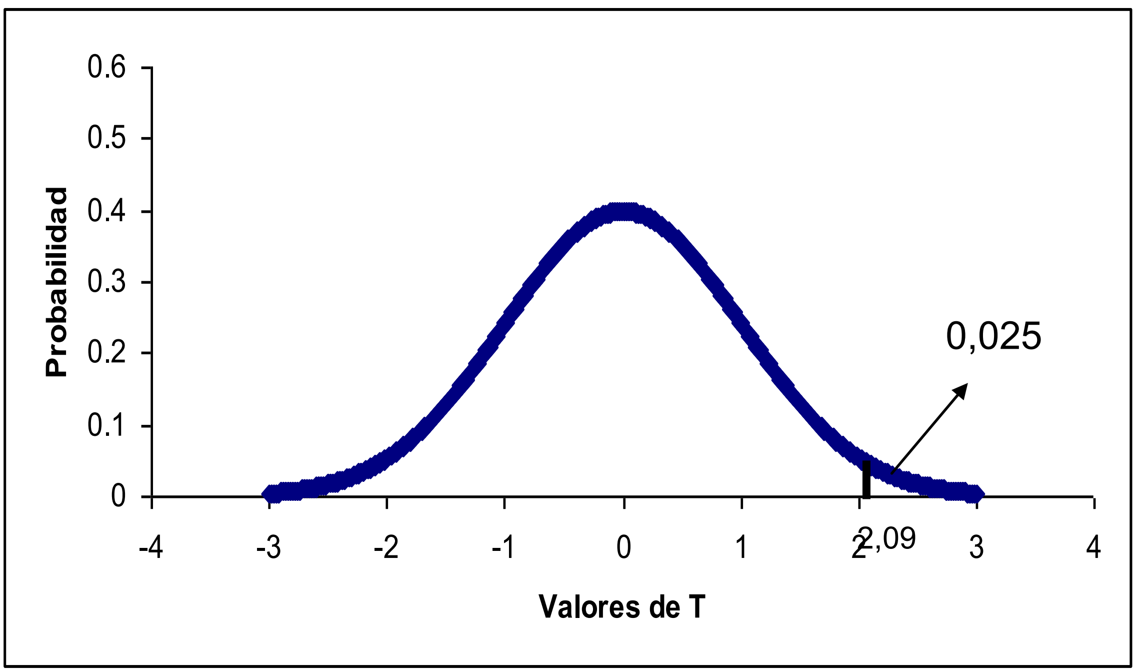
TENER EN CUENTA SIEMPRE:

La distribución t difiere de la de Z en que la varianza de t depende del tamaño de la muestra y siempre es mayor a uno.

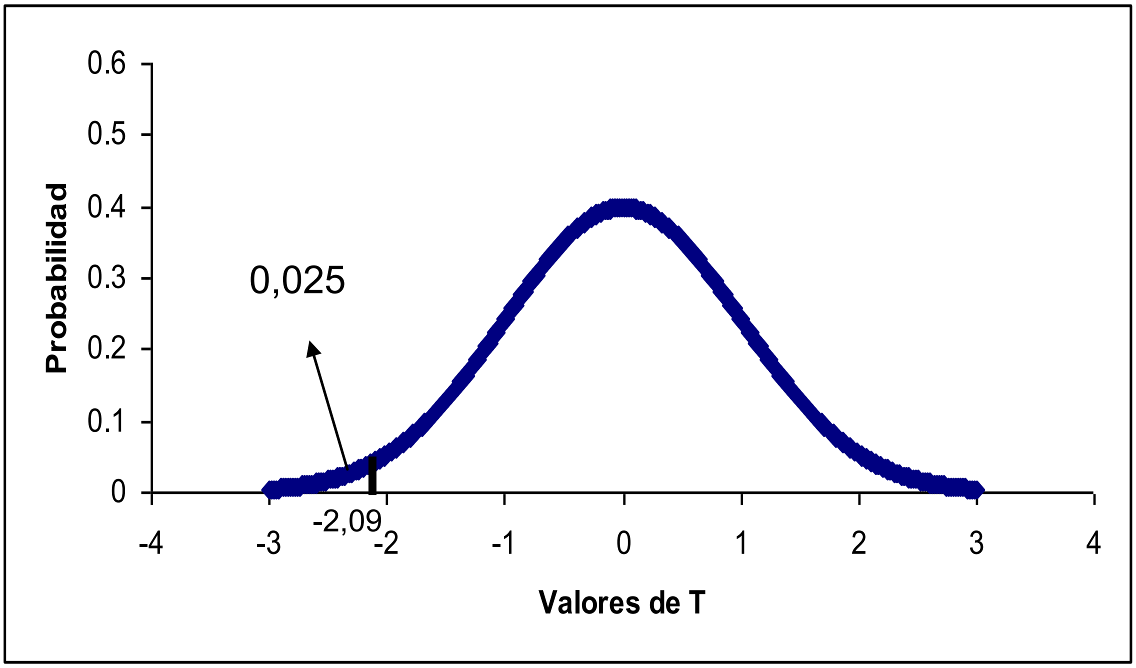
Únicamente cuando el tamaño de la muestra tiende a infinito las dos distribuciones serán las mismas.



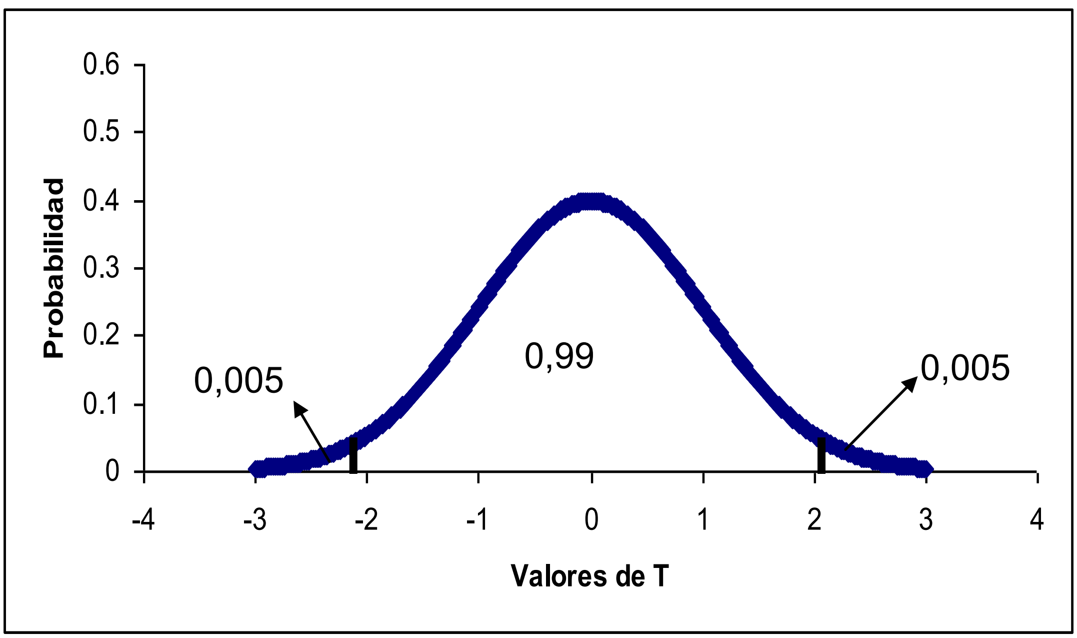
P (t > 1.72) = 0,05 con 20 grados de libertad



P (t > 2,09) = 0,025 con 20 grados de libertad



P (t < - 2,09) = 0,025 con 20 grados de libertad



P (-2,85 < t < 2,85) = 0,99 con 20 grados de libertad

PRUEBA DE HIPOTESIS SOBRE LA MEDIA DE UNA DISTRIBUCION NORMAL, VARIANZA DESCONOCIDA

Ciertamente sospechamos que las pruebas sobre una media poblacional μ con σ^2 desconocida, debe incluir el uso de la distribución t de Student.

La estructura de la prueba es idéntica a la del caso de σ^2 conocida, con la excepción de que el valor σ en la estadística de prueba se reemplaza por la estimación de S calculada y la distribución normal estándar se reemplaza con una distribución t.

- Muchos problemas del área de la Geografía y Ambiente, requieren que se tome una decisión entre **aceptar o rechazar** una **proposición** sobre algún **parámetro**.
- Esta proposición recibe el nombre de **hipótesis.**
- Este es uno de los aspectos más útiles de la inferencia estadística, puesto que muchos tipos de problemas de toma de decisiones, pruebas o experimentos pueden formularse como problemas de prueba de hipótesis.
- Una **hipótesis estadística** es una proposición o supuesto sobre los parámetros de una o más poblaciones.
- Un procedimiento que conduce a una decisión sobre una hipótesis en particular recibe el nombre de **prueba de hipótesis**.
- Los procedimientos de prueba de hipótesis dependen del empleo de la información contenida en la muestra aleatoria de la población de interés.
- Es necesario desarrollar un procedimiento de prueba de hipótesis teniendo en cuenta la probabilidad de llegar a una conclusión equivocada.
- La hipótesis nula, representada por Ho, es la afirmación sobre una o más características de poblaciones que al inicio se supone cierta (es decir, la "creencia a priori").
- La hipótesis alternativa, representada por H1, es la afirmación contradictoria a Ho, y ésta es la hipótesis del investigador.

La hipótesis nula se rechaza en favor de la hipótesis alternativa, sólo si la evidencia muestral sugiere que Ho es falsa. Si la muestra no contradice decididamente a Ho, se continúa creyendo en la validez de la hipótesis nula. Entonces, las dos conclusiones posibles de un análisis por prueba de hipótesis son *rechazar Ho o no rechazar Ho.*

Los "parámetros" mas utilizados en la investigación son:

a) Ho: μ = K, cuando se desea evaluar la media o promedio de una población

b) Ho: $\mu_1 = \mu_2$, se quiere estudiar diferencias entre dos medias provenientes de dos grupos o poblaciones

c) Ho: σ^2 = K, La varianza de la población

d) Ho: $\sigma_1^2 = \sigma_2^2$, diferencias entre varianzas de dos grupos o poblaciones

e) Ho: p = k, cuando se quiere estudiar la proporción de una población

d) Ho: $p_1 = p_2$, se quiere estudiar diferencias entre dos proporciones provenientes de dos grupos o poblaciones

PRUEBA DE HIPOTESIS SOBRE LA MEDIA de LA POBLACION

a) Ho : μ = K vs H_1: $\mu \neq K$ (Prueba Bilateral o dos colas)
b) Ho : μ = K vs H_1: $\mu > K$ (Prueba Unilateral o una cola derecha)
c) Ho : μ = K vs H_1: $\mu < K$ (Prueba Unilateral o una cola izquierda)

PRUEBA DE HIPOTESIS SOBRE LA MEDIA de LA POBLACION

Establecer el nivel de significancia de la prueba: se define como la máxima probabilidad de rechazar Ho cuando ésta es verdadera. Será denotado por la letra griega α

El nivel de significancia representa la máxima probabilidad de equivocarse en el sentido de concluir que Ho es falsa cuando en realidad no lo es. Este error se denomina Error Tipo I

Valores de α de mayor uso: 0,10 ; 0,05; y 0,01….pero algunas veces de acuerdo a la naturaleza de la investigación este valor puede ser aun mas alto a 0,10 (Exactamente, ¿De Que depende?)

LA PRUEBA ESTADISTICA PARA UNA MEDIA (Varianza desconocida y muestras pequeñas (n ≤ 30)

$$t_c = \frac{X - k}{S / \sqrt{n}}$$

Una vez fijados el estadístico de la prueba, su distribución y el nivel de significación, el próximo paso consiste en establecer las *regiones de aceptación y de rechazo de* Ho.

Región o zona de rechazo de Ho

La región de rechazo de Ho es uno o más intervalos de la recta real que describen al evento que conduce al rechazo de Ho y cuya probabilidad, cuando Ho es verdadera, es α

Región o zona de aceptación de Ho

Es un intervalo de la recta real que describe al evento que conduce al no rechazo de Ho con probabilidad 1-α ,cuando Ho es cierta.

a) $H_o : \mu = K$ vs $H_1 : \mu \neq K$ **Prueba Bilateral**

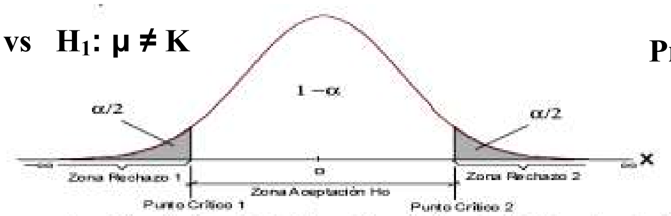


Figura 6.1: Representación de la distribución del estadístico bajo H_0 en una prueba bilateral

b) $H_o : \mu = K$ vs $H_1 : \mu > K$ **Prueba Unilateral**

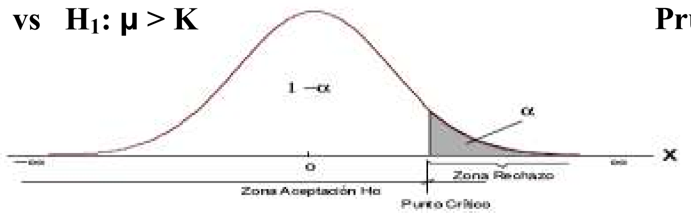


Figura 6.2: Representación de la distribución del estadístico bajo H_0 en una prueba unilateral derecha

c) $H_o : \mu = K$ vs $H_1 : \mu < K$ **Prueba Unilateral**

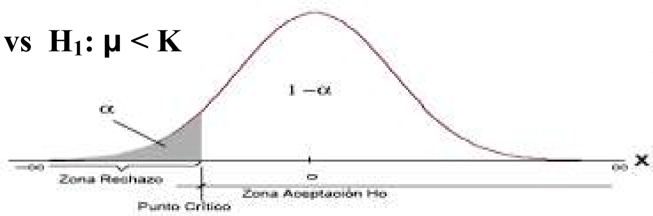


Figura 6.3: Representación de la distribución del estadístico bajo H_0 en una prueba unilateral izquierda

Ejemplo 1

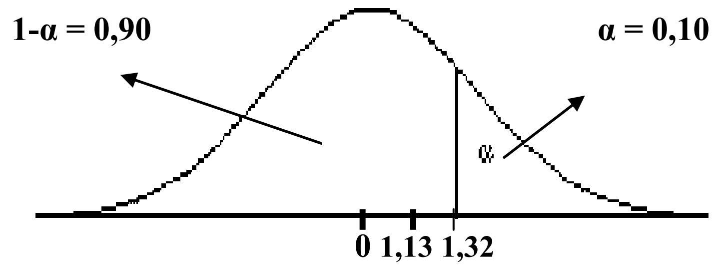
a) Hace algunos años, la media de estatura de los españoles adultos varones era de 170 cm. Pasado el tiempo, un muestreo realizado a 26 adultos da una medida de 172 cm y su desviación estándar 9 cm. ¿Puede afirmarse que esa diferencia de 2 cm es debida al azar o realmente la estatura media ha aumentado?.

Ho : μ = 170cm vs H1: μ > 170cm (Prueba Unilateral o una cola)

α = 0,10

$$t_c = \frac{X - k}{S / \sqrt{n}} = \frac{172 - 170}{9 / \sqrt{26}} = \frac{2}{1,765} = 1,13$$

Región de Aceptación de H_o **Región de Rechazo de la H_o**



Regla de decisión:
Si $t_c \geq t_\alpha$ se rechaza la H_o y se concluye a favor de H_1
Si $t_c < t_{(\alpha;n-1)}$ se Acepta la H_o

En este ejemplo, como $t_c = 1{,}13 < t_{(0{,}10;25)} = 1{,}32$
Se acepta que la estatura promedio de los españoles adultos sigue siendo de 170 cm

Ejemplo 2

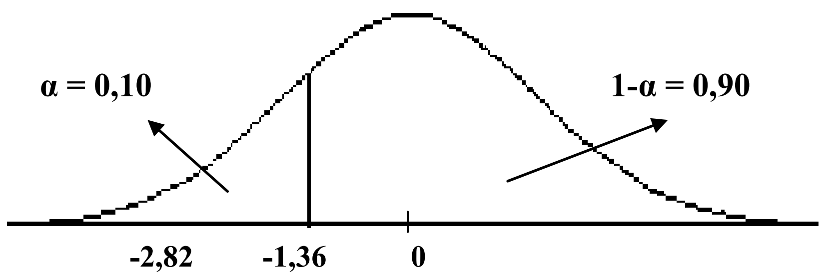
b) Sobre una cuenca se determino el calculo de un índice llamado de calidad de aguas (la combinación de elementos tóxicos químicos y sólidos) . Este índice toma valores entre 0 y 1; donde 0 implica aguas de mala calidad ambiental y 1 de excelente calidad ambiental. Se tomaron 13 muestras de 1 litro c/u de aguas de un río que circula por dicha cuenca cubriendo aguas arribas como aguas abajo. Los datos obtenidos después de los análisis de laboratorio fueron los siguientes:

0,6 0,5 0,3 0,4 0,4 0,4 0,4 0,4 0,2 0,5 0,3 0,5 0,5

Se quiere verificar si la calidad del agua es inferior a 0,5

$H_o : \mu = 0,5$ vs $H_1: \mu < 0,5$ (Prueba Unilateral o una cola)
$\alpha = 0,10$

Región de Rechazo de H_o Región de Aceptación de la H_o



$$t_c = \frac{X-k}{S/\sqrt{n}} = \frac{0,42-0,50}{0,11/\sqrt{13}} = \frac{-0,08}{0,03} = -2,82$$

El valor t de la tabla se ubica con $\nu = 13 - 1 = 12$ gl
y $\alpha = 0,10$ de una cola, …… $t_{(0,10;12)} = -1,36$

Regla de decisión:

Si $tc \le t_{\alpha}$ se rechaza la Ho y se concluye a favor de H1

Si $tc > t_{(\alpha;n-1)}$ se Acepta la Ho

En este ejemplo, como tc = -2,82 < t(0,10;12) = -1,36

Se rechaza Ho, por tanto se concluye favor de la H1 donde el índice de calidad de aguas promedio sobre el río de la cuenca es inferior a 0,5 por tanto de baja calidad ambiental

Ejemplo 3

c) Un supervisor de planta quiere conocer si el valor promedio del llenado de envases de larga duración contiene exactamente 0,75 L de una bebida. Para ello, se toman aleatoriamente 25 envases y se les mide el contenido. El promedio de esta muestra fue 0,88 L con una varianza de la muestra de 0,01 L2.

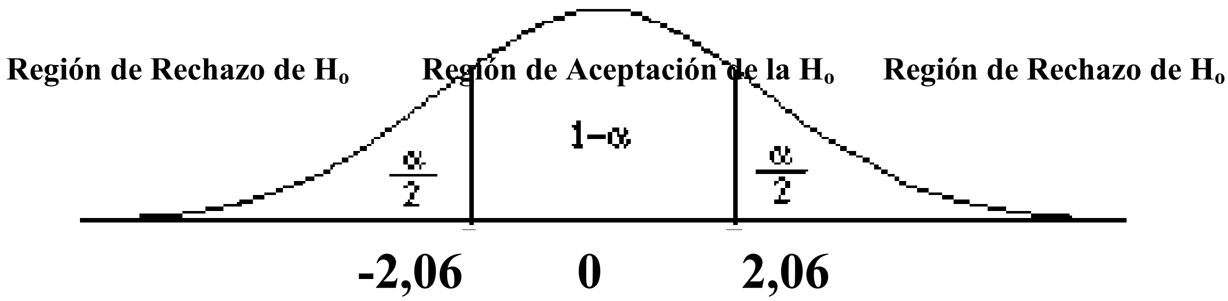
Ho : $\mu = 0,75$ L vs H1: $\mu \neq 0,75$ L (Prueba Bilateral o dos colas)

$\alpha = 0,05$

$$t_c = \frac{X-k}{S/\sqrt{n}} = \frac{0,88-0,75}{0,08/\sqrt{25}} = \frac{0,13}{0,02} = 6,50$$

El valor t tabulado se ubica con la probabilidad 0,05 de dos colas

Y con $\nu = 25 - 1 = 24$ g.l $t_{(0,05/2;24)} = 2,06$ a la derecha y -2,06 a la izquierda



Regla de decisión:

Si $|tc| \geq |t_{(\alpha;n-1)}|$ se rechaza la Ho y se concluye a favor de H1

Si $|tc| < |t_{(\alpha;n-1)}|$ se Acepta la Ho

En este ejemplo, como $|tc| = 6,50 > |t_{(0,05;24)}| = 2,06$

Se rechaza Ho, por tanto se concluye favor de la H1 donde el el proceso del llenado de los envases no esta cumpliendo con la cuota de 0,75 L por envase

•Inferencia sobre la diferencia de dos medias poblacionales ($\mu_1 - \mu_2$)

Caso A: Las varianzas son desconocidas e iguales

Al plantear pruebas de hipótesis donde se quiera comparar dos medias y solo se cuenta con muestras menores a 30 y las varianzas de ambos grupos son desconocidas y estadísticamente iguales entre ellos, el estadístico apropiado para la prueba de hipótesis de igualdad de medias es el siguiente:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - k}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

que se distribuye según una *T* de Student con ($n_1 + n_2 - 2$) grados de libertad.

Prueba t para diferencia de medias bajo el supuesto de Varianzas Iguales (Regla de Decisión)

Hipótesis nula	Valor de la prueba estadística
H_o: $\mu_1 = \mu_2$ equivalente H_o: $\mu_1 - \mu_2 = \boldsymbol{k}$	$$t_c = \frac{(x_1 - x_2) - k}{\sqrt{\frac{S_1^2 * (n_1 - 1) + S_2^2 * (n_2 - 1)}{(n_1 + n_2 - 2)} * (\frac{1}{n_1} + \frac{1}{n_2})}}$$
Hipótesis alternativa	Criterio de rechazo de H_o
H_1: $\mu_1 \neq \mu_2$ H_1: $\mu_1 > \mu_2$ H_1: $\mu_1 < \mu_2$	Rechazar H_o cuando $\lvert t_c \rvert \geq \lvert t_{(\alpha/2;(n-2))} \rvert$ Rechazar H_o cuando $t_c \geq \mathbf{t}_{(\alpha,(n-2))}$ Rechazar H_o cuando $t_c \leq \mathbf{t}_{(\alpha,(n-2))}$

Ejemplo I- **Prueba t para diferencia entre dos medias**

Para probar la eficacia de un tratamiento de poda en un bosque de ***Inga edulis***, un investigador decide comparar el **incremento del diámetro** de los fustes de los **árboles podados**, con el incremento en **árboles sin poda**. Para ello se localizan 20 lotes de los cuales a 10 se los poda y al resto no. Al cabo de 3 años se obtienen **los incrementos promedio para cada lote** siendo los resultados los siguientes (en cm):

Stand con poda:	0.29	0.305	0.28	0.32	0.35	0.297	0.30	0.298	0.315	0.324
Stand sin poda:	0.30	0.303	0.27	0.30	0.32	0.31	0.28	0.302	0.298	0.301

¿Existe efecto de la poda sobre el incremento del diámetro de los fustes?
Trabaje con un nivel de significación del 5%.

Planteamiento de hipótesis

$H_0: \mu_1 = \mu_2$
$H_1: \mu_1 > \mu_2$

Prueba Estadística

$\overline{X_1} = 0,308$ cm
$\overline{X_2} = 0,298$ cm
$S_1^2 = 0,0004$
$S_2^2 = 0,0002$
$n_1 = n_2$

$$T = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

$$T = \frac{0,0095}{\sqrt{\frac{0,0036+0,0018}{18}(0,2)}} = 1,23$$

El **nivel de significancia** seleccionado es de **α = 0,05** también expresado en porcentaje **5 %** .....refiere al chance de equivocarnos al rechazar la hipótesis nula cuando en realidad esta es cierta

Y como la hipótesis alternativa fue planteada de forma **unilateral** o de **una cola**, la **región de rechazo** de la hipótesis nula se refleja en la curva de la distribución normal o la distribución t de Student (cola de la derecha)

$$H_1: \mu_1 > \mu_2$$

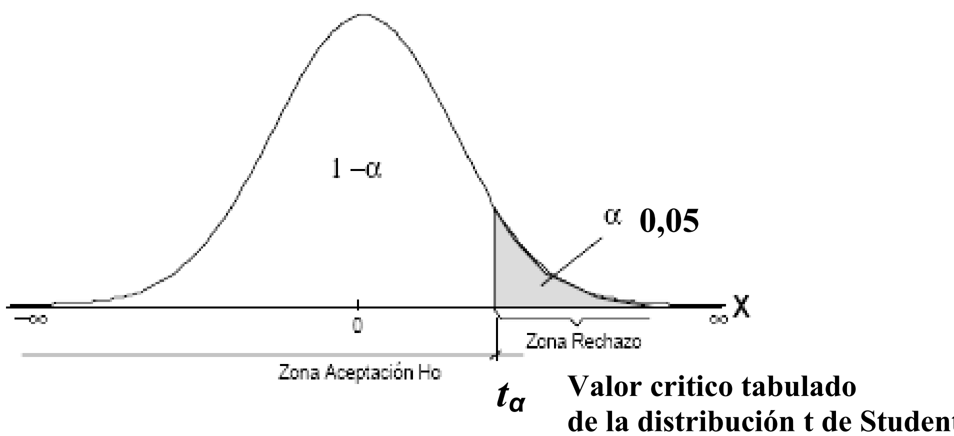
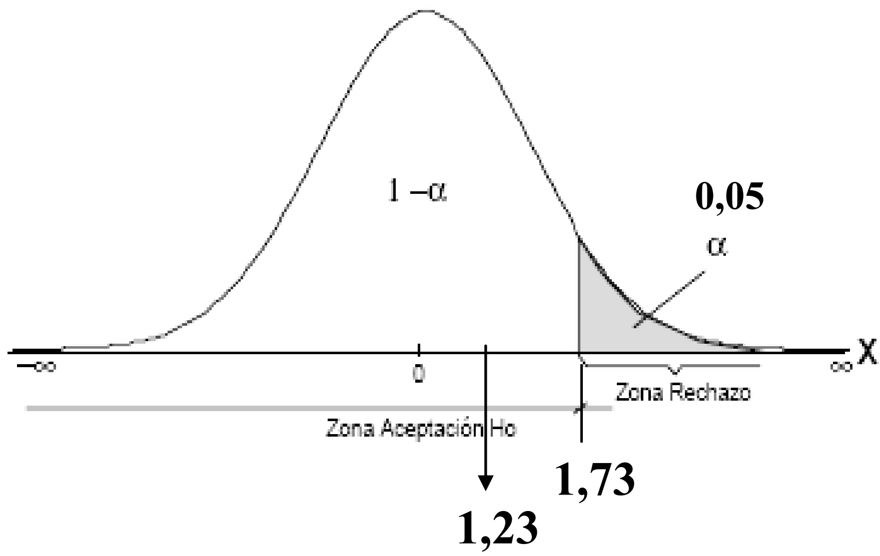


Tabla de valores criticos de la distribucion t de Student

Grados de libertad	Niveles de Significancia DOS COLA							
	0.500	0.250	0.200	0.100	0.050	0.025	0.010	0.005
1	1.00	2.41	3.08	6.31	12.71	25.45	63.66	127.32
2	0.82	1.60	1.89	2.92	4.30	6.21	9.92	14.09
3	0.76	1.42	1.64	2.35	3.18	4.18	5.84	7.45
4	0.74	1.34	1.53	2.13	2.78	3.50	4.60	5.60
5	0.73	1.30	1.48	2.02	2.57	3.16	4.03	4.77
6	0.72	1.27	1.44	1.94	2.45	2.97	3.71	4.32
7	0.71	1.25	1.41	1.89	2.36	2.84	3.50	4.03
8	0.71	1.24	1.40	1.86	2.31	2.75	3.36	3.83
9	0.70	1.23	1.38	1.83	2.26	2.69	3.25	3.69
10	0.70	1.22	1.37	1.81	2.23	2.63	3.17	3.58
11	0.70	1.21	1.36	1.80	2.20	2.59	3.11	3.50
12	0.70	1.21	1.36	1.78	2.18	2.56	3.05	3.43
13	0.69	1.20	1.35	1.77	2.16	2.53	3.01	3.37
14	0.69	1.20	1.35	1.76	2.14	2.51	2.98	3.33
15	0.69	1.20	1.34	1.75	2.13	2.49	2.95	3.29
16	0.69	1.19	1.34	1.75	2.12	2.47	2.92	3.25
17	0.69	1.19	1.33	1.74	2.11	2.46	2.90	3.22
18	0.69	1.19	1.33	1.73	2.10	2.45	2.88	3.20
19	0.69	1.19	1.33	1.73	2.09	2.43	2.86	3.17
20	0.69	1.18	1.33	1.72	2.09	2.42	2.85	3.15
21	0.69	1.18	1.32	1.72	2.08	2.41	2.83	3.14
22	0.69	1.18	1.32	1.72	2.07	2.41	2.82	3.12
23	0.69	1.18	1.32	1.71	2.07	2.40	2.81	3.10
24	0.68	1.18	1.32	1.71	2.06	2.39	2.80	3.09
25	0.68	1.18	1.32	1.71	2.06	2.38	2.79	3.08
26	0.68	1.18	1.31	1.71	2.06	2.38	2.78	3.07
27	0.68	1.18	1.31	1.70	2.05	2.37	2.77	3.06
28	0.68	1.17	1.31	1.70	2.05	2.37	2.76	3.05
29	0.68	1.17	1.31	1.70	2.05	2.36	2.76	3.04
30	0.68	1.17	1.31	1.70	2.04	2.36	2.75	3.03
31	0.68	1.17	1.31	1.70	2.04	2.36	2.74	3.02
32	0.68	1.17	1.31	1.69	2.04	2.35	2.74	3.01
33	0.68	1.17	1.31	1.69	2.03	2.35	2.73	3.01
34	0.68	1.17	1.31	1.69	2.03	2.35	2.73	3.00
35	0.68	1.17	1.31	1.69	2.03	2.34	2.72	3.00
36	0.68	1.17	1.31	1.69	2.03	2.34	2.72	2.99
37	0.68	1.17	1.30	1.69	2.03	2.34	2.72	2.99
38	0.68	1.17	1.30	1.69	2.02	2.33	2.71	2.98
39	0.68	1.17	1.30	1.68	2.02	2.33	2.71	2.98
40	0.68	1.17	1.30	1.68	2.02	2.33	2.70	2.97
	0.250	0.125	0.100	0.050	0.025	0.013	0.005	0.003
	Niveles de Significancia UNA COLA							

Como este valor está dentro de la región de aceptación, se concluye que no hay evidencia para rechazar H_0.



La conclusión para este caso en particular es que la poda aplicada a los árboles de *I. edulis* no afecta significativamente el incremento del diámetro después de los 3 años de crecimiento

Caso B: Las varianzas son desconocidas y diferentes
Recuérdese que según lo estudiado para el caso de la distribución normal de la diferencia de dos medias, la desviación estándar de la diferencia de medias muestrales, se calcula como:

$$S_{\overline{x}1-\overline{x}2} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

La “estandarización” que se obtiene utilizando las estimaciones de las varianzas muestrales es la siguiente:

$$T' = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_{\overline{x}1 - \overline{x}2}}$$

que tiene distribución T de Student con los grados de libertad que se especifican a continuación:

$$\nu = \frac{\left({S_1^2}/{n_1} + {S_2^2}/{n_2} \right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1+1} + \frac{\left(S_2^2/n_2\right)^2}{n_2+1}} - 2$$

Esto significa que si se desea realizar una prueba de diferencia de medias, pero se evidencia que las varianzas de estas dos muestras son estadísticamente diferentes la prueba T de Student a utilizar difiere un poco de la prueba T para cuando se tiene muestras con varianzas iguales, y esta diferencia básica recae en la manera en que se obtienen los grados de libertad (ajustados) para encontrar el valor T de student tabulado.

Prueba t para diferencia de medias bajo el supuesto de Varianzas desiguales (Regla de Decisión)

Hipótesis nula	Valor de la prueba estadística
H_o: $\mu_1 = \mu_2$ equivalente H_o: $\mu_1 - \mu_2 = \boldsymbol{k}$	$t_c = \dfrac{(x_1 - x_2) - k}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ $\qquad \nu = \dfrac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\dfrac{\left(S_1^2/n_1\right)^2}{n_1+1} + \dfrac{\left(S_2^2/n_2\right)^2}{n_2+1}} - 2$
Hipótesis alternativa	Criterio de rechazo de H_o
H_1: $\mu_1 \neq \mu_2$ H_1: $\mu_1 > \mu_2$ H_1: $\mu_1 < \mu_2$	Rechazar H_o cuando $\lvert t_c \rvert \geq \lvert t_{(\alpha/2;\nu)} \rvert$ Rechazar H_o cuando $t_c \geq \mathbf{t}_{(\alpha,\nu)}$ Rechazar H_o cuando $t_c \leq \mathbf{t}_{(\alpha,\nu)}$

Planteamiento de hipótesis

H_0: $\mu_1 = \mu_2$
H_1: $\mu_1 > \mu_2$

X_1= 0,308 cm; S_1^2 = 0,0004
X_2= 0,298 cm; S_2^2 = 0,0002
$n_1 = n_2$

Prueba Estadística

$$t_c = \frac{(x_1 - x_2) - k}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$v = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1+1} + \frac{\left(S_2^2/n_2\right)^2}{n_2+1}} - 2$$

$$t_c = \frac{0,0095}{\sqrt{\frac{0,0004}{10} + \frac{0,0002}{10}}} = 1,23$$

$$\upsilon = \frac{\left(\frac{0,0004}{10} + \frac{0,0002}{10}\right)^2}{\frac{\left(\frac{0,0004}{10}\right)^2}{11} + \frac{\left(\frac{0,0002}{10}\right)^2}{11}} - 2 = 17,8$$

El resultado en este caso coincide no solo en el valor del tc = 1,23; sino también en los grados de libertad (17,8 ~ 18) para hallar el valor t tabulado con α = 0,05 de una cola, y como consecuencia la conclusión es la misma también

Inferencia para muestras relacionadas o pareadas (*d*)

Se analizará ahora la diferencia entre las medias de dos grupos cuando los datos se obtienen de muestras que están **relacionadas**; es decir, los resultados del primer grupo no son independientes de los del segundo.

Como ejemplo se tiene la recolección de datos de suelos tomados sobre ***n*** parcelas, se quiere estimar el contenido de plomo existente en esas parcelas. A su vez, esta recolección involucra la toma de muestras en dos capas del suelo, es decir, de 0- 20 cm y 20- 40 cm.

De esta manera, se tiene que en cada ***n*** parcelas se obtendrán ***n*** pares de datos del contenido de plomo, donde ambos datos están relacionados espacialmente y se quiere conocer si los contenidos de este elemento difieren entre profundidades.

Otra situación similar, es cuando se tiene mediciones en el tiempo sobre ***n*** unidades de observación o de muestras, es decir, si se desea conocer el contenido de contaminantes aguas arriba de un río perteneciente a una cuenca, se seleccionan ***n*** puntos de muestreos pero se considera repetir la toma de muestras en dos épocas o periodos de tiempo, época lluviosa y época seca sobre las mismas unidades de observación. La estructura de datos seria la siguiente:

Unidad de muestra /observación	Característica medida Tiempo1/espacio1	Característica medida Tiempo2/espacio2	Diferencia entre valores medidos dentro de la unidad
1	X1	Y1	d1
2	X2	Y2	d2
3	X3	Y3	d3
.	.	.	.
.	.	.	.
n	Xn	Yn	dn

Los ***n*** pares de observaciones (antes y después) obtenidas en cada unidad de muestreo u observación **no son independientes** ya que el valor de las características medidas en cada par están correlacionadas por provenir de la misma unidad de observación. Con el término **observaciones apareadas** se hace referencia al diseño de experimentos que produce observaciones "de a pares" de las dos distribuciones que se comparan. En este tipo de diseño la variable de interés es la **diferencia** entre los valores de cada uno de los pares observados.

Sea ***Xi*** el primer miembro del par i-ésimo y ***Yi*** el segundo miembro, para ***n*** pares de observaciones se tendrá: (***X1,Y1***) , (***X2,Y2***) , (***X3,Y3***) , ... , (***Xn***,***Yn***). Si se toman las diferencias ***di*** = ***Xi- Yi***, se tendrá un conjunto de ***n*** observaciones, cada una de las cuales es una diferencia entre dos observaciones originales.

Prueba T de Student para observaciones apareadas

Esta prueba se basa en la distribución de la variable diferencia entre los pares de observaciones. Si Xi y Yi tienen distribución normal, entonces, las ***di*** = Xi-Yi tendrán distribución normal con media μd y varianza σ^2_d. El estimador de

μ_1-μ_2 es $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ y el estimador de σ_d es $S_d = \sqrt{\frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}{n-1}}$

Si la hipótesis nula que se quiere probar es $\mu d = 0$, luego para probar esta hipótesis el estadístico apropiado es:

$$T = \frac{\bar{d}}{S_d / \sqrt{n}} \sim T_{(n-1)}$$

Prueba t para observaciones relacionadas o pareadas (Regla de Decisión)

Hipótesis nula	Valor de la prueba estadística
Ho: µd = ***k***	$t_c = \dfrac{\bar{d} - k}{\sqrt{\dfrac{S_d^2}{n}}}$
Hipótesis alternativa	Criterio de rechazo de Ho
H1: µd ≠ k	Rechazar Ho cuando $\lvert tc \rvert \geq \lvert t\ (\alpha/2;\nu) \rvert$
H1: µd > k	Rechazar Ho cuando $tc \geq \mathbf{t}\ (\alpha,\nu)$
H1: µd < k	Rechazar Ho cuando $tc \leq \mathbf{t}\ (\alpha,\nu)$

Ejemplo. A un grupo de 12 estudiantes seleccionados aleatoriamente de un curso les fue aplicado dos pruebas de conocimiento sobre una asignatura. Las dos pruebas fueron aplicadas "antes" y "después" de participar en una practica de laboratorio correspondiente

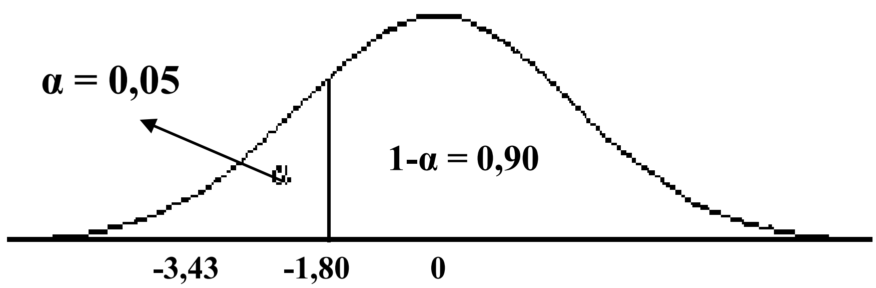
Estudiante	**Antes**	**Después**	Diferencia (**Después-** antes)
1	12,8	13,4	**-0,6**
2	17,6	17,4	**0,2**
3	11,0	11,8	**-0,8**
4	14,9	15,2	**-0,3**
5	18,3	18,7	**-0,4**
6	13,6	13,6	**0**
7	11,8	12,5	**-0,7**
8	15,8	16,8	**-1,0**
9	15,0	15,2	**-0,2**
10	13,0	12,8	**0,2**
11	12,6	13,0	**-0,4**
12	16,2	16,7	**-0,5**
		Media de las diferencias	**0,375**

$$\mathbf{S_d^2 = 0,1439}$$
$$\mathbf{S_d = 0,3793}$$

Se quiere verificar si la actividad practica tuvo un efecto satisfactorio reflejado en una mayor calificación del grupo de la asignatura después de aplicar la actividad

H_o: μ_d $= 0$
H_1: μ_d < 0

$$t_c = \frac{0{,}375}{\sqrt{\frac{0{,}1439}{12}}} = 3{,}43$$



El valor tabulado con (n -1) = 12 -1 = 11 grados de libertad y usando α = 0,05, se tiene t = -1,80