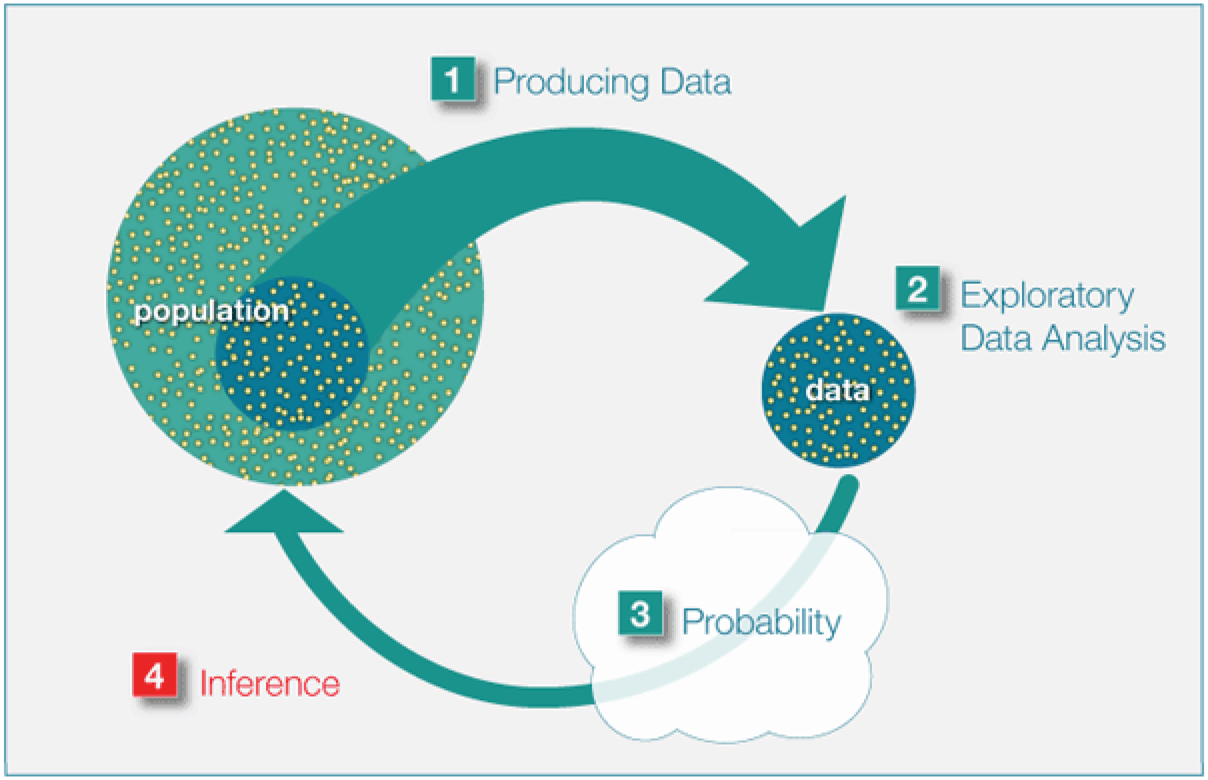
Inferencia Estadística



Profesor: Pedro Saa (pnsaa@uc.cl)
Año: 1-2025

OBJETIVOS DE APRENDIZAJE

- **O1**: Comprender y analizar el **poder de una prueba** y su relación con el el **tamaño del efecto** y nivel de significancia

- **O2**: Incorporar el **tamaño del efecto** como una métrica cuantitativa de impacto de investigación que su ejercicio mejora la planificación y análisis de los resultados.

- **O3**: Entender bajó qué condiciones es recomendable el uso de un test-Z o test de t mediante la **distribución t**

- **O4:** Formular hipótesis atingente a **comparación de promedios** e interpretar los resultados de las pruebas estadísticas o intervalo de confianza

- **O5:** Comprender la diferencia entre **test pareados y no pareados**

Distribución t

¿Por qué requerimos grandes muestras?

Si asumimos que las observaciones son independientes y la distribución de la población no es extremadamente asimétrica, tener grandes muestras nos permite:

- **Aproximar la distribución muestral sea normal**
- **Que la estimación del error muestral sea fidedigna**

$$SE = \frac{\sigma}{\sqrt{n}}$$

…Al no tener la desviación estándar de población, la desviación estándar de la muestra es nuestra mejor estimación si hay un número grande observaciones

(n > 30)

¿Qué hacemos cuando tenemos poca muestra n < 30?

Gosset fue pionero en el desarrollo de diseños y análisis experimentales con **muestras pequeñas**

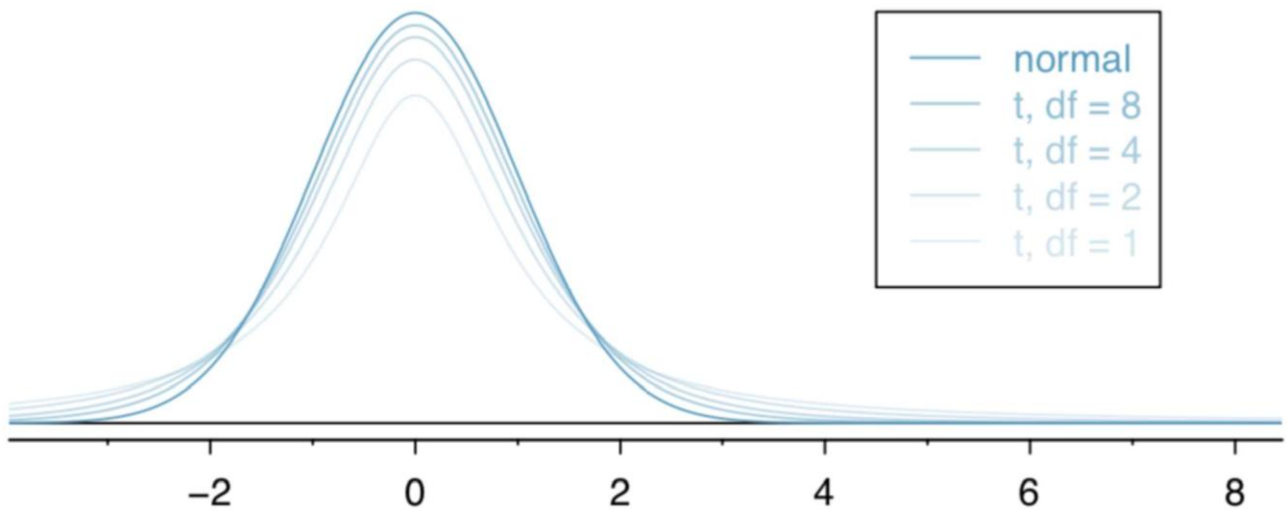
Gosset fue director del departamento de fermentación en la cervecería Guiness en Dublin durante su toda su vida profesional comenzando en 1899

Gosset estaba interesado en mejorar de forma **consistente** y **económica** un barril de cebada

William Sealy Gosset

Es conocido ampliamente por el desarrollo de la distribución **t-Student**

La distribución t es similar a la normal, no obstante, tiene colas mas gruesas y el parámetro que la define son los **grados de libertad**



Los **grados de libertad** se ocupan para contabilizar la **incerteza adicional** por trabajar con muestras pequeñas, esto se calcula como:

$$gl = n - 1$$

Para estimaciones un promedio

$$gl = n_1 + n_2 - 2$$

Para estimaciones de Una diferencia de promedios

¿Cuándo ocupamos la distribución de t-Student?

Independencia de las observaciones (lo mismo que hemos visto para las otras pruebas)

Observaciones proviene de una distribución casi normal (la distribución t es bastante robusta a la asimetrías)

Cuando no conoces la desviación estándar de la población (..casi siempre) y el número de muestras es pequeño ($n < 30$) que no te permite estimar adecuadamente este parámetro

La prueba de hipótesis de un promedio con **test de t** es **muy parecido** a la **prueba de z**, solamente cambia la distribución que consultamos esta vez sujeto a los **grados de libertad**

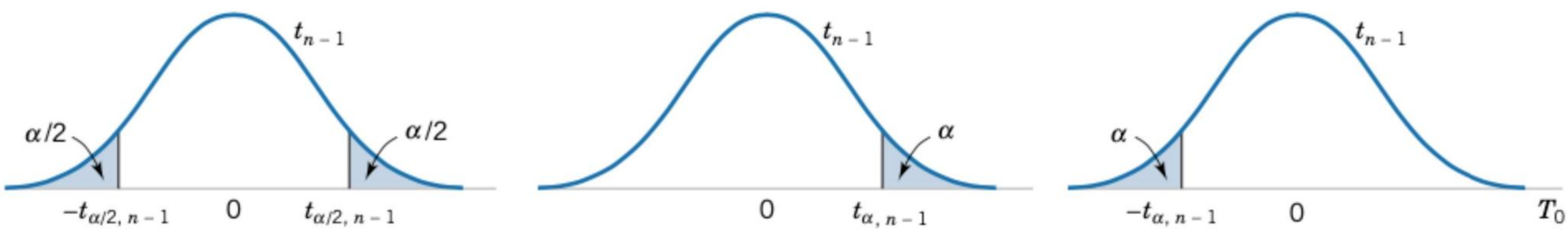
$$gl = n - 1$$

The One-Sample t-Test

Null hypothesis: $H_0\colon \mu = \mu_0$

Test statistic: $T_0 = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$

Alternative hypothesis	Rejection criteria
$H_1\colon \mu \neq \mu_0$	$t_0 > t_{\alpha/2,n-1}$ or $t_0 < -t_{\alpha/2,n-1}$
$H_1\colon \mu > \mu_0$	$t_0 > t_{\alpha,n-1}$
$H_1\colon \mu < \mu_0$	$t_0 < -t_{\alpha,n-1}$



Comparación de dos promedios

¿Qué diferencias tengo con el calculo de Z-test?

Solo cambia el uso de los **grados de libertad** como se menciono antes y la **estimación del SE** en comparación de dos promedios independientes

$$T = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{V(\bar{x}_1 - \bar{x}_2)}}$$

Para estimaciones de Una diferencia de promedios

Definition: The Two-Sample or Pooled *t*-Test*

Null hypothesis: $H_0\colon \mu_1 - \mu_2 = \Delta_0$

Test statistic: $$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (10\text{-}14)$$

Alternative Hypothesis	Rejection Criterion
$H_1\colon \mu_1 - \mu_2 \neq \Delta_0$	$t_0 > t_{\alpha/2,n_1+n_2-2}$ or $t_0 < -t_{\alpha/2,n_1+n_2-2}$
$H_1\colon \mu_1 - \mu_2 > \Delta_0$	$t_0 > t_{\alpha,n_1+n_2-2}$
$H_1\colon \mu_1 - \mu_2 < \Delta_0$	$t_0 < -t_{\alpha,n_1+n_2-2}$

$$gl = n_1 + n_2 - 2$$

Condiciones de prueba estadística t (t-test)

Independencia

La muestra del grupo consiste en observaciones independientes (asignación aleatoria o muestra aleatoria)

Revisar independencia entre grupos, si hay dependencia ocupar **t-test pareado**

Asimetría

Revisar normalidad con qqplots o con **prueba de Shapiro-Wilk**

Varianza semejante (homocedasticidad)

Revisar varianza semejante mediantes comparación visual en boxplots o mediante una **prueba de F o prueba de Levene**. No obstante, la **prueba t de Welch** es una adaptación de la usual para aceptar varianzas desiguales

Es **recomendable** revisar homocedasticidad de varianzas cuando se quiera utilizar el **test de t student**, si no puedes proceder a saltarlo y ocupar la **adaptación de Welch**

¿Por qué debo revisar homocedasticidad para prueba de t student?

Volviendo a la ecuación de SE combinada:

$$V(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n1} + \frac{\sigma_2^2}{n2}$$

Si $\sigma_1 = \sigma_2$:

$$V(\bar{x}_1 - \bar{x}_2) = \sigma^2\left(\frac{1}{n1} + \frac{1}{n2}\right)$$

Si no conocemos σ pero sabemos que son semejantes, entonces el mejor estimador es un combinado equilibrado de las varianzas de ambas muestras

$$gl = n_1 + n_2 - 2$$

The **pooled estimator** of σ^2, denoted by S_p^2, is defined by

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

¿Qué sucede si no tiene varianzas semejantes?

En este caso volvemos atrás y nos quedamos con la ecuación inicial para la comparación de muestras:

$$V(\bar{x}_1 - \bar{x}_2) = \frac{s_1^2}{n1} + \frac{s_2^2}{n2}$$

Pero, para la estimación de **T** se ajustan los **grados de libertad** por la **ecuación de Welch-Satterthwaite:**

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

En general la **prueba de t** es más **precisa** que la **adaptación de Welch** cuando se cumplen todas las condiciones, es menos robusta a las desviaciones mientras la de **Welch** mantiene un **buen performance**

. . . ¿y los intervalos de confianza?

Margen de Error para varianzas **semejantes**

$$= t_{\alpha/2,n1+n2-2} Sp \sqrt{\frac{1}{n1}+\frac{1}{n2}}$$

If $\bar{x}_1, \bar{x}_2, s_1^2$ and s_2^2 are the sample means and variances of two random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown but equal variances, then a **100(1 − α)% confidence interval on the difference in means $\mu_1 - \mu_2$ is**

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,n_1+n_2-2}\, s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$\leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,n_1+n_2-2}\, s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \qquad (10\text{-}19)$$

where $s_p = \sqrt{[(n_1-1)s_1^2 + (n_2-1)s_2^2]/(n_1+n_2-2)}$ is the pooled estimate of the common population standard deviation, and $t_{\alpha/2,n_1+n_2-2}$ is the upper $\alpha/2$ percentage point of the t distribution with $n_1 + n_2 - 2$ degrees of freedom.

Margen de Error para Varianza **desiguales**

$$= t_{\alpha/2,v} \sqrt{\frac{S_1^2}{n1}+\frac{S_2^2}{n2}}$$

If $\bar{x}_1, \bar{x}_2, s_1^2$, and s_2^2 are the means and variances of two random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown and unequal variances, an approximate $100(1-\alpha)\%$ confidence interval on the difference in means $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 - t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + t_{\alpha/2,v}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \qquad (10\text{-}20)$$

where v is given by Equation 10-16 and $t_{\alpha/2,v}$ is the upper $\alpha/2$ percentage point of the t distribution with v degrees of freedom.

Ejercicio

Evaluar el efecto del ejercicio regular en el peso de los sujetos de estudio

...Recordemos el caso de estudio

Primero tomamos una muestra al **azar** de la población, de esta muestra **asignamos de forma aleatoria** individuos que:

- Entrarán a un programa de **ejercitación específica**
- Aquellos que no ejercitarán y tendrán una actividad **sedentaria**

Buscamos medir el **peso promedio** que tiene cada grupo y los comparamos entre sí mediante alguna herramienta estadística

. . . Algo más realista

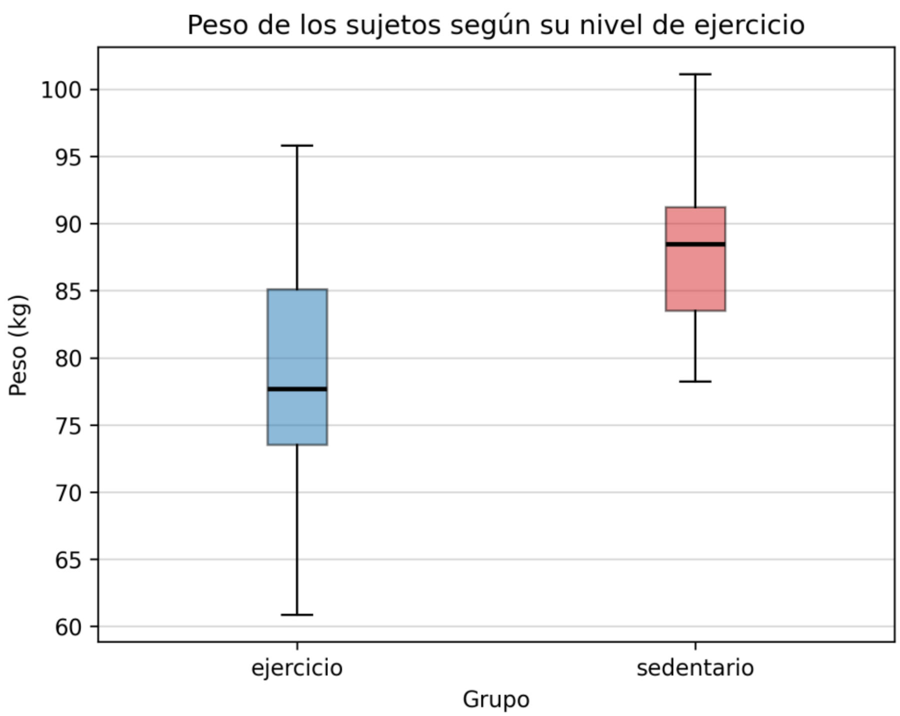
De una muestra de 40 voluntarios **asignamos de forma aleatoria** 20 individuos en:

- Un programa de **ejercitación** mediante una trotadora por 30 min x 3 veces a la semana
- Aquellos que no ejercitarán y tendrán una actividad **sedentaria**

Buscamos saber si hay diferencias en el **peso promedio** entre ambos grupos

¿Qué puedo concluir de los resultados?

Resultados Estudio Experimental



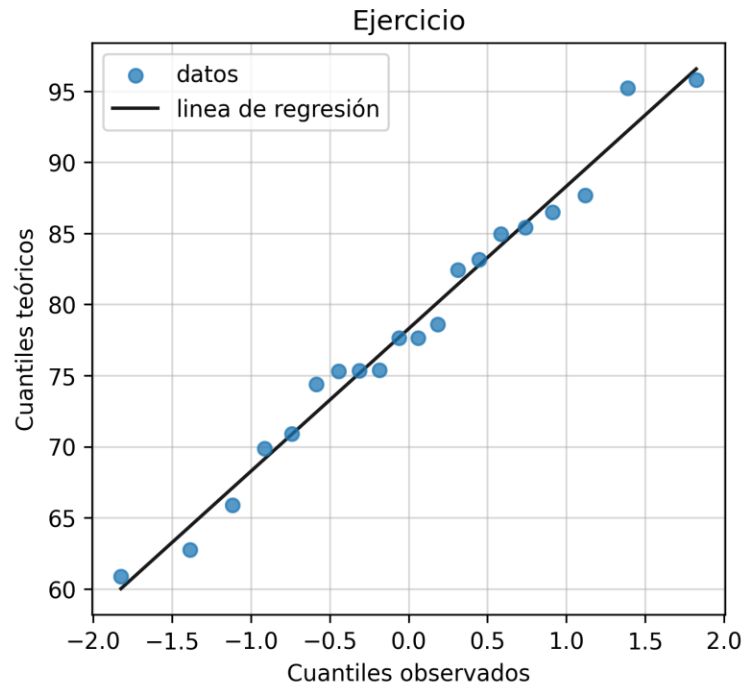
Vamos asumir variantes semejantes dado el gráfico, no obstante, también vamos a ver a posteriori que se pueden hacer un prueba de homocedasticidad de varianzas.

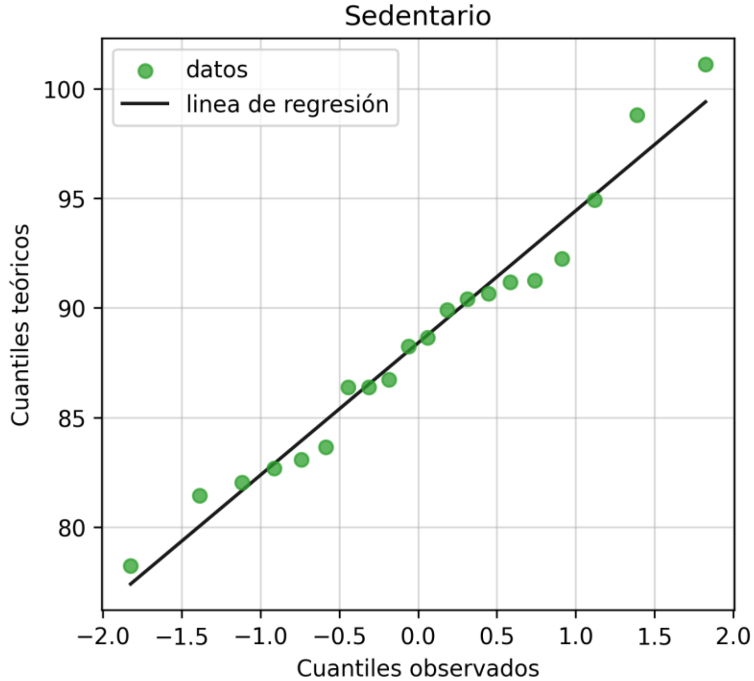
Conclusión estudio

```
# Test de hipótesis
# 1. Hipótesis
# H0: μ_ejercicio  - μ_sedentario = 0
# Ha: μ_ejercicio  - μ_sedentario ≠ 0
# 2. Calcular estimadores puntuales - promedio y erorr estándar
ejercicio = df_ejercicio['ejercicio'].values
sedentario = df_ejercicio['sedentario'].values

mean_ejercicio = np.mean(ejercicio)
mean_sedentario = np.mean(sedentario)
std_ejercicio = np.std(ejercicio)
std_sedentario = np.std(sedentario)
n_ejercicio = len(ejercicio)
n_sedentario = len(sedentario)
diferencias_esperadas = 0
```

```
Promedio ejercicio: 78.29; Promedio sedentadio: 88.40
Desviación estándar ejercicio: 9.36; Desviación estándar sedentadio: 5.66
Tamaño ejercicio: 20; Tamaño sedentadio: 20
```






```
# 3. Verificar supuestos
# Independencia: dada asignación aleatoria
# Normalidad: Revisamos los qqplots
```

Revisar código en Colab de la clase

```
# 4. calculamos el estadístico correspondiente y el valor-p
gdl = n_ejercicio + n_sedentario - 2
Sp = np.sqrt(
    ( ((n_sedentario -1 )*(std_sedentario**2))+ ((n_ejercicio - 1) * std_ejercicio**2) )/gdl
)
Sp_n = Sp * np.sqrt(1/n_ejercicio + 1/n_sedentario )
# calculo valor t
valor_t = (mean_ejercicio - mean_sedentario - diferencias_esperadas)/Sp_n
```

```
Grados de Libertad: 38
Sp: 7.73
Sp_n: 2.45
Valor-t: -4.14
Valor-p: 1.88e-04
```

```
# 5. Calculamos los intervalos de 95% de confianza para la diferencia

x1_x2 = mean_ejercicio - mean_sedentario
t_critico = st.t.ppf(1 - 0.05/2, df = gdl) # t_alpha/2 para ambas colas
corte_inferior = x1_x2 - t_critico * Sp_n
corte_superior = x1_x2 + t_critico * Sp_n
```

```
-15.07 ≤ μ_ejercicio - μ_sedentario (-10.12) ≤ -5.17
```

Con una prueba estadística t de -4.14 y un valor-p de 0.0002, rechazamos la hipótesis nula a un 5% nivel de significancia. Concluimos que hay suficiente evidencia estadística para decir que existe diferencia de peso entre aquellos que hacen ejercicio y no.

Con un 95% de confianza, la diferencia promedio de peso entre los que ejercitan y no está contenida entre -15.07 y -5.17 kg. Por ende, rechazamos la hipótesis nula dado que el intervalo no incluye la diferencia nula entre grupos.

La **adaptación de Welch** al test de t es una alternativa más **robusta** al tradicional que viene por default en la prueba de t

```
--------Prueba t de varianzas semejantes--------
tt_eq = ttest_ind(ejercicio, sedentario, equal_var = True)
Estadístico de la prueba: -4.03
Grados de libertad: 38.00
Valor-p de la prueba: 2.57e-04
```

Prueba de t con opción De varianzas semejantes

```
--------Prueba t de Welch de varianzas desiguales--------
tt_noneq = ttest_ind(ejercicio, sedentario, equal_var = False)
Estadístico de la prueba: -4.03
Grados de libertad: 31.27
Valor-p de la prueba: 3.30e-04
```

Prueba de t de Welch con varianzas desiguales

¿Qué sucede cuando no se cumple la condición de normalidad?

Es posible ocupar una prueba **no-paramétrica** llamada **Mann-Whitney U test** (también llamada Mann-Whitney-Wilcoxon, Wilcoxon-Mann-Whitney, and the Wilcoxon Rank Sum), es una prueba equivalente al test de t pero **con menor poder** debido a los menos supuestos.

```
from scipy.stats import mannwhitneyu
mwu_peso = mannwhitneyu(ejercicio, sedentario, method = 'exact')
```

```
Estadistico: 74.00
Valor-p: 4.19e-04
```

Los resultados de la prueba mantienen las conclusiones que obtuvimos con el test de t.

Test de t pareados

¿Qué sucede cuando hay dependencias entre grupos?

Este se llama prueba estadística **t pareada**. En este caso, buscamos ver la diferencia entre el par de datos:

Table 10-2 Strength Predictions for Nine Steel Plate Girders (Predicted Load/Observed Load)

Girder	Karlsruhe Method	Lehigh Method	Difference d_j
S1/1	1.186	1.061	0.119
S2/1	1.151	0.992	0.159
S3/1	1.322	1.063	0.259
S4/1	1.339	1.062	0.277
S5/1	1.200	1.065	0.138
S2/1	1.402	1.178	0.224
S2/2	1.365	1.037	0.328
S2/3	1.537	1.086	0.451
S2/4	1.559	1.052	0.507

Por ejemplo, si comparamos la estimación de resistencia de vigas de acero entre dos tipos de métodos (Karlsruhe o Lehigh) es posible diseñar un experimento aplicando cada método a la misma muestra, entonces las observaciones ya no son independiente y se consideran pareadas.

El parámetro a evaluar por el test de hipotesis es la diferencia del par: $diff = x1 - x2$

μ_{diff} : Parametro a estimar

$\bar{x}_{diff}$: Estimador puntual

Mismo procedimiento contemplado grados de libertad

$$H_0 : \mu_{diff} = 0$$

$$H_A : \mu_{diff} \neq 0$$

El uso o no del test de t pareado dependerá de como ustedes **diseñen su experimento**, por ejemplo, si tienen limitación de muestra y esta puede ser ocupada nuevamente, o por diferencia en intervalos de tiempo

The Paired *t*-Test

Null hypothesis: $H_0: \mu_D = \Delta_0$

Test statistic: $$T_0 = \frac{\overline{D} - \Delta_0}{S_D/\sqrt{n}} \quad (10\text{-}22)$$

Alternative Hypothesis	Rejection Region
$H_1: \mu_D \neq \Delta_0$	$t_0 > t_{\alpha/2,n-1}$ or $t_0 < -t_{\alpha/2,n-1}$
$H_1: \mu_D > \Delta_0$	$t_0 > t_{\alpha,n-1}$
$H_1: \mu_D < \Delta_0$	$t_0 < -t_{\alpha,n-1}$

1. The parameter of interest is the difference in mean shear strength between the two methods, say, $\mu_D = \mu_1 - \mu_2 = 0$.
2. $H_0: \mu_D = 0$
3. $H_1: \mu_D \neq 0$
4. $\alpha = 0.05$
5. The test statistic is

$$t_0 = \frac{\overline{d}}{s_D/\sqrt{n}}$$

6. Reject H_0 if $t_0 > t_{0.025,8} = 2.306$ or if $t_0 < -t_{0.025,8} = -2.306$.
7. Computations: The sample average and standard deviation of the differences d_j are $\overline{d} = 0.2736$ and $s_D = 0.1356$, so the test statistic is

$$t_0 = \frac{\overline{d}}{s_D/\sqrt{n}} = \frac{0.2736}{0.1356/\sqrt{9}} = 6.05$$

8. Conclusions: Since $t_0 = 6.05 > 2.306$, we conclude that the strength prediction methods yield different results. Specifically, the data indicate that the Karlsruhe method produces, on the average, higher strength predictions than does the Lehigh method. The P-value for $t_0 = 6.05$ is $P = 0.0002$, so the test statistic is well into the critical region.

En el caso de necesitar utilizar una prueba no-paramétrica se puede hacer uso del equivalente **Wilcoxon signed-rank test**

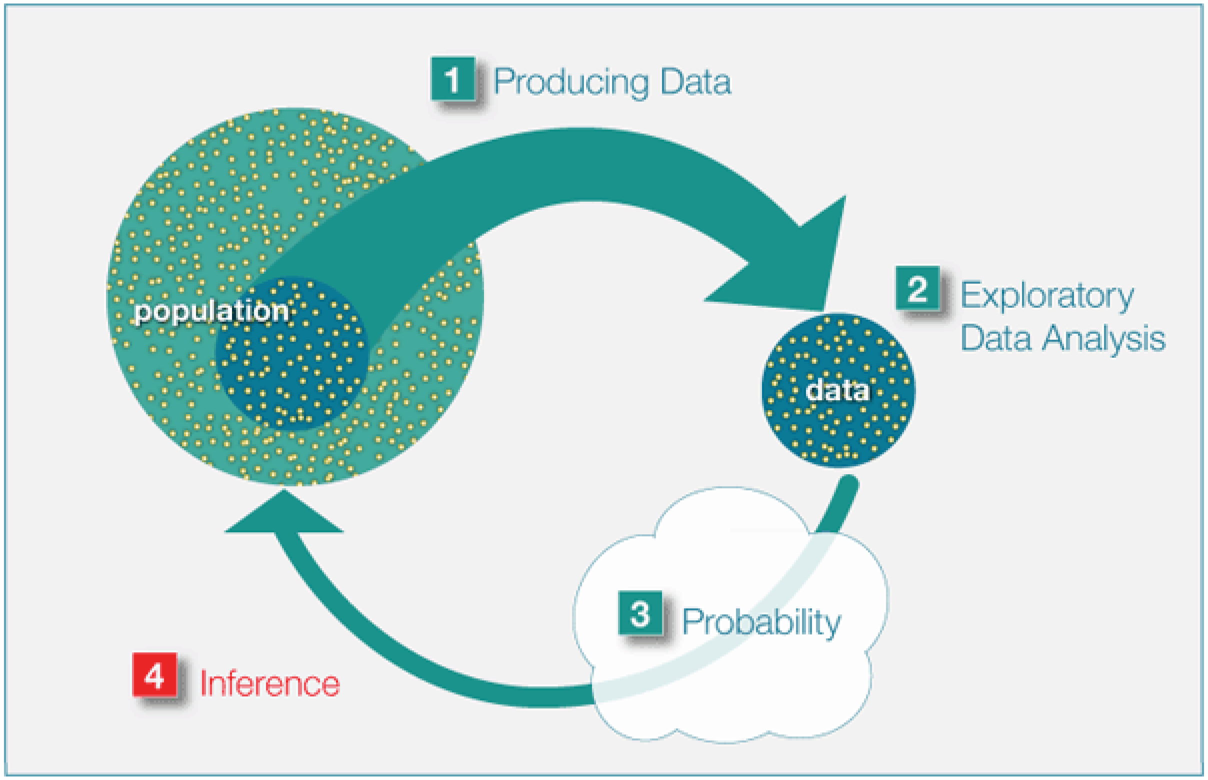
```
# ttest pareado (paramétrico)
tt_rel = st.ttest_rel(a,b)

# wilcoxon (no paramétrico)
wil_test = st.wilcoxon(a,b) # asume por defecto relación
```

Resumen

- Introdujimos formalmente el **test T para el análisis de medias**.
- **El test T se basa en la distribución t de Student y tiene mayor flexibilidad que el estadístico Z.**
- **Su aplicación se justifica en casos en que el número de muestras es pequeño** ($n \leq 30$), y la incertidumbre en la desviación estándar de la población no se puede pasar por alto.
- Analizamos casos de la aplicación del **test T para muestras independientes y pareadas.**

Inferencia Estadística



Profesor: Pedro Saa (pnsaa@uc.cl)
Año: 1-2025