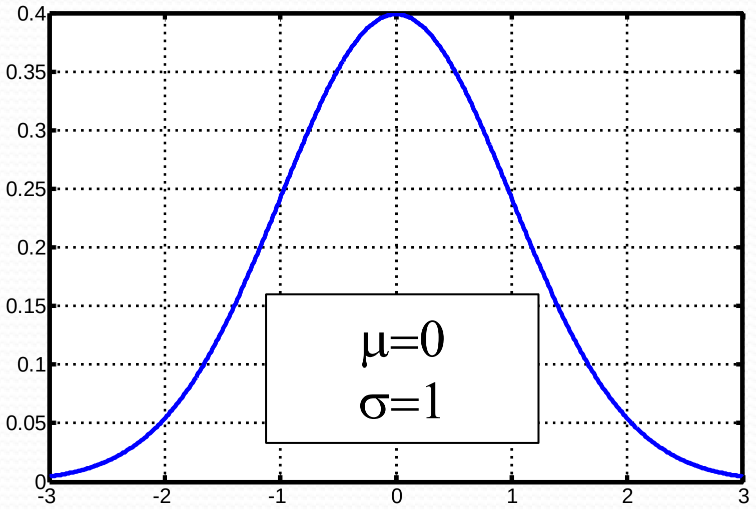
4.3. Comparación de prestaciones en presencia de aleatoriedad

Repaso de Estadística: Distribución Normal

- Independientemente de qué índice se escoja, un buen ingeniero debería, en primer lugar, determinar si las diferencias entre las medidas obtenidas por un test de rendimiento en presencia de aleatoriedad son **estadísticamente significativas** → Necesitaremos repasar algunos conceptos de estadística.
- **Distribución gaussiana o normal:** Es una distribución de probabilidad caracterizada por su media μ y su varianza σ^2 cuya función de probabilidad viene dada por:

$$Prob(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

La probabilidad de obtener un elemento en el rango $[\mu - 2\sigma, \mu + 2\sigma]$ es del 95%



- Teorema del límite central: la media de un conjunto grande de muestras aleatorias de cualquier distribución e independientes entre sí pertenece una distribución normal.

Repaso de Estadística: Distribución t de Student

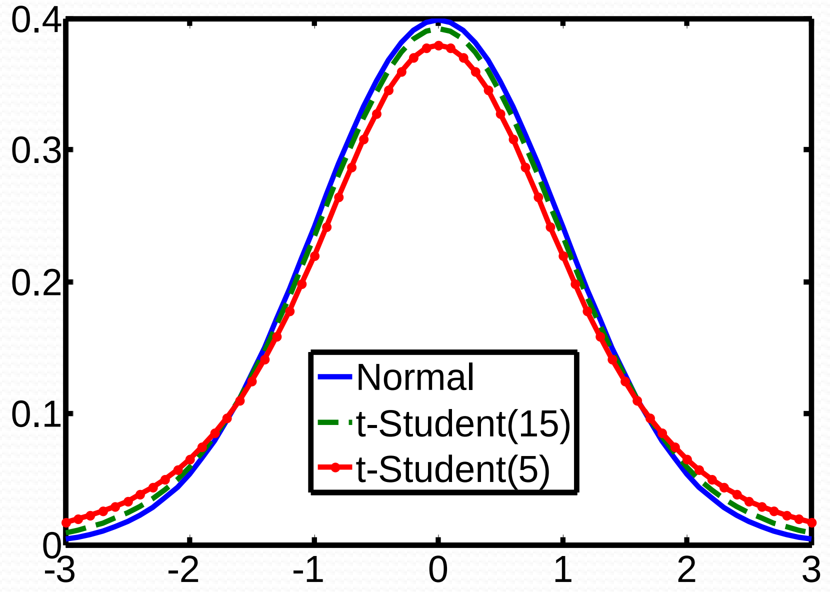
Si extraemos **n** muestras $\{d_1, d_2, \ldots, d_n\}$ pertenecientes a una distribución Normal de media $\mu = \bar{\boldsymbol{d}}_{\boldsymbol{real}}$, y calculo la siguiente medida (=estadístico):

$$t_{exp} = \frac{\bar{d} - \bar{d}_{real}}{s/\sqrt{n}}$$

siendo $\bar{\boldsymbol{d}}$ la media muestral y $\boldsymbol{s}$ la desviación típica muestral

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}{n-1}}$$

y repetimos el experimento muchas veces, veremos que el histograma de esos t_{exp} converge a una distribución t-Student con n-1 grados de libertad (*degrees of freedom*, df).



$$Prob(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

$s/\sqrt{n} \equiv$ Error estándar

¿Para qué me puede servir esto?

Ejemplo 1: Comparación de rendimiento entre A y B

- Tiempos de ejecución (en segundos) de 6 programas (P1…P6) en dos máquinas diferentes (A y B) en condiciones donde puede haber alta aleatoriedad.

Programa	t_A (s)	t_B (s)	d = t_A-t_B (s)
P1	142	100	42
P2	139	92	47
P3	152	128	24
P4	112	82	30
P5	156	148	8
P6	166	171	-5

$\overline{t_A} = 144{,}5s$
$\overline{t_B} = 120{,}2s$
¿Es significativa esta diferencia?

$\bar{d} = 24{,}3\ s \quad s = 19{,}9\ s \quad s/\sqrt{n} = 8{,}12\ s$

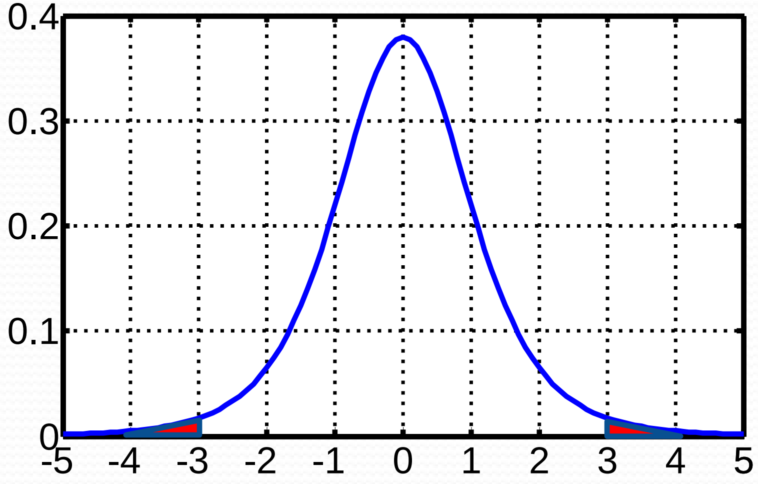
- **Si partimos de la hipótesis (hipótesis "nula", $\boldsymbol{H_0}$) de que las máquinas tienen rendimientos equivalentes**, entonces las diferencias se deben a una suma (=una media) de factores aleatorios independientes. En ese caso d_i serán muestras de una distribución normal de media cero ($\bar{d}_{real} = 0$). Por tanto:

$$t_{exp} = \frac{\bar{d}}{s/\sqrt{n}} = \frac{24{,}3s}{8{,}12s} = 2{,}99$$

pertenecerá a una distribución t de Student con 6-1=5 grados de libertad. **¿Qué probabilidad hay de que esto sea realmente así?**

Nivel o Grado de Significatividad (α)

- Distribución t de Student con 5 grados de libertad (T5).



$$\boldsymbol{P-value} = P(|t| \geq |t_{exp}|) \text{ en } T_{n\text{-}1}$$
$$= 2\times P(t \leq -|t_{exp}|) \text{ en } T_{n\text{-}1}$$

=DISTR.T.2C(2,99;5) = 0,03 *(Excel).*
=DISTR.T(2,99;5;2) = 0,03 *(Calc).*
=2·tcdf(-2.99,5) = 0,03 *(Matlab).*

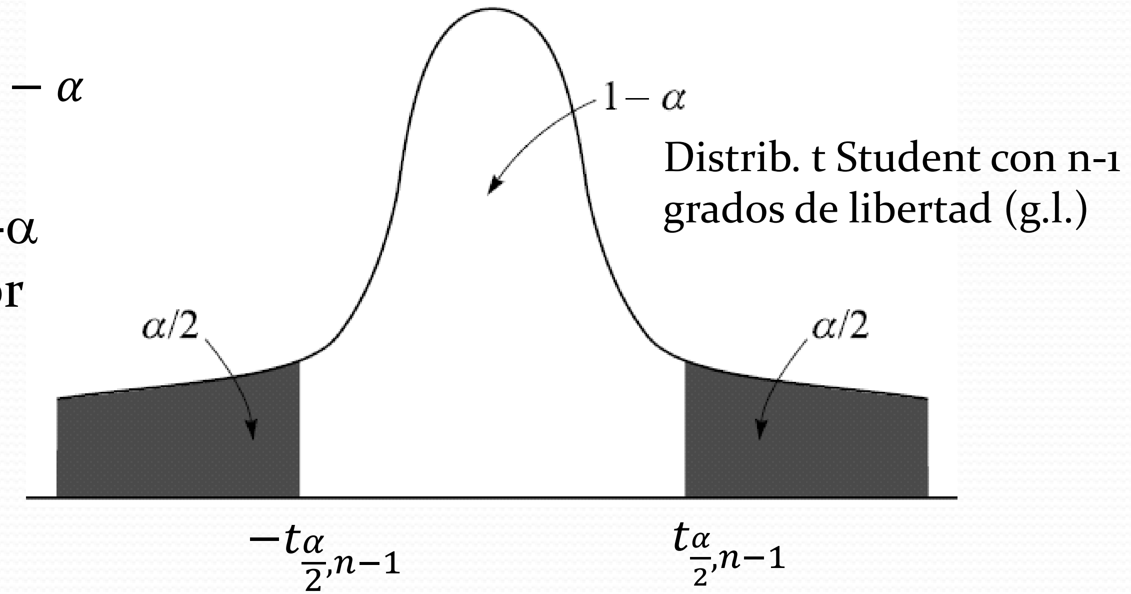
La probabilidad de obtener un valor de |t| igual o superior a 2,99 de una distribución t de Student con 5 grados de libertad es de 0,03 (**P-value** (Valor-P) = 0,03). ¿Es eso mucho o poco? Debemos definir un umbral: **nivel o grado de significatividad α**. Normalmente, α=0,05 (5%).

Conclusión para el ejemplo 1: Como P-value < α (0,03 < 0,05) diremos que: "para un grado de significatividad α=0,05 o para un **nivel de confianza** (1-α)*100 (95%), las máquinas A y B tienen rendimientos estadísticamente diferentes." En ese caso, B sería, de media, 1,2 veces más rápida que A en ejecutar cada programa (144,5s/120,2s = 1,2). En caso contrario, no habríamos podido descartar la hipótesis de que las máquinas tengan rendimientos equivalentes para α=0,05.

Intervalos de confianza para t_{exp}

- Para un nivel de significatividad α (típ. 0,05 = 5%), buscamos el valor $t_{\alpha/2,n-1}$ que cumpla $Prob\left(|t| > t_{\frac{\alpha}{2},n-1}\right) = \alpha$ o equivalentemente:

$$Prob\left(-t_{\frac{\alpha}{2},n-1} \leq t \leq t_{\frac{\alpha}{2},n-1}\right) = 1 - \alpha$$



- Diremos que para un nivel de confianza 1-α (típ. 0,95 = 95%), **para aceptar H_0** el valor de t_{exp} debería situarse en el intervalo:

$$\left[-t_{\frac{\alpha}{2},n-1}, t_{\frac{\alpha}{2},n-1}\right]$$

- A dicho intervalo se le denomina **intervalo de confianza** de la medida para un nivel de significatividad α. Teniendo en cuenta que:

$$Prob\left(-t_{\frac{\alpha}{2},n-1} \leq t \leq t_{\frac{\alpha}{2},n-1}\right) = 1 - 2 \times Prob\left(t \leq -t_{\frac{\alpha}{2},n-1}\right) = 1 - 2 \times Prob\left(t > t_{\frac{\alpha}{2},n-1}\right)$$

es fácil demostrar que $t_{\alpha/2,n-1}$ cumple que (ver figura):

$$Prob\left(t \leq -t_{\frac{\alpha}{2},n-1}\right) = Prob\left(t > t_{\frac{\alpha}{2},n-1}\right) = \alpha/2$$

Intervalos de confianza para t_{exp} (Ejemplo 1)

- En el caso del *Ejemplo 1*, para un nivel de significatividad de α=0,05, buscamos $t_{\alpha/2,n-1}$ tal que:

$$Prob\left(t \leq -t_{\frac{\alpha}{2},n-1}\right) = \alpha/2 = 0{,}025$$

para una distribución t de Student con 5 grados de libertad. Eso se puede obtener, por ejemplo:

  - Consultando tablas estadísticas. P.ej. en este enlace (2-*Tail Alpha* = 0,05, *df* = n-1 = 5).
  - En *Excel*, haciendo: ABS(INV.T(alfa/2;n-1)) = ABS(INV.T(0,025;5)) =2,57.
  - En *Calc*, DISTR.T.INV(alfa;n-1)=DISTR.T.INV(0,05;5)=2,57.
  - En *Matlab*, haciendo: abs(tinv(alfa/2,n-1))= abs(tinv(0,025,5)) =2,57.

$t_{\alpha/2,n-1}$

- Dicho de otra manera, si las diferencias entre los tiempos de ejecución de ambas máquinas se debieran a factores aleatorios, existiría un 95% de probabilidad de que

$$t_{exp} = \frac{\bar{d}}{s/\sqrt{n}}$$

se encuentre en el rango $\left[-t_{\frac{\alpha}{2},n-1}, t_{\frac{\alpha}{2},n-1}\right] = \left[-t_{0,025,5}, t_{0,025,5}\right]$= [-2,57 , 2,57].

Como t_{exp}=2,99 **no** está en ese rango, concluiremos nuevamente que **rechazamos la hipótesis de que ambas máquinas tienen rendimientos equivalentes con el 95% de confianza**.

Intervalos de confianza para $\bar{d}_{real}$

- Acabamos de ver que si las diferencias entre los tiempos de ejecución de ambas máquinas se debieran a factores aleatorios, existiría un 95% de probabilidad de que t_{exp} se encuentre en el rango $[-t_{\frac{\alpha}{2},n-1}, t_{\frac{\alpha}{2},n-1}]$ = [-2,57, 2,57].
- Como

$$t_{exp} = \frac{\bar{d} - \bar{d}_{real}}{s/\sqrt{n}} \in \left[-t_{\frac{\alpha}{2},n-1}, t_{\frac{\alpha}{2},n-1}\right] = [-2{,}57, 2{,}57]$$

sin más que identificar t_{exp} con los valores límite $\pm t_{\frac{\alpha}{2},n-1}$ sabemos que, de ser Ho cierta, habrá un 95% de probabilidad de que el valor medio real $\bar{d}_{real}$ de las diferencias entre los tiempos de ejecución se encuentre en el intervalo:

$$\bar{d}_{real} \in \left[\bar{d} - \frac{s}{\sqrt{n}} \times t_{\frac{\alpha}{2},n-1}, \bar{d} + \frac{s}{\sqrt{n}} \times t_{\frac{\alpha}{2},n-1}\right] = 24{,}3 \mp 20{,}9 = [3{,}4\ , 45{,}2]\ s$$

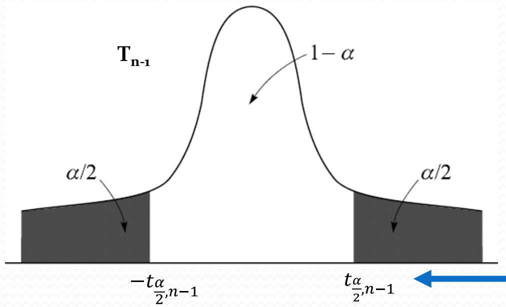
Y el problema se transforma simplemente en comprobar si ese valor medio real $\bar{d}_{real}$ puede o no ser **cero**.

En nuestro ejemplo, como el intervalo **no** incluye el cero, concluiremos una vez más que **la hipótesis de que ambas máquinas pueden tener rendimientos equivalentes no es cierta al 95% de confianza.**

Resumen: Test t para muestras pareadas

- Partimos de:

Exp.	tA	tB	$d_i = tA_i - tB_i$
P_1	tA_1	tB_1	d_1
P_2	tA_2	tB_2	d_2
…	…	…	…
P_n	tA_n	tB_n	d_n



α

df = n-1

df	0.20	0.10	0.05	0.02	0.01	0.001
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079
8	1.3968	1.8595	2.3060	2.8965	3.3554	5.0413
9	1.3830	1.8331	2.2622	2.8214	3.2498	4.7809
10	1.3722	1.8125	2.2281	2.7638	3.1693	4.5869
11	1.3634	1.7959	2.2010	2.7181	3.1058	4.4370
12	1.3562	1.7823	2.1788	2.6810	3.0545	4.3178
13	1.3502	1.7709	2.1604	2.6503	3.0123	4.2208
14	1.3450	1.7613	2.1448	2.6245	2.9768	4.1405
15	1.3406	1.7531	2.1314	2.6025	2.9467	4.0728
16	1.3368	1.7459	2.1199	2.5835	2.9208	4.0150
17	1.3334	1.7396	2.1098	2.5669	2.8982	3.9651
18	1.3304	1.7341	2.1009	2.5524	2.8784	3.9216
19	1.3277	1.7291	2.0930	2.5395	2.8609	3.8834
20	1.3253	1.7247	2.0860	2.5280	2.8453	3.8495
21	1.3232	1.7207	2.0796	2.5176	2.8314	3.8193
22	1.3212	1.7171	2.0739	2.5083	2.8188	3.7921
23	1.3195	1.7139	2.0687	2.4999	2.8073	3.7676
24	1.3178	1.7109	2.0639	2.4922	2.7969	3.7454
25	1.3163	1.7081	2.0595	2.4851	2.7874	3.7251
26	1.3150	1.7056	2.0555	2.4786	2.7787	3.7066
27	1.3137	1.7033	2.0518	2.4727	2.7707	3.6896
28	1.3125	1.7011	2.0484	2.4671	2.7633	3.6739
29	1.3114	1.6991	2.0452	2.4620	2.7564	3.6594
30	1.3104	1.6973	2.0423	2.4573	2.7500	3.6460

- Ho: Rendimiento A ≡ Rendimiento B, es decir, $d_i \sim \mathcal{N}\left(\bar{d}_{real}, \sigma^2\right)$ con $\bar{d}_{real} = 0$
- Calculo $t_{exp} = \frac{\bar{d} - \bar{d}_{real}}{s/\sqrt{n}} \sim T_{n-1}$ siendo $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$ $s = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$
- Definimos el nivel o grado de significatividad α.
- Rechazamos Ho para un nivel de confianza (1- α)*100(%) si:
  1. Método 1: p-value < α. Siendo p-value = $P(|t| \geq |t_{exp}|)$ en T_{n-1} ≈ Prob (Ho podría ser cierta).
  2. Método 2: $t_{exp} \notin \left[-t_{\frac{\alpha}{2},n-1}, t_{\frac{\alpha}{2},n-1}\right]$. Siendo $t_{\frac{\alpha}{2},n-1}$ el valor que hace que $Prob\left(|t| > t_{\frac{\alpha}{2},n-1}\right) = \alpha$ para una distribución t de Student con n-1 grados de libertad.
  3. Método 3: $0 \notin \left[\bar{d} - \frac{s}{\sqrt{n}} \times t_{\frac{\alpha}{2},n-1}, \bar{d} + \frac{s}{\sqrt{n}} \times t_{\frac{\alpha}{2},n-1}\right]$. Intervalo de confianza para $\bar{d}_{real}$.

Ejemplo 2: ¿Influye el parámetro *proxy_cache_min_uses* en este servidor?

- Productividades (en páginas web/s) obtenidas por el servidor en 5 experimentos diferentes para dos valores diferentes (A y B) del parámetro *proxy_cache_min_uses* de Nginx en condiciones donde puede haber alta aleatoriedad.

Experimento	X_A (pág/s)	X_B (pág/s)	d = X_A-X_B (pág/s)
Exp1	23	15	8
Exp2	28	22	6
Exp3	19	20	-1
Exp4	29	27	2
Exp5	36	39	-3

- Usando como criterio la media aritmética ($\overline{X_A}$=27 pág/s, $\overline{X_B}$=25 pág/s) parece que el parámetro A obtiene mejor productividad que el B pero, ¿son significativas las diferencias para un nivel de confianza del 95%? $(1-\alpha)\times 100 = 95$ → **grado de significatividad α = 0,05**.

Ejemplo 2: Realizo el test t para muestras pareadas

- Hago la siguiente hipótesis (Ho):

  Rendimiento A ≡ Rendimiento B, es decir, $d_i \sim \mathcal{N}\left(\bar{d}_{real}, \sigma^2\right)$ con $\bar{d}_{real} = 0$

- Calculo $t_{exp} = \frac{\bar{d}-\bar{d}_{real}}{s/\sqrt{n}} \sim T_{n-1}$

Experimento	X_A (pág/s)	X_B (pág/s)	d = X_A-X_B (pág/s)
Exp1	23	15	8
Exp2	28	22	6
Exp3	19	20	-1
Exp4	29	27	2
Exp5	36	39	-3

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{8+6-1+2-3}{5} = 2{,}4\ pág/s$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}{n-1}} = \sqrt{\frac{(8-2{,}4)^2 + \cdots + (-3-2{,}4)^2}{5-1}} = 4{,}6\ pág/s$$

$$\frac{s}{\sqrt{n}} = \frac{4{,}6}{\sqrt{5}} = 2{,}06\ pág/s \qquad \boldsymbol{t_{exp} = \frac{\bar{d}}{s/\sqrt{n}} = \frac{2,4}{2,06} = 1,16}$$

- Método 1:
  - p-value = $P(|t| \geq |t_{exp}|)$ en T_{n-1} = $P(|t| \geq |1{,}16|)$ en T_4 ≈ Prob (Ho podría ser cierta).
  - Uso *Calc* (por ejemplo): p-value = DISTR.T(1,16;4;2) = 0,31.
  - Como p-value > α (0,31>0,05) no podemos rechazar la hipótesis Ho al 95% de nivel de confianza (los parámetros A y B sí podrían tener rendimientos equivalentes).

Ejemplo 2: Otros métodos para hacer el test

- Método 2 (intervalos de confianza para t_{exp}):

Calculo $t_{\frac{\alpha}{2},n-1}$, que es el valor que hace que $Prob\left(|t| > t_{\frac{\alpha}{2},n-1}\right) = \alpha$ para una distribución t de Student con n-1 grados de libertad.

En nuestro caso: α = 0,05, *df* = n-1 = 4:

$$Prob\left(|t| > t_{0{,}025,\,4}\right) = 0{,}05$$

Mirando la tabla: $t_{0{,}025,\,4} = 2{,}78$

df = n-1; α

df	0.20	0.10	0.05	0.02	0.01	0.001
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688

➢ Como t_{exp} = 1,16 ∈ [-2,78, 2,78] no podemos rechazar la hipótesis Ho al 95% de nivel de confianza (los parámetros A y B sí podrían tener rendimientos equivalentes).

- Método 3 (intervalos de confianza para $\bar{d}_{real}$):

$$\left[\bar{d} - \frac{s}{\sqrt{n}} \times t_{\frac{\alpha}{2},n-1}, \bar{d} + \frac{s}{\sqrt{n}} \times t_{\frac{\alpha}{2},n-1}\right] = [2{,}4 - 2{,}06 \times 2{,}78 \;\;,\;\; 2{,}4 + 2{,}06 \times 2{,}78] = [-3{,}3,\ 8{,}1] \text{ pág/s.}$$

➢ Como 0 ∈ [-3,3, 8,1] no podemos rechazar la hipótesis Ho al 95% de nivel de confianza (los parámetros A y B sí podrían tener rendimientos equivalentes).

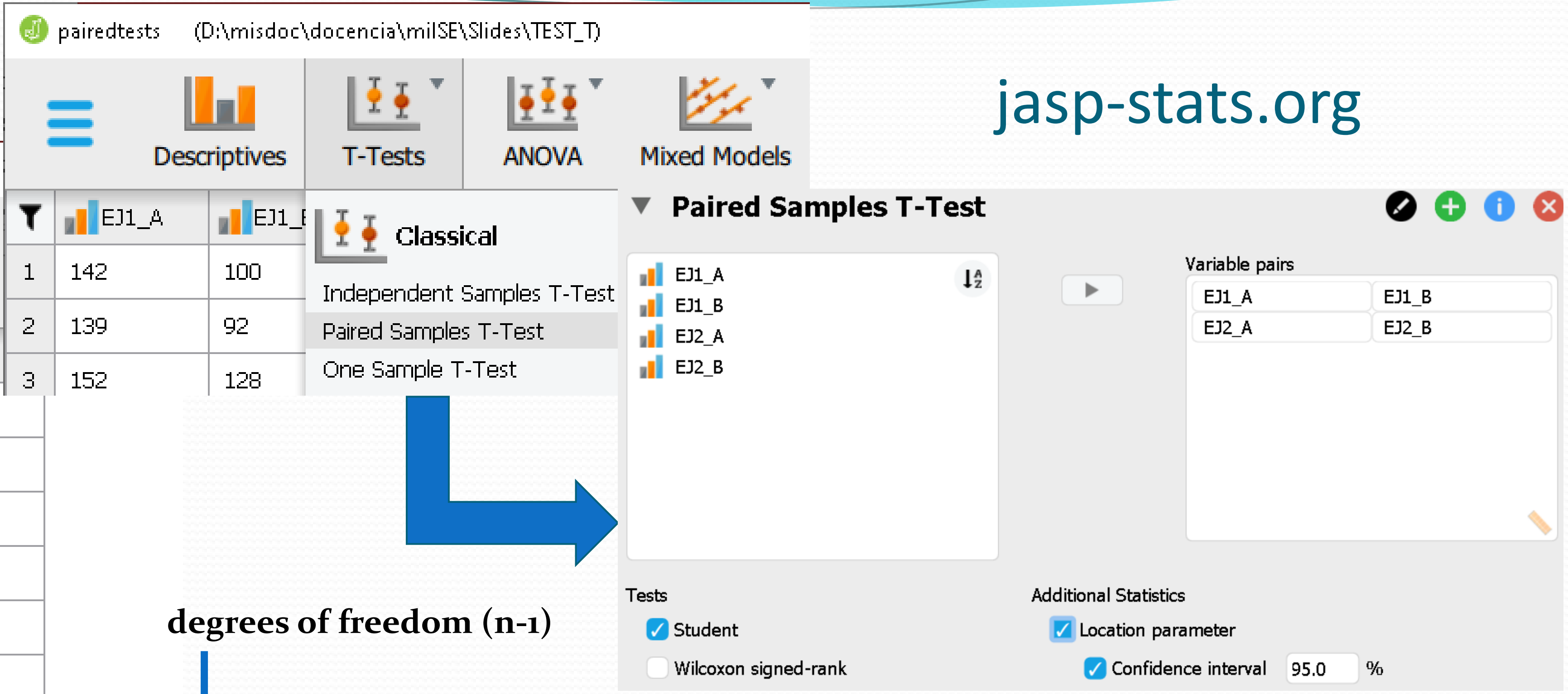
Test t con JASP

jasp-stats.org

pairedtests (D:\misdoc\docencia\milSE\Slides\TEST_T)

Descriptives · T-Tests · ANOVA

	EJ1_A	EJ1_B	EJ2_A	EJ2_B
1	142	100	23	15
2	139	92	28	22
3	152	128	19	20
4	112	82	29	27
5	156	148	36	39
6	166	171		



degrees of freedom (n-1)

t_{exp} $\quad$ $p - value$ $\quad$ $\bar{d}$ $\quad$ Standard Error $\frac{s}{\sqrt{n}}$

Paired Samples T-Test

Measure 1		Measure 2	t	df	p	Mean Difference	SE Difference	95% CI for Mean Difference Lower	95% CI for Mean Difference Upper
EJ1_A	-	EJ1_B	2.991	5	0.030	24.333	8.135	3.422	45.245
EJ2_A	-	EJ2_B	1.163	4	0.310	2.400	2.064	-3.331	8.131

Intervalo de confianza (95%) para $\overline{d_{real}}$: $\bar{d} \pm \frac{s}{\sqrt{n}} \times t_{\frac{\alpha}{2},n-1}$

Otra utilidad del test t: Estimación de intervalos de confianza de medias de medidas experimentales

Hipótesis: Realizamos n medidas {d_1, d_2, …, d_n} de un mismo fenómeno (p.ej. tiempos de ejecución de un programa, tiempos acceso de un disco duro, productividades de red,…). Si éstas pueden diferir debido a una suma de efectos aleatorios, podemos suponer que se distribuyen según una normal de media $\bar{d}_{real}$, que es el valor que buscamos. En ese caso, sabemos que

$$t_{exp} = \frac{\bar{d} - \bar{d}_{real}}{s/\sqrt{n}}$$

pertenece a la distribución t-Student con n-1 grados de libertad, siendo $\bar{d}$ y s la media y la desviación típica muestrales, respectivamente.

Por tanto, hay un (1-α)*100% de probabilidad de que el valor medio real $\bar{d}_{real}$ se encuentre en el intervalo:

$$\bar{d} \pm \frac{s}{\sqrt{n}} t_{\alpha/2,n-1}$$

Utilidad: Podemos usar esta información para determinar un intervalo de confianza para $\bar{d}_{real}$, y no quedarnos simplemente con el valor medio muestral.

Ejemplo

Queremos determinar un intervalo de confianza del 95% para el tiempo medio de escritura de un determinado fichero en un disco duro. Para ello, se han realizado n=8 medidas experimentales:

#exp	d (ms)
1	835
2	798
3	823
4	803
5	834
6	825
7	813
8	829

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = 820ms$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}{n-1}} = 14ms$$

$$t_{\frac{\alpha}{2}, n-1} = t_{\frac{0,05}{2}, 8-1} = 2,36$$

df = 8-1

α = 0,05

df	0.20	0.10	0.05	0.02	0.01	0.001
1	3.0777	6.3138	12.7062	31.8205	63.6567	636.6192
2	1.8856	2.9200	4.3027	6.9646	9.9248	31.5991
3	1.6377	2.3534	3.1824	4.5407	5.8409	12.9240
4	1.5332	2.1318	2.7764	3.7469	4.6041	8.6103
5	1.4759	2.0150	2.5706	3.3649	4.0321	6.8688
6	1.4398	1.9432	2.4469	3.1427	3.7074	5.9588
7	1.4149	1.8946	2.3646	2.9980	3.4995	5.4079

O bien:

- En *Excel*, haciendo: ABS(INV.T(alfa/2;n-1)).
- En *Calc*, DISTR.T.INV(alfa;n-1).

Por tanto, hay un 95% (α=0,05) de probabilidad de que el tiempo medio de escritura **real** de ese fichero se encuentre en el intervalo:

$$\bar{d} \pm \frac{s}{\sqrt{n}} t_{\alpha/2,n-1} = 820 \pm \frac{14}{\sqrt{8}} \times 2,36 = [808, 832]ms$$