Biostatistics BootCamp Lecture 1:

Statistical Uncertainty and Hypothesis testing: Errors, p-value & Confidence Intervals

August 9, 2023

Wonsuk Yoo, PhD

Ivy Brain Tumor Center | Barrow Neurological Institute

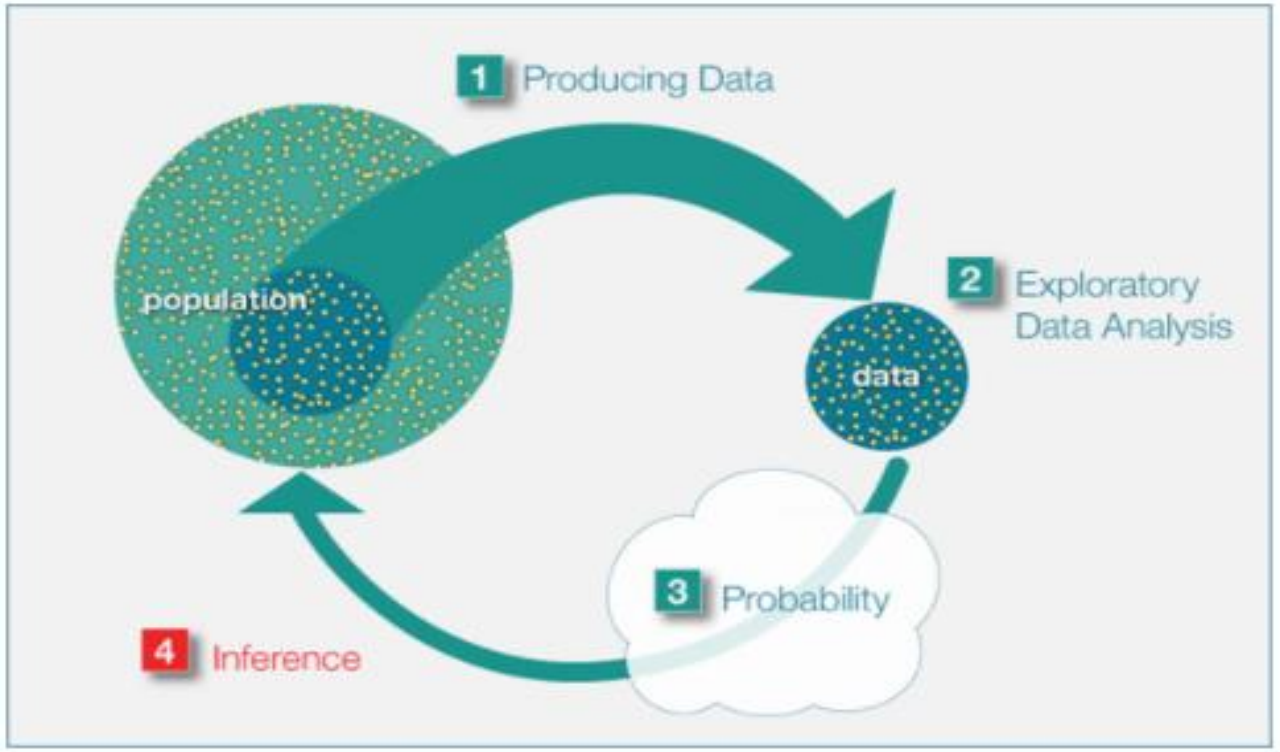
A scientific research based on the population investigation:

- What is a scientific research?

  ➡ It is research performed by a systematic and scientific method to obtain, analyze and interpret (experimental) data.

- Let’s assume we are planning a scientific research about a new treatment affecting a biomarker on Parkinson’s disease (PD).
- What is the study population and the parameter? How can we prove the scientifically significant efficacy of our new treatment?
- Is the whole population-based investigation possible? And if it is not possible to measure all subjects in population, what should we do?

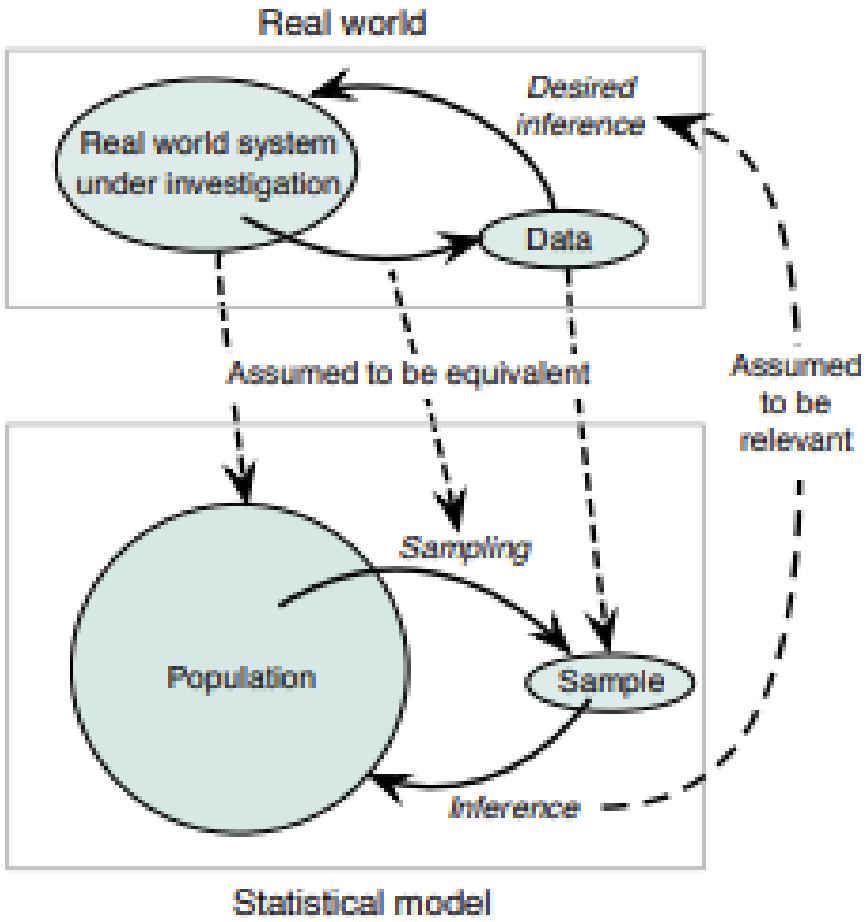
Scientific research and statistical inference:



Open learning Initiative at Carnegie Mellon University, https://oli.cmu.edu/courses/concepts-of-statistics/

- Define the **target population** and the **characteristics of our interest** based on research goal.
- Generate "**sample data**" through experiment: Experimental design
- Calculate the **sample statistics** on sample data: Exploratory data analysis
- **Infer the population parameters** based on the observed statistics using probability theory.
- What is the most important on experiment and inference?

Statistical Uncertainty and Hypothesis Testing:



Liu et al (2016). A reckless guide to P-values: local evidence, errors

- We have two fundamental results:
- First, we do use the sample data rather than whole population data. It is not usually feasible to include all individuals from a target population in a scientific study.

  ➡ Statistical uncertainty
- Second, we may think that our data (the founding) looks scientifically meaningful. But people have different opinions by looking at data. Some people think our findings look meaningful, but others do not. Thus, we need to prove that our findings are scientifically meaningful? How?

  ➡ Hypothesis testing

Statistical Uncertainty

- What is the ideal scenario/best strategy for your research?
- Is it possible to include the whole population patients in a study? Why do we use a sample rather than a population?
- We calculate the (sample) statistics based on the sample data. That is, the scientific results/findings depend on sub-population (sample), not whole population.

  ➡ When we draw statistical inference, we encounter the **statistical uncertainty**.
- How can we measure statistical uncertainty?

  ➡ We need the statistical inference through hypothesis testing procedures by quantifying the statistical uncertainty.

Why do we perform hypothesis testing?

- Through the experiment, the researchers have sample data and expect that the data (their findings) look scientifically meaningful. But ….
- But our results include the statistical uncertainty because our results are obtained through lack of full information. Thus, we need to quantify the uncertainty.
- Thus, our findings need to be supported by objective and scientific evidence.

➡ A hypothesis testing procedure provides us a uniform decision-making criterion that is consistent for all people. How? ….

➡ We make decision on our findings by quantifying the statistical uncertainty

➡ Now our question is what is the statistical uncertainty in hypothesis testing?

Uncertainty in Hypothesis testing (I)

		Hypothesis	
		Null (H_0)	Alternative (H_1)
Decision	Accept H_0 (null)	Good	$\boldsymbol{\beta}$
	Accept H_1 (alter)	$\boldsymbol{\alpha}$	Good

- We always encounter two types of uncertainty (i.e., errors, α and β) in hypothesis testing.

➡ $\text{Type 1 error } (\alpha) = \Pr(Reject\ H_0 | H_0 \text{ is true})$
$= \text{False positive}$

➡ $\text{Type 2 error } (\beta) = \Pr(Reject\ H_1 | H_1 \text{ is true})$
$= \text{False negative}$

Uncertainty in Hypothesis testing (II)

		Hypothesis	
		Null (H_0)	Alternative (H_1)
Decision	Accept H_0 (null)	Good	$\boldsymbol{\beta}$
	Accept H_1 (alter)	$\boldsymbol{\alpha}$	Good

- What should we do for a good study design?
  - ➡ Thus, our goal is to minimize the two errors.
  - ➡ What's the relationship between α and β?
  - ➡ In practice, we usually set α=0.05 ($\alpha \leq 0.05$) and minimize the β.

Uncertainty in Hypothesis testing (III)

		Hypothesis	
		Null (H_0)	Alternative (H_1)
Decision	Accept H_0 (null)	Good	β
	Accept H_1 (alter)	α	Good

- What does the α=0.05 mean?
  - ➡ When we assume to perform the study 20 times, one study might make one wrong conclusion.
- In strictly speaking, we need to express "α" is less than and equal to 5%.

Uncertainty in Hypothesis testing (IV)

		Hypothesis	
		Null (H_0)	Alternative (H_1)
Decision	Accept H_0 (null)	Good	$\boldsymbol{\beta}$
	Accept H_1 (alter)	$\boldsymbol{\alpha}$	Good

- Second, we need to minimize the type 2 error (β).

  $\boldsymbol{\beta} = \Pr(Reject\ H_1 | H_1 \text{ is true})$ ➡ False negative

  $= 1 - \Pr(Accept\ H_1 | H_1 \text{ is true})$ ➡ $1 - \boldsymbol{True\ Positive}\ (\boldsymbol{TP})$

  TP $= \Pr(Accept\ H_1 | H_1 \text{ is true})$ ➡ **Power (= 1- $\boldsymbol{\beta}$)**

- To minimize type 2 error means to maximize the (study) power.

Uncertainty in Hypothesis testing (V) :

- We know these two errors should be pre-specified in the study protocol before the study. Why?
- The information related to the two errors I protocol is a study experimenter's expectation and/or belief on their studies.
- After completing the experiment, we should assess whether our expectations are acceptable. How? The data shows everything!
- We evaluate the pre-specified errors through the data in hypothesis testing procedure. We need to estimate/quantify the errors with the data.

  ➡ Yes, now we will discuss about the **p-value**.

p-value (I): type 1 error applied by data

- The type 1 error (α) is the probability of accepting the alternative hypothesis when the null is true, which means the prespecified false positive rate.

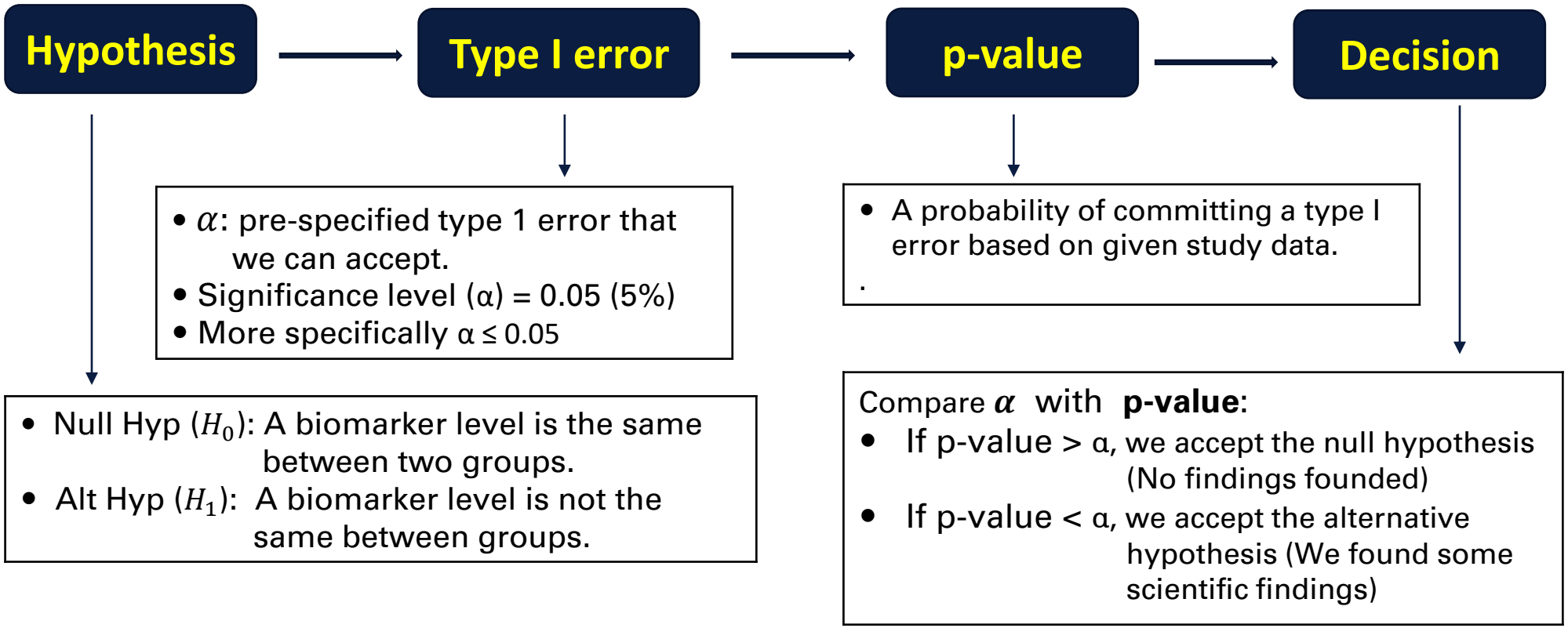
$$\alpha = P(accept\ H_1 |\ H_0\ is\ true)$$

- The p-value is calculated under the assumption that H_0 is true and represents the probability of the given parameter(mean) of the sample being observed given H_0 is true.

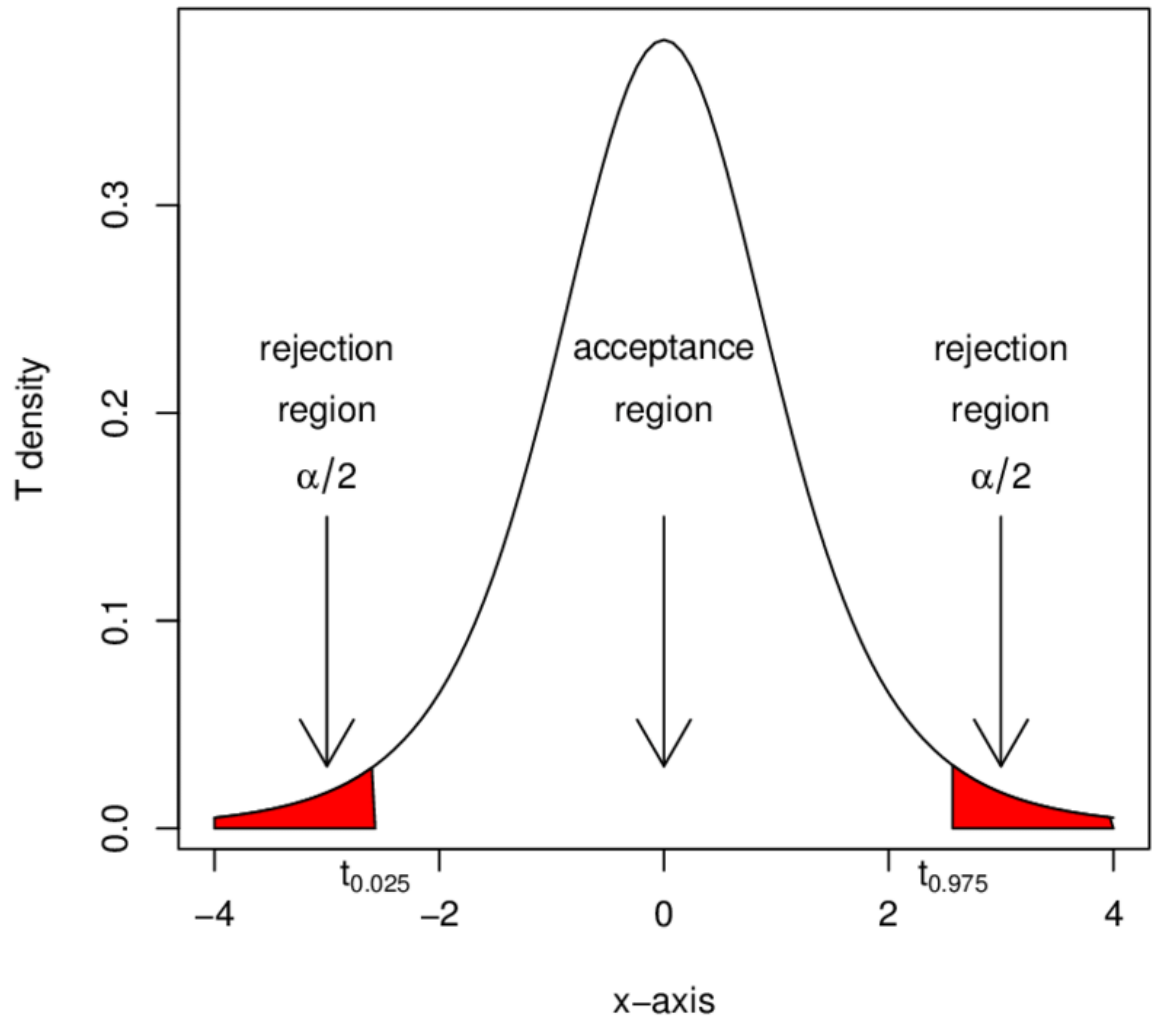
$$p - value = P(|Z| \geq TS |\ H_0\ is\ true)$$

- The α is a pre-specified cutoff error, while the p-value is the type I error based on our experimental data.

The procedure of Hypothesis testing



The procedure of Hypothesis testing (cont.)



p-value (II): if we fail to reject H_0?

- In superior tests, we want to have the significant results.

  ➡ Reject the H_0 (no treat effect), so accept the H_1(significant treat effect).

- We need small p-values: The smaller the p-value, the more unlikely it is for the null hypothesis (say, no treatment effect) to be true in light of the data that will be collect.

- If one fails to get a significant result, it is not correct to conclude that there is no treatment effect. Why?

  ➡ We can say that we have not enough evidence to be able to reject the null hypothesis. Why?

p-value (III): Problem with hypothesis tests

- The p-value is a function of sample size when the null hypothesis is false.
- When $H_1\ is\ true$, a statistically significant result can always be obtained if a sufficiently large sample is used.

  ➡ The p-value does not relate to the scientific importance of findings.
- What is the scientific finding?

  ➡ The "**effect size (ES)**" means the degree to which the phenomenon is present in the population

  ➡ clinically meaningful difference of the primary endpoint.

		Effect size	
		Large	Small
Statistical significance	Small p-value (< 0.05)	No problem	Mistaking statistical significance for scientific importance
	Large p-value (>0.05)	Failure to detect a scientific effect	No problem

- A large study may find a <u>small & unimportant ES</u> that's highly significant
- A small study may <u>fail to find important differences</u> even though the study has large effect size.
- In summary, statistical significance does not necessarily mean the result is clinically significant.
- We need to overcome the problem of p-value in hypothesis testing.
  ➡ Yes, now we will discuss about the confidence interval (CI).

Confidence intervals (I): Why does CIs be needed?

- "p-value=0.05" is an artificial cut-value between significant and non-significant results
- "p < 0.05" or "p=NS" do not describe the results of a study well.
  - ➡ P-value does not tell us what the population mean difference is or how large the mean difference is.
  - ➡ P-values do not measure the success of the study.
- To solve the problem of p-value, we need to supplement the hypothesis test with a confidence interval (CI).

Confidence intervals (II): What are CIs?

- Def: A confidence interval is a range of values calculated from the data that will contain the true population mean difference (true ES).
  - ➡ That is, CI is the range of values in which the investigator believes a true mean difference will lie with some probability.
- What are the 95% confidence intervals?
  - ➡ The chance that the true value lie in the interval is 95%
  - ➡ The interval would enclose the true value 95% of the times if the experiment are repeated a large number of times.

Confidence intervals (III): Magnitude of ES

- The CI gives an estimate of the precision with which a statistic estimates a population value, which is useful information for the reader.

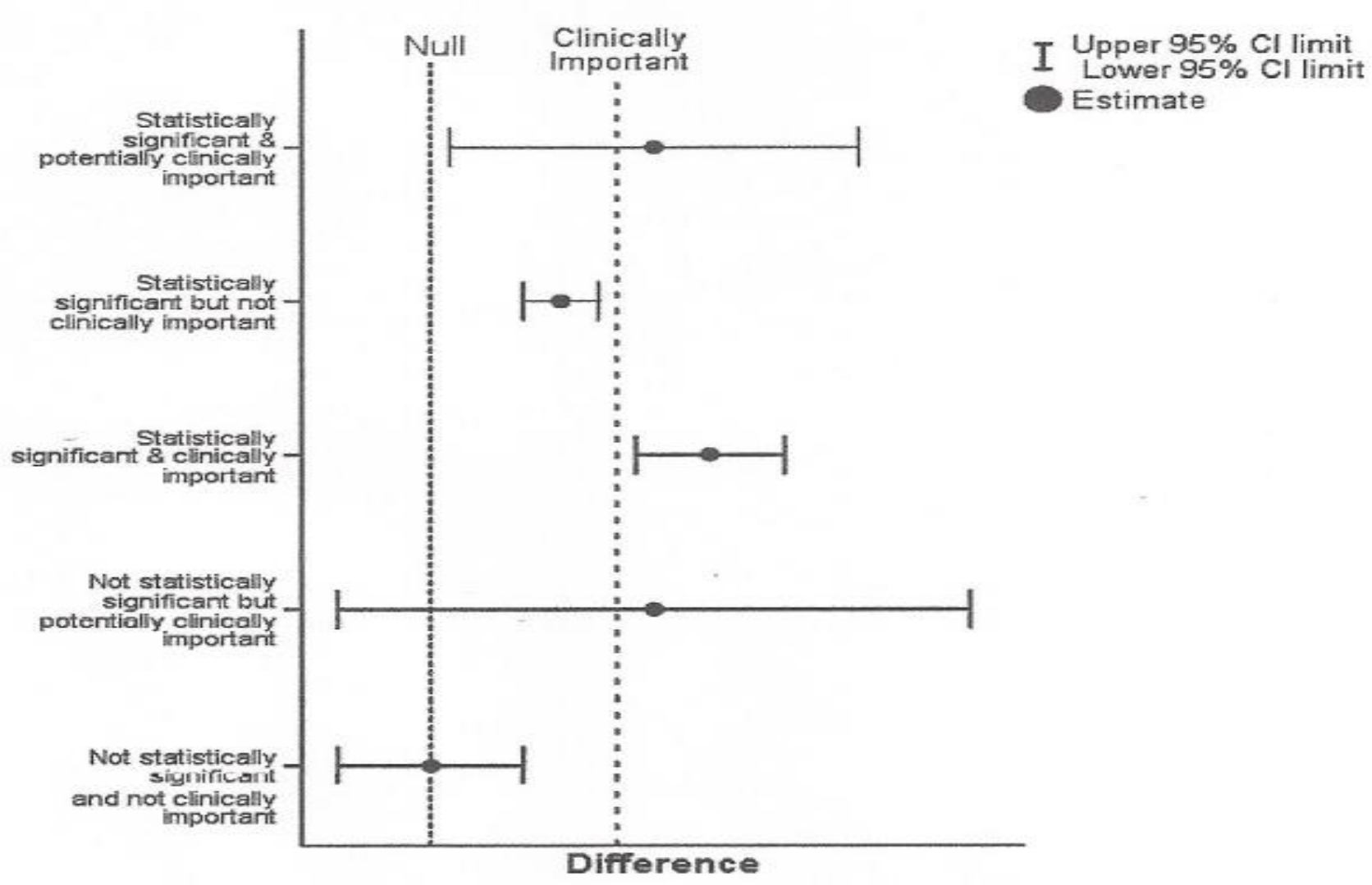
  ➡ CIs measure the uncertainty in the estimate of interest.

- The CIs provide the magnitude of the true difference of interest scientifically or clinically.

  ➡ CIs help distinguishing statistical significance from scientific importance.

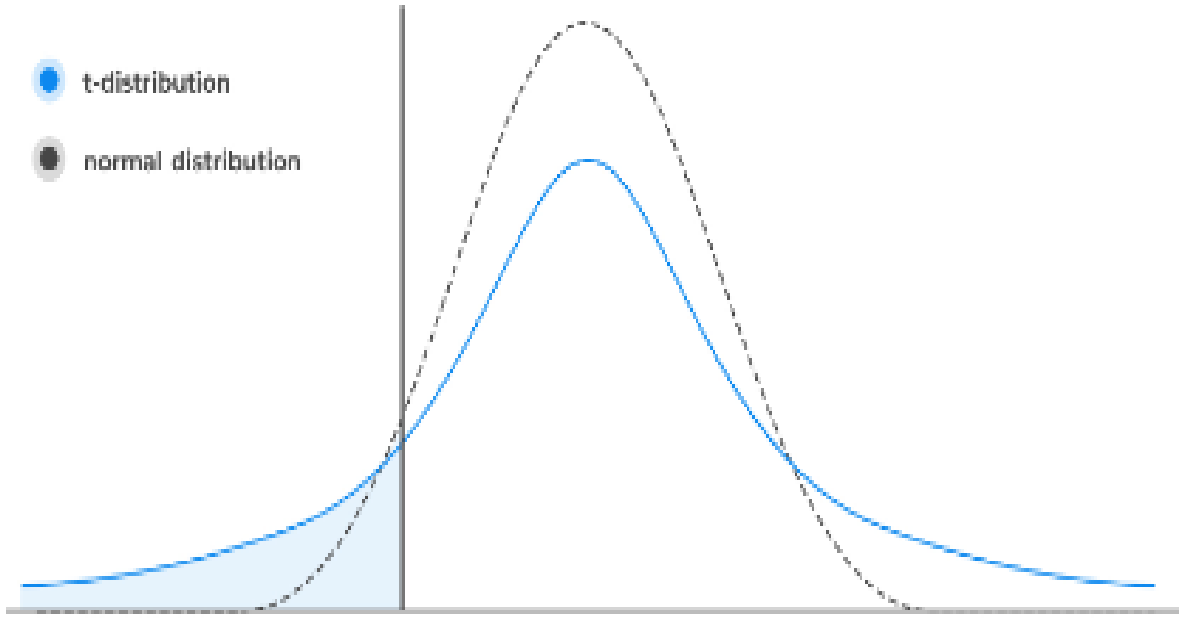
  ➡ CIs convey useful information (the magnitude and precision of the difference). So, CI should always be reported with p-value.

CIs (IV): Statistical& Clinical Importance



Thought 1: Normal .vs. t-distribution

The critical value at 97.5 percentile in normal distribution is $Z_{0.975}$=**1.96** while the critical value at 97.5 percentile in t-distribution is $t_{0.975}$ = **2.23**. Even though the t-distribution seems to be similar to the normal for more than 30 samples, why do these two numbers be different?



Thought 2:

The confidence interval (CI) conveys more useful information that p-values. The CIs provide the magnitude of the true difference of interest scientifically or clinically. How can we the range/interval in the magnitude of the true effect size (i.e., mean difference of two groups) interpret?

Conditions for A good study design:

- During the hypothesis testing, we set $\alpha = 0.05$ (fixed) and minimize β in practice.
- Conditions for a good experimental design are the study designs that
  (1) controlling α up to 5% and
  (2) maximizing the study power (1-β).
- Now what scenarios should we avoid in the hypothesis testing?
  ➡ [1] Study designs with inflated type 1 error: $\alpha > 0.05$
  ➡ [2] Study designs with low study power

Final Comments:

- When performing the hypothesis testing, we encounter statistical unceratinty.
- The uncertainty is pre-specified using two errors (α and β), which are quantified using p-value and the power through hypothesis testing.
- For good research designs, we need to control the type 1 error (at 5%). The multiplicity issues inflate the type 1 error rate.
- The p-value has a fundamental problem. Why? It is a function of sample size when the alternative hypothesis is true.
- A large study may find a highly significant but small/unimportant difference, while a small study may fail to find important differences.
- To solve the problem of p-value, we need to supplement the hypothesis test with a confidence interval (CI). The CI provides the magnitude of the true difference of our interest scientifically or clinically.

Further reading:

Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. **Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations**. Eur J Epidemiol. 2016 Apr;31(4):337-50. doi: 10.1007/s10654-016-0149-3. Epub 2016 May 21. PMID: 27209009;. URL: Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations - PubMed (nih.gov)

Price R, Bethune R, Massey L, **Problem with p values: why p values do not tell you if your treatment is likely to work**, *Postgraduate Medical Journal*, Volume 96, Issue 1131, January 2020, Pages 1–3, https://doi.org/10.1136/postgradmedj-2019-137079

Althouse, A.D., Soman, P. Understanding the true significance of a *P* value. *J. Nucl. Cardiol.* **24**, 191–194 (2017). https://doi.org/10.1007/s12350-016-0605-1


Q&A