2020

Clear-Sighted Statistics: Module 13: Introduction to Null Hypothesis Significance Testing (NHST) (slides)

Edward Volchok
CUNY Queensborough Community College

Introduction to NHST

(Null Hypothesis Significance Testing)

Module 13

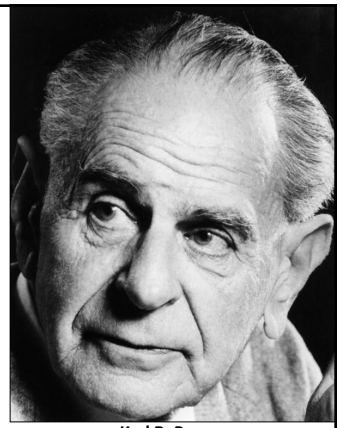
CLEAR-SIGHTED STATISTICS

EDWARD VOLCHOK

0

"…I shall not require of a scientific system that it shall be capable of being singled out once and for all, in a positive sense; but I shall require that its logical form shall be such that it can be singled out, by means of empirical tests, in a negative sense: it must be possible for an empirical scientific system to be refuted by experience."

Falsification, nullification, refutation



Karl R. Popper
(1902 – 1994)

Karl R. Popper, *The Logic of Scientific Discovery*, (Mansfield Centre, CT: Martino Publishing, 2014), pp. 40-41

1

1

Lecture Objectives

Define Null and Alternate Hypotheses	Define the significance level	Define Type I (α) errors
Define Type II (β) errors	Define statistical power	Describe the purpose of a decision rule
Define p-values	List the steps in the testing process	

2

2

The Origins of NHST

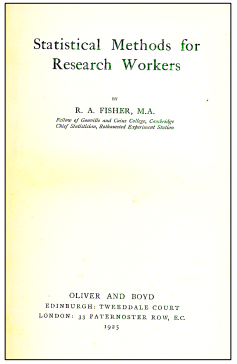
3

3

1925: Fisher published the first edition of *Statistical Methods for Research Workers**

Laid the foundation for "significance tests"

Only one hypothesis
(Today NHST has 2 hypotheses)


Statistical Methods for Research Workers

BY

R. A. FISHER, M.A.

Fellow of Gonville and Caius College, Cambridge
Chief Statistician, Rothamsted Experimental Station

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 33 PATERNOSTER ROW, E.C.
1925

*The 14th and final edition was published in 1970, 8 years after Fisher's death

1928: Jerzy Neyman and Egon Pearson introduced Hypothesis Tests

Goal: Correct flaws in Fisher's approach

Added a 2nd hypothesis, Alternate Hypothesis (H_1)

Defined Type I or α Errors (false positive), and

Type II or β errors (false negative)

Jerzy Neyman, and Ego S. Pearson, E. S. "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I." *Biometrika* 20A, 1928, pp. 175–240.

Fisher and Neyman waged an acrimonious debate about the merits of their approaches until Fisher's death in 1962

Textbook authors combined Fisher's significance testing and Neyman-Pearson's hypothesis testing into a "unified" approach

Not everyone agrees that this amalgam is "unified"

"It might be tempting to regard a mixture of the two approaches [Fisher and Neyman-Pearson] as possibly combing the best of both worlds, but the two frameworks are based on incompatible conceptions of probability. The mixture is indeed incoherent, and so it's not surprising that misconceptions about NHST are so widespread."

-- Geoff Cumming

Geoff Cumming, *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. (New York: Routledge, 2012), p. 25.

8

What NHST does

Uses sample data and probability theory to determine whether a proposition about population data should be rejected

Does not verify a hypothesis

9

NHST: Only 2 possible decisions

There is insufficient evidence to reject the Null Hypothesis

There is sufficient evidence to reject the Null Hypothesis

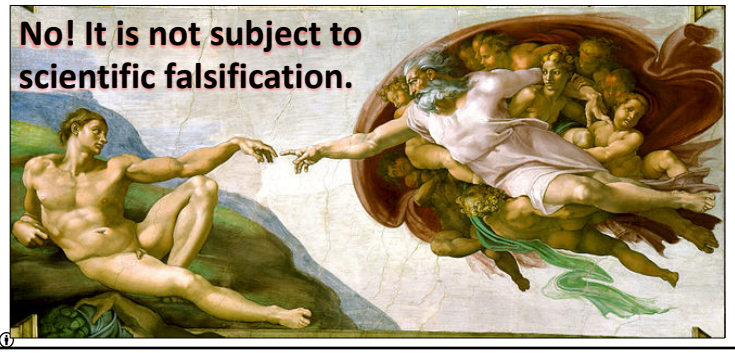
These decisions are mutually exclusive

10

What an Hypothesis is not

11

Is Michelangelo's painting of God creating Adam an hypothesis?



12

A Hypothesis is not a Theory

13

What is an Hypothesis?

Preliminary proposition that provides an explanation of a phenomenon

Can be tested, refuted, falsified

14

What is a Theory?

Unified explanation of phenomena

A theory was initially only a hypothesis

Withstood repeated attempts at falsification

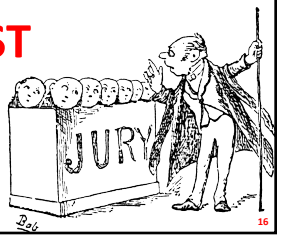
Yet, theories can also be falsified

Paradigm Shift:
Scientific revolution occurs when a theory is falsified

15

Non-Mathematical NHST

Criminal Jury Trial



16

Trials & NHST have a shared premise

Defendant is assumed not guilty until the prosecution can convince a jury of the defendant's guilt beyond a reasonable doubt

Null Hypothesis is presumed an acceptable explanation until the data proves otherwise

17

Null Hypothesis, H_0

With parametric techniques: H_0 always pertains to the population

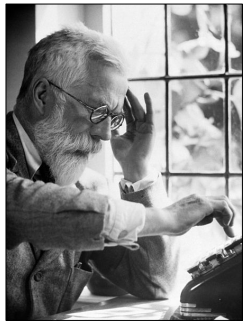
Sample statistic ($\bar{X}$, p, s^2, or r) = population parameter (μ, π, σ^2, or ρ)

With a nonparametric technique like Chi-square: Observed Frequencies = Expected Frequencies

Implies "no difference" or "no effect

18

H_0 never proven, but may be disproven



Ronald A. Fisher
(1890 – 1962)

"…the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis."*

*Ronald Aylmer Fisher, *The Design of Experiments*, (Edinburgh, UK: Oliver and Boyd, 1935), p. 19.

19

Hypotheses & Skepticism

The truth of any hypothesis is always subject to doubt

When the H_0 is not rejected, we never say it is proven

We say, "we failed to reject it"

When the H_0 is rejected, we do not consider the H_1 true

20

Alternate Hypothesis, H_1

This is the proposition the prosecutor hopes to prove: Defendant is guilty

Falsification of the H_0

Sample statistic ($\bar{X}$, p, s^2, or r) ≠ population parameter (μ, π, σ^2, or ρ)

Observed Frequencies do not fit the Expected Frequencies

21

H_0 & H_1 are…

Mutually Exclusive

Collectively Exhaustive

Either/Or proposition

22

Significance Level, α

Jury trial verdicts are based on the "reasonable doubt" standard

NHST decisions are based on the significance level, α

α is the inverse of the confidence level (1 – CL = α)

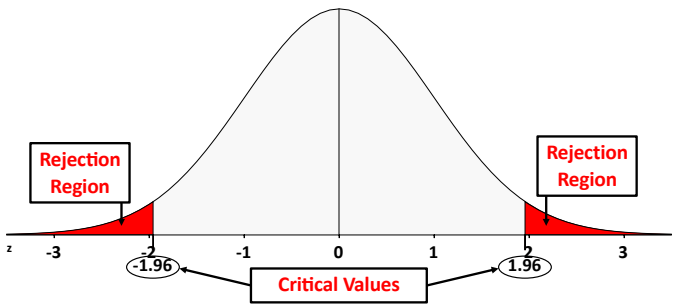
α delineates the rejection region from the region where the H_0 is not rejected

α is set at the start of the NHST

Neyman-Pearson & Fisher used a 5% α most frequently

23

Rejection regions & critical values



24

Two possible outcomes, two possible errors

Two Outcomes:

1) Reject H_0, 2) Fail to reject the H_0

Two Possible Errors:

1) False Positive (Type I or α)
2) False Negative (Type II or β)

25

Jury Trial: Correct Decision #1

Innocent man declared not guilty

H_0 is not rejected

Any difference between the statistic and parameter is merely sampling error

26

Jury Trial: Correct Decision #2

Guilty man is convicted

H_0 is rejected

Difference between the statistic and parameter is statistically significant (not sampling error)

27

Type I Error: False Positive

Innocent man is convicted

H_0 is wrongfully rejected

α sets the acceptable risk of a Type I error

28

Type II Error: False Negative

Guilty man is acquitted

Wrongfully failing to reject the H_0

29

2 x 2 NHST Outcomes Matrix

	Accept Null Hypothesis	Reject Null Hypothesis
Correct Decision	Jury correctly acquits the defendant. The analyst correctly fails to reject the null hypothesis.	Jury correctly convicts the defendant. The analyst correctly rejects the null hypothesis.
Incorrect Decision	Type I Error, α error, or false positive. An innocent man is convicted. The analyst wrongly rejects the null hypothesis.	Type II Error, β error, or false negative. An guilty man is acquitted. The analyst wrongly fails to reject the null hypothesis.

30

A philosopher on Type I & Type II Errors

Søren Kierkegaard
(1813 – 1855)

Type I Error
Reject "True" H_0

*"…one can be deceived in believing what is untrue, but on the other hand, one is also deceived in not believing what is true."

Type II Error
Retain "False" H_0

*Kierkegaard, Søren 1962. *Works of Love: Some Christian Reflections in the Form of Discourse*. Translated by Howard V. and Enda H. Hong. New York, NY, Harper Torchbooks. p. 23.

31

What affects the P(α) & P(β)?

β ↑ α ↓ β goes up, α goes down

σ^2 ↑ β ↑ σ^2 goes up, β goes up

n ↑ β ↓ n goes up, β goes down

32

Calculating Type I & Type II Errors

The allowable risk of a Type I error is set when the significance level is chosen

The probability of a Type II error is a calculation that will be explored

Goal for many analysts: P(α) ≤ 0.05 and the P(β) ≤ 0.20

(Type I Errors considered more serious than Type II Errors)

33

Statistical Power: (1 - β)

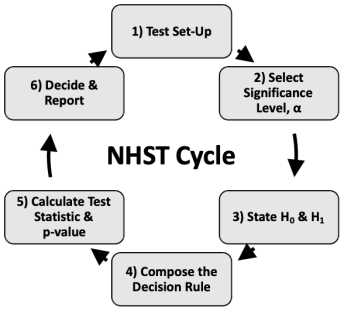
Power of a test to find a significant difference between the parameter and the statistic when one exists

Minimum level of statistical power is 0.80 [P(β) = 0.20]

Low powered tests are unreliable: Miss important effect and confuse sampling error for an effect

34

6 Steps of the NHST Cycle



35

Step 1: Test Set-Up

- Determine the research method
- Determine the operational definitions (how variables will be measured)
- Define the data collection process: Surveys, experiments, etc.
- Decide on the sampling methods
- Determine sample size
- Collect data

36

What's a test statistic?

Two meanings

- The formula for the test
- The result of NHST formula

37

Test statistics: 4 probability distributions

4 Probability Distribution

- Normal (z)
- Student-t
- F
- Chi-Square (χ^2)

38

Sample size is a big issue

- Sample size should be large enough to ensure sufficient statistical power
- Power should be estimated *a priori* (before the test)
- Power can be measured *post hoc* (after the test)

39

Significance Levels, α

- Acceptable risk of committing a Type I Error
- Chosen before data is analyzed
- Typically set at 0.05, but sometimes 0.01 or 0.10
- The lower the α, the harder to reject H_0
- The lower the α, the less statistical power
- Reject H_0 when the probability of obtaining a test statistic is ≤ to α

40

α & Critical Value(s)

The critical value or values of the test statistic that mark the boundary between the region where the H_0 is rejected or not rejected

Critical values are based on the α and the test statistic

41

Interpreting Statistical Significance

When the probability of the test statistic is ≤ α, the results are "statistically significant"

Reject the H_0

Results unlikely due to random sampling error

42

Statistical Significance ≠ Practical Significance

- Practical significance refers to the magnitude of the effect
- Having statistical significance does not mean results have any real application (practical significance)
- With a large enough n, all results have statistical significance, but may lack practical significance
- Practical significance without statistical significance is possible

43

Null Hypothesis (H_0)

- Hypothesis that researchers seek to "nullify"
- No difference between the statistic and parameter
- No effect
- Takes an equal sign: =, ≤, or ≥

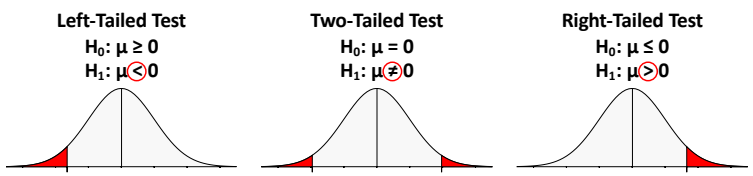
44

Alternate Hypothesis (H_1)

- Opposite of the H_0
- Something is happening:
  Statistic ≠ Parameter or there is an effect
- Takes a not equal sign: ≠, <, or >
- Does not imply practical significance

45

z and t tests are directional



Note the direction of the sign in the H_1

46

Decision Rules

States the criterion for rejecting H_0	Based on critical value(s)

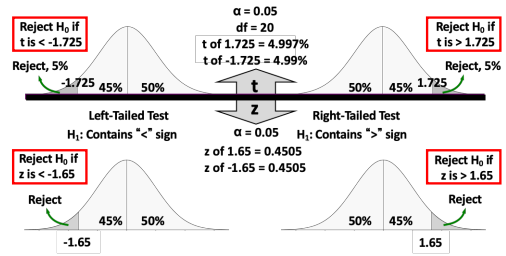
47

Decision Rule Syntax

Reject the Null Hypothesis if (z, t, F, χ^2) is (<, >, or < and >) the critical value(s)

48

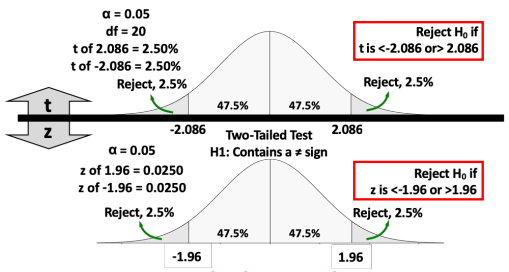
One-Tailed Tests



Note the decision rules

49

Two-Tailed Tests



Note the decision rules

50

Critical Values for z-tests

α	Left-Tailed Test	Two-Tailed Test	Right-Tailed Test
0.01	-2.33 (-2.326)	-2.58 & 2.58 (-2.576 (2.576)	2.33 (2.326)
0.05	-1.65 (-1.645)	-1.96 & 1.96 (-1.960 & 1.960)	1.65 (1.645)
0.10	-1.28 (-1.283)	-1.65 & 1.65 (-1.645 & 1.645)	1.28 (1.283)

Values within the parentheses are used by Excel

51

Calculating the Test Statistic

- Formula varies with the type of test
- Typically a fraction
- Numerator: Sampling error
- Denominator: Standard Error of the Mean or Proportion

52

Making a Decision

- Significance Level
- Value of Test Statistic
  - Decision Rule, or
  - p-Value

53

What are p-Values?

- Slippery concept 
- Measures compatibility of the data with the H_0
- p-value: Probability of getting a test statistics that is as extreme or more extreme than the one you just calculated

54

Interpreting p-Values

Fail to reject the H_0 when the p-value is > than α

α or P(Type 1 Error) = the probability of wrongfully rejecting a H_0

Reject the H_0 when the p-value is ≤ than α

55

p-Values & the decision to reject the H_0

Possible scenarios: $\alpha = 0.05$ and the p-value is 0.05 or <0.001

H_0 is rejected in both scenarios

We would have more confidence that the results are statistically significant when $p < 0.001$ than $p = 0.05$

56

P-values are widely misunderstood

57

March 2016: American Statistical Association's statement on p-values

58

ASA on p-Values

"The contents of the ASA statement and the reasoning behind it are not new—statisticians and other scientists have been writing on the topic for decades. But this is the first time that the community of statisticians, as represented by the ASA Board of Directors, has issued a statement to address these issues."

-- Jessica Utts, President, ASA

Jessica Utts, "American Statistical Association Releases Statement on Statistical Significance and P-Values," March 7, 2016. https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf

59

6 Principles

#1: P-values can indicate how incompatible the [sample] data are with a specified statistical model [or the null hypothesis].

#2: P-values do not measure the probability that the studied [null] hypothesis is true, or the probability that the data were produced by random chance alone.

#3: Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold [significance level].

60

6 Principles (Continued)

#4: Proper inference requires full reporting and transparency.

#5: A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

#6: By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

61

What's Next?

Test of Means, μ	Tests of Proportions, π	Test of Variance, σ^2	Tests of Relationships
• z-Tests • t-Tests • ANOVA Tests	• z-Tests	• F-test for Equality σ^2	• Chi-Square • Correlation & Regression

62

63

63

CLEAR-SIGHTED STATISTICS

EDWARD VOLCHOK