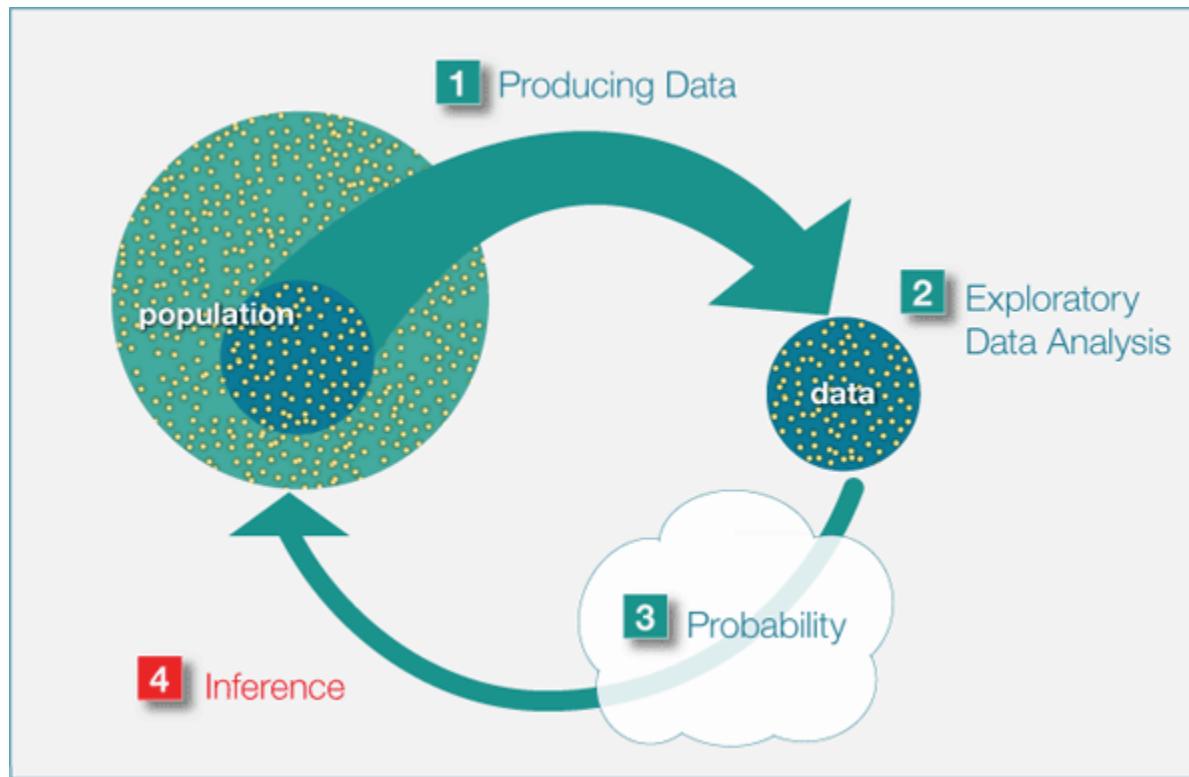


Inferencia Estadística



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 bajo 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**

Poder Estadístico

...¿Cómo lo calculamos?

Caso de estudio

Revisión de la velocidad de combustión de un propelente

Estudio: Los sistemas de escape de la tripulación aérea están impulsados por un propulsor sólido. La velocidad de combustión de este propelente es una característica importante del producto. Las especificaciones requieren que la velocidad de combustión promedio sea de 50 cm/s. Sabemos que la desviación estándar de la velocidad de combustión es 2 cm/s. El investigador decide **especificar una probabilidad de error tipo I o un nivel de significación de 5%** y selecciona una muestra aleatoria de $n= 25$ y obtiene una tasa de combustión promedio de la muestra de $\bar{x} = 51.3$ cm/s (Asuma que esta variable aleatoria se comporta normalmente)
¿Qué conclusiones puede sacar?

Seguiremos el procedimiento de estándar de test de hipótesis

1. Establecer la hipótesis:

- $H_0: \mu = 50$ cm/s
- $H_1: \mu \neq 50$ cm/s

2. Calcular el estimador puntual (\bar{x})

- $\bar{x} = 51.3$ cm/s

3. Verificar supuestos y condiciones

- Independencia: muestra aleatoria $n = 25$
- Normalidad: se asume distribución normal

4. Calcular una prueba estadística correspondiente y valor-p

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{51.3 - 50}{2/\sqrt{25}} = 3.25$$

$$P = \begin{cases} 2[1 - \Phi(|z_0|)] & \text{for a two-tailed test: } H_0: \mu = \mu_0 & H_1: \mu \neq \mu_0 \\ 1 - \Phi(z_0) & \text{for a upper-tailed test: } H_0: \mu = \mu_0 & H_1: \mu > \mu_0 \\ \Phi(z_0) & \text{for a lower-tailed test: } H_0: \mu = \mu_0 & H_1: \mu < \mu_0 \end{cases}$$

$\Phi(z)$ = Distribución normal acumulativa

$$P\text{-value} = 2[1 - \Phi(3.25)] = 0.0012$$

$$50.52 \leq \mu \leq 52.08$$

5. Tomar una decisión e interpretarla en contexto.

- Si valor-p < α, rechazo H_0 , datos proveen evidencia de H_1
- Si valor-p > α, no rechazo H_0 , datos no proveen evidencia de H_1

$$P\text{-value} = 2[1 - \Phi(3.25)] = 0.0012$$

$$50.52 \leq \mu \leq 52.08$$

Con una prueba estadística Z de 3.25 y un valor-p de 0.0012, rechazamos la hipótesis nula a un 5% nivel de significancia. Concluimos que hay suficiente evidencia estadística para decir que la velocidad de combustión del propelente en producción es distinta a 50 cm/s

Con un 95% de confianza, la velocidad de combustión de propálgelante estaría contenido entre 50.52 a 52.08 cm/s. Por ende, rechazamos la hipótesis nula dado que el intervalo no incorpora el valor especificado de 50 cm/s

¿Cuál fue el poder de mi prueba?

Supongamos que la H_0 es falsa y la diferencia entre el valor verdadero y la H_0 sería:

$$\delta = \mu - \mu_0$$

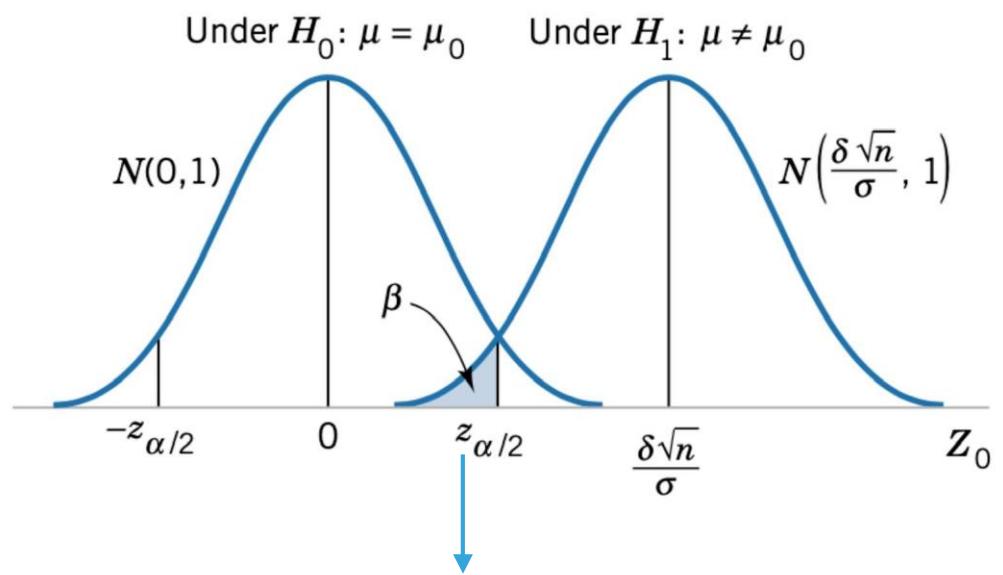
Entonces:

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{\bar{X} - (\mu_0 + \delta)}{\sigma/\sqrt{n}} + \frac{\delta\sqrt{n}}{\sigma}$$

Por ende, la distribución de Z cuando H_1 es verdadera es:

$$Z_0 \sim N\left(\frac{\delta\sqrt{n}}{\sigma}, 1\right)$$

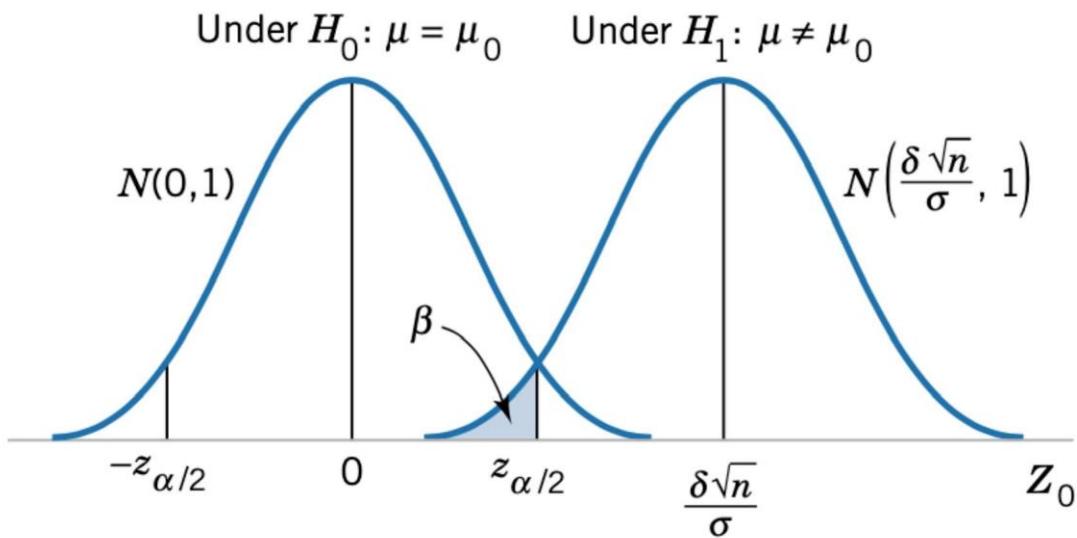
↓
¡Esto es cero para H_0 !



Se define como **error tipo 2** cuando no rechazas la hipótesis nula cuando en realidad la hipótesis alternativa era verdadera, la probabilidad que ocurra es β

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right)$$

¿Cuál fue el poder de mi prueba?



$$z_{\frac{\alpha}{2}} = \mu_0 + z_{critico} * SE$$

$$z_{\frac{\alpha}{2}} = 50 + 1.96 * \frac{2}{\sqrt{(25)}}$$

$$z_{\frac{\alpha}{2}} = 50.78$$

Si H_1 es verdadera entonces:

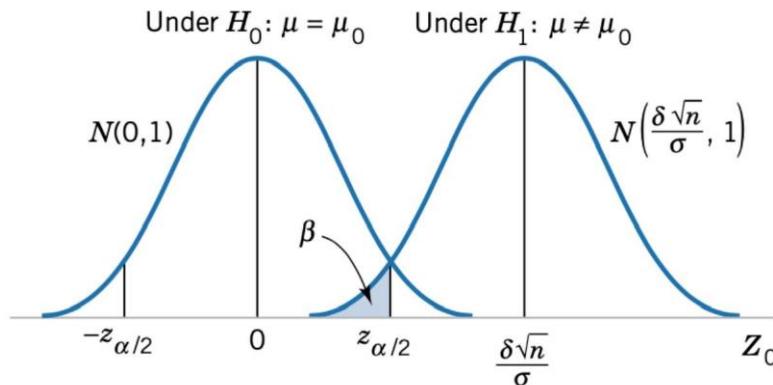
$$z_{h1} = \frac{50.78 - 51.3}{2/\sqrt{(25)}} = -1.29$$

$$\beta = \phi(-1.29) = 0.099$$

$$Poder_{1-\beta} = 90.01 \%$$

Por norma se busca que los test estadístico tengan al menos un **80% de poder estadístico** para ser capaces de rechazar la H_0

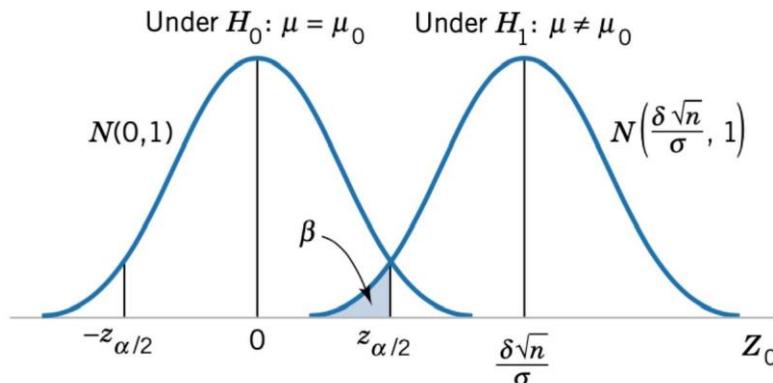
El poder estadístico depende de 3 factores: el **nivel de significancia** con el **tamaño de la muestra** y el **tamaño del efecto** observado en nuestro estudio



↑ *Nivel de significancia* → ↑ *Poder estadístico*

A medida que relajamos el nivel de significancia aumenta la capacidad de rechazar la H_0 (resultados estadísticamente significativo), por tanto aumenta nuestro poder estadístico a costa de aumentar el **error tipo I (α)** y **reducir el error tipo II (β)**

El poder estadístico depende de 3 factores: el **nivel de significancia** con el **tamaño de la muestra** y el **tamaño del efecto** observado en nuestro estudio



↑ **Tamaño de la muestra** → ↑ **Poder estadístico**

Es posible tener control sobre el margen de error esperado, esto implica que podemos manipular el poder de una prueba escogiendo **suficiente muestra** para rechazar nuestra H_0

$$n \simeq \frac{(z_{\alpha/2} + z_{\beta})^2 \sigma^2}{\delta^2}$$

Estudio. Considere el mismo problema del propelente, el analista desea ser capaz de detectar diferencias significativas hasta una tasa de combustión de 1 cm/s, con un alto poder (90%), ¿Cuánta muestra necesitaría?

$$\mu_0 = 50$$

$$\mu_1 = 51$$

$$\sigma = 2$$

$$\delta = 1$$

$$\alpha = 0.05$$

$$\beta = 0.1$$

$$z_{\frac{\alpha}{2}} = 1.96$$

$$z_\beta = 1.28$$

$$n = \frac{(z_{\frac{\alpha}{2}} + z_\beta)^2 \sigma^2}{\delta^2} = \frac{(1.96 + 1.28)^2 2^2}{1^2} = 42$$

El analista requeriría tomar **al menos 42 observaciones** para poder detectar diferencias de 1 cm/s

El poder estadístico depende de 3 factores: el **nivel de significancia** con el **tamaño de la muestra** y el **tamaño del efecto** observado en nuestro estudio

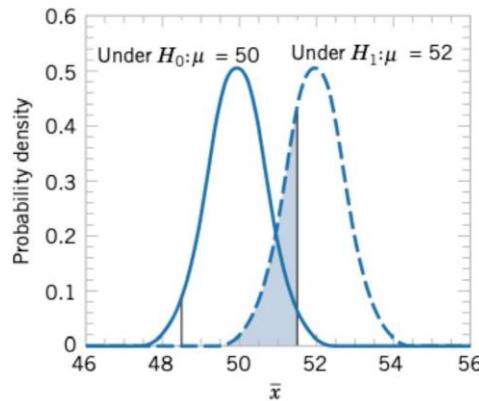


Figure 9-3 The probability of type II error when $\mu = 52$ and $n = 10$.

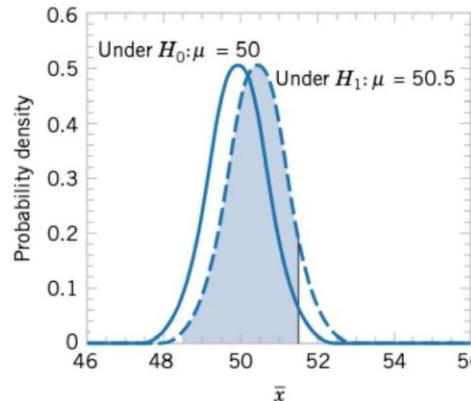


Figure 9-4 The probability of type II error when $\mu = 50.5$ and $n = 10$.



El **tamaño del efecto** es una medida cuantitativa de la magnitud de un fenómeno, en contexto de esta clase, el tamaño del efecto es la diferencia entre nuestra estimación puntual y nuestro valor nulo
No obstante, es más robusto estandarizar este tamaño de efecto para contabilizar la variación natural de las variables de estudio

A medida que aumento mi tamaño de muestra, nuestra estimación puntual se hace más **preciso (menor error estándar)**, permitiendo que cualquier diferencia real entre el nuestro promedio y valor nulo sea más fácil de **detectar (mayor poder)**

Si bien podemos decir que la diferencia es
estadísticamente significativa,
esta puede no ser **prácticamente
significativa**

Using Effect Size—or Why the P Value Is Not Enough

GAIL M. SULLIVAN, MD, MPH
RICHARD FEINN, PhD

Statistical significance is the least interesting thing about the results. You should describe the results in terms of measures of magnitude—not just, does a treatment affect people, but how much does it affect them.

-Gene V. Glass¹

The primary product of a research inquiry is one or more measures of effect size, not P values.

-Jacob Cohen²

These statements about the importance of effect sizes were made by two of the most influential statistician-researchers of the past half-century. Yet many submissions to *Journal of Graduate Medical Education* omit mention of the effect size in quantitative studies while

appears clear, the effect size in the second example is less apparent. Is a 0.4 change a lot or trivial? Accounting for variability in the measured improvement may aid in interpreting the magnitude of the change in the second example.

Thus, effect size can refer to the raw difference between group means, or absolute effect size, as well as standardized measures of effect, which are calculated to transform the effect to an easily understood scale. Absolute effect size is useful when the variables under study have intrinsic meaning (eg, number of hours of sleep). Calculated indices of effect size are useful when the measurements have no intrinsic meaning, such as numbers on a Likert scale; when studies have used different scales so no direct comparison is possible; or when effect size is examined in the context of

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/>

La siguiente tabla es una recomendación de como estimar el **tamaño del efecto** para distintos **tipos de estudios**

TABLE 1

COMMON EFFECT SIZE INDICES^a

| Index | Description ^b | Effect Size | Comments |
|------------------------------------|--|--|---|
| Between groups | | | |
| Cohen's d^a | $d = M_1 - M_2 / s$ $M_1 - M_2$ is the difference between the group means (M); s is the standard deviation of either group | Small 0.2 Medium 0.5 Large 0.8 Very large 1.3 | Can be used at planning stage to find the sample size required for sufficient power for your study |
| Odds ratio (OR) | Group 1 odds of outcome Group 2 odds of outcome If OR = 1, the odds of outcome are equally likely in both groups | Small 1.5 Medium 2 Large 3 | For binary outcome variables Compares odds of outcome occurring from one intervention vs another |
| Relative risk or risk ratio (RR) | Ratio of probability of outcome in group 1 vs group 2; If RR = 1, the outcome is equally probable in both groups | Small 2 Medium 3 Large 4 | Compares probabilities of outcome occurring from one intervention to another |
| Measures of association | | | |
| Pearson's r correlation | Range, -1 to 1 | Small ± 0.2 Medium ± 0.5 Large ± 0.8 | Measures the degree of linear relationship between two quantitative variables |
| r^2 coefficient of determination | Range, 0 to 1 ; Usually expressed as percent | Small 0.04 Medium 0.25 Large 0.64 | Proportion of variance in one variable explained by the other |

Caso. Un caso habitualmente citado para este tipo de problema es el estudio del *Physicians Health Study* sobre el consumo de aspirina para prevenir infartos al miocardio (MI). Se estudiaron 22.000 individuos sobre una ventana de 5 años. Se observó que el grupo que consumía aspirina estaba asociado a una **reducción significativa de MI** (No así mortalidad cardiovascular), con un valor-p de **0.00001**. El estudio terminó tempranamente debido a la evidencia conclusiva que se observó, y la aspirina se recomendó como método de prevención general. No obstante, el **tamaño del efecto era extremadamente pequeño, la diferencia de riesgo entre los grupos era OR = 0.77%**. Como resultado del estudio, muchas personas comenzaron a consumir aspirina sin experimentar un beneficio importante y expuesto a los efectos adversos. Estudios posteriores encontraron que el efecto era aún menor, por lo que la recomendación de la aspirina ha sido modificada.

Bartolucci AA, Tendera M, Howard G. Meta-analysis of multiple primary prevention trials of cardiovascular events using aspirin. Am J Cardiol. 2011;107(12):1796-801.

Buenas prácticas estadísticas

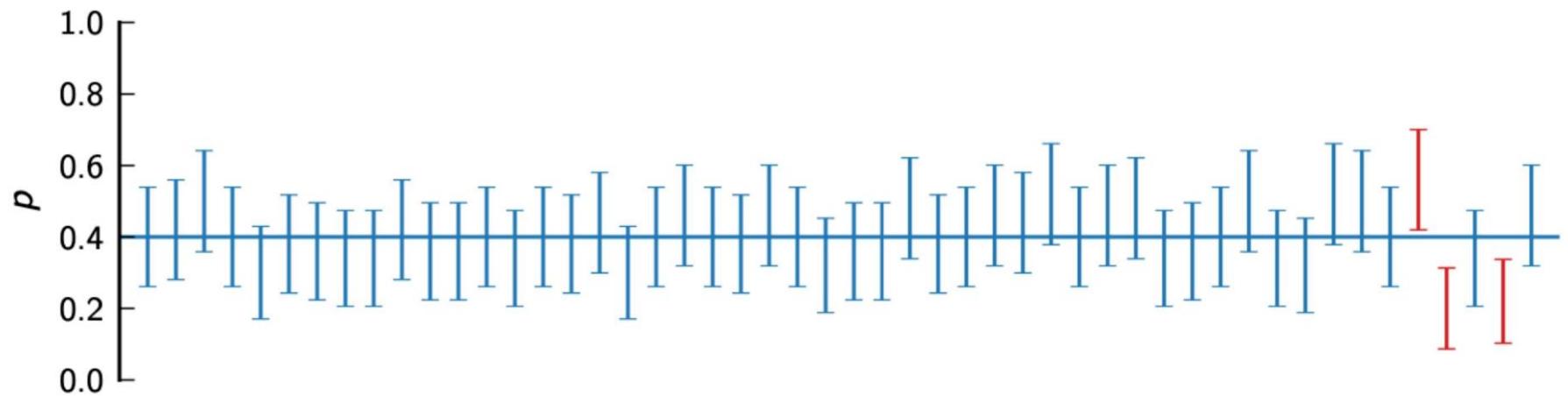
Verificar supuestos de los datos (**independencia y normalidad**) y reportarlos

Reportar **p-valor, intervalos de confianza, errores estándar** y el **procedimiento** que se hizo para el muestreo

¿Es relevante? Reportar el **tamaño de efecto** del fenómeno de estudio

Evitar el **p-hacking** o **data dredging** o **data snooping**, sea todos los métodos estadísticos que buscan descubrir resultados significativos. Sean estos, influenciar el muestreo, modificar la hipótesis alternativa en luz de los resultados, volver a muestrear continuamente hasta obtener un valor significativo, probar distintas hipótesis alternativa con una misma muestra

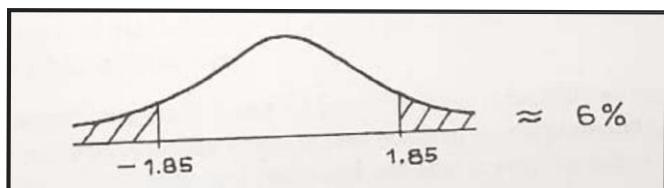
¿Por qué es relevante el **data snooping**? En caer en este tipo de prácticas provocan que las conclusiones sean difíciles de **interpretar y no reproducibles**



El nivel de significancia admite un 5% de error que puede suceder naturalmente...

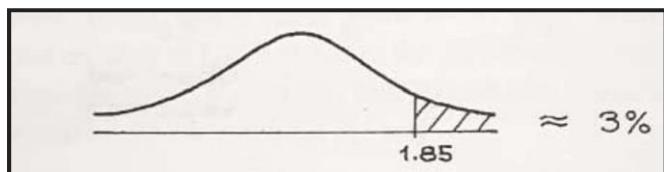
¿Por qué es relevante el **data snooping**? En caer en este tipo de prácticas provocan que las conclusiones sean difíciles de **interpretar** y **no reproducibles**

Imaginen que un investigador quiere saber el efecto de un tratamiento X sobre el rendimiento de un proceso. Su H_1 es distinto al nulo ($\mu \neq \mu_0$). Luego de su experimento y su análisis le da $z = 1.85$.



No cruza el nivel de significancia estandar (5%)

El investigador en vez de refinar su diseño experimental, obtener mas datos o afinar sus métodos analíticos **podría cambiar H_1** y volverlo un test de **una cola**



En principio da igual el test que aplique mientras comunique lo que hizo, no obstante, es discutible

Redefine statistical significance

We propose to change the default P -value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.

Daniel J. Benjamin, James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman and Valen E. Johnson

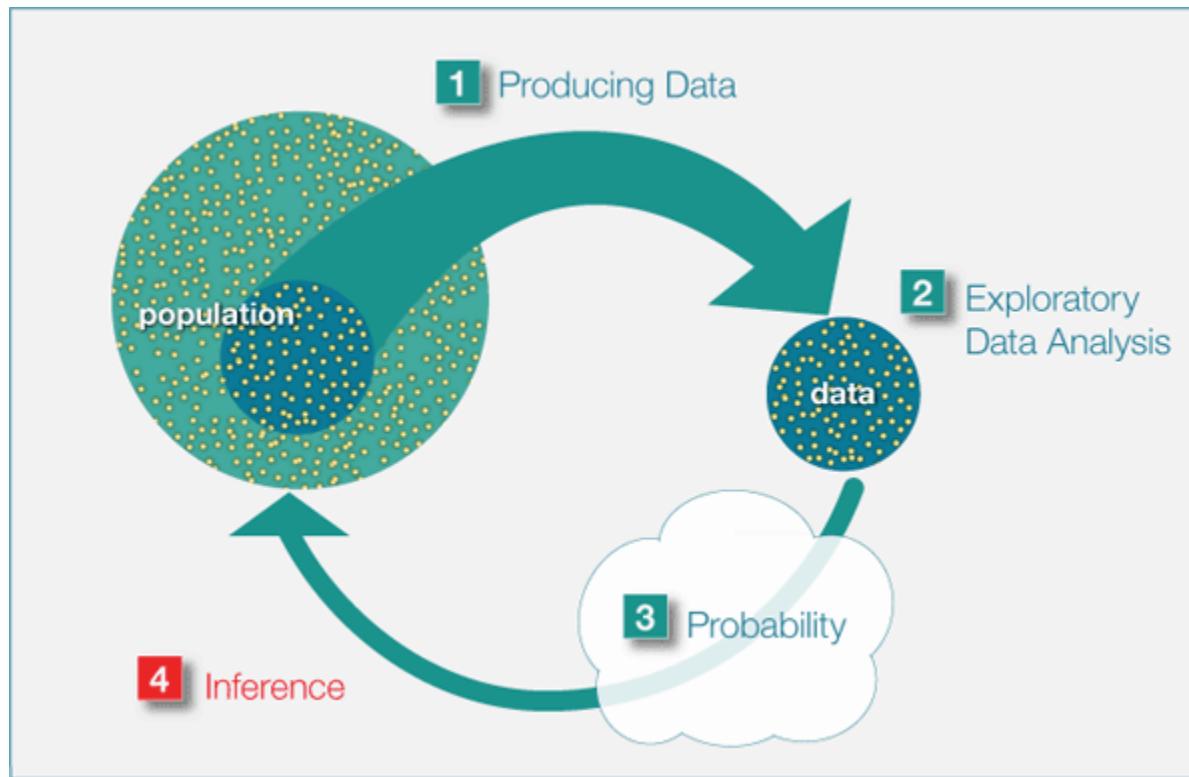
<https://www.nature.com/articles/s41562-017-0189-z.pdf>

<https://www.vox.com/2016/3/15/11225162/p-value-simple-definition-hacking>

Resumen

- Introdujimos técnicas para estimar el **poder estadístico de un test de hipótesis**.
- El **poder estadístico tiene se relaciona con la capacidad de detectar un efecto real**.
- El **poder estadístico depende del tamaño de la muestra, nivel de significancia y tamaño del efecto**.
- Es **difícil en general reducir la probabilidad de error tipo I sin aumentar la probabilidad de error tipo II**.

Inferencia Estadística



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