Hypothesis testing and p-value pitfalls

Sandra Taylor, Ph.D.

August 14 & 28, 2019

UCDAVIS
CLINICAL AND TRANSLATIONAL
SCIENCE CENTER

UCDAVIS
MIND INSTITUTE

UCDAVIS
COMPREHENSIVE
CANCER CENTER

UCDAVIS
ENVIRONMENTAL HEALTH
SCIENCES CENTER

We are video recording this seminar so please hold questions until the end.

Thanks

Seminar Objectives

- Understand framework of traditional null hypothesis significance testing
- Be able to correctly interpret p-values
- Understand confidence intervals
- Appreciate multiple testing issues and know corrections

Cardiovascular Disease Dataset

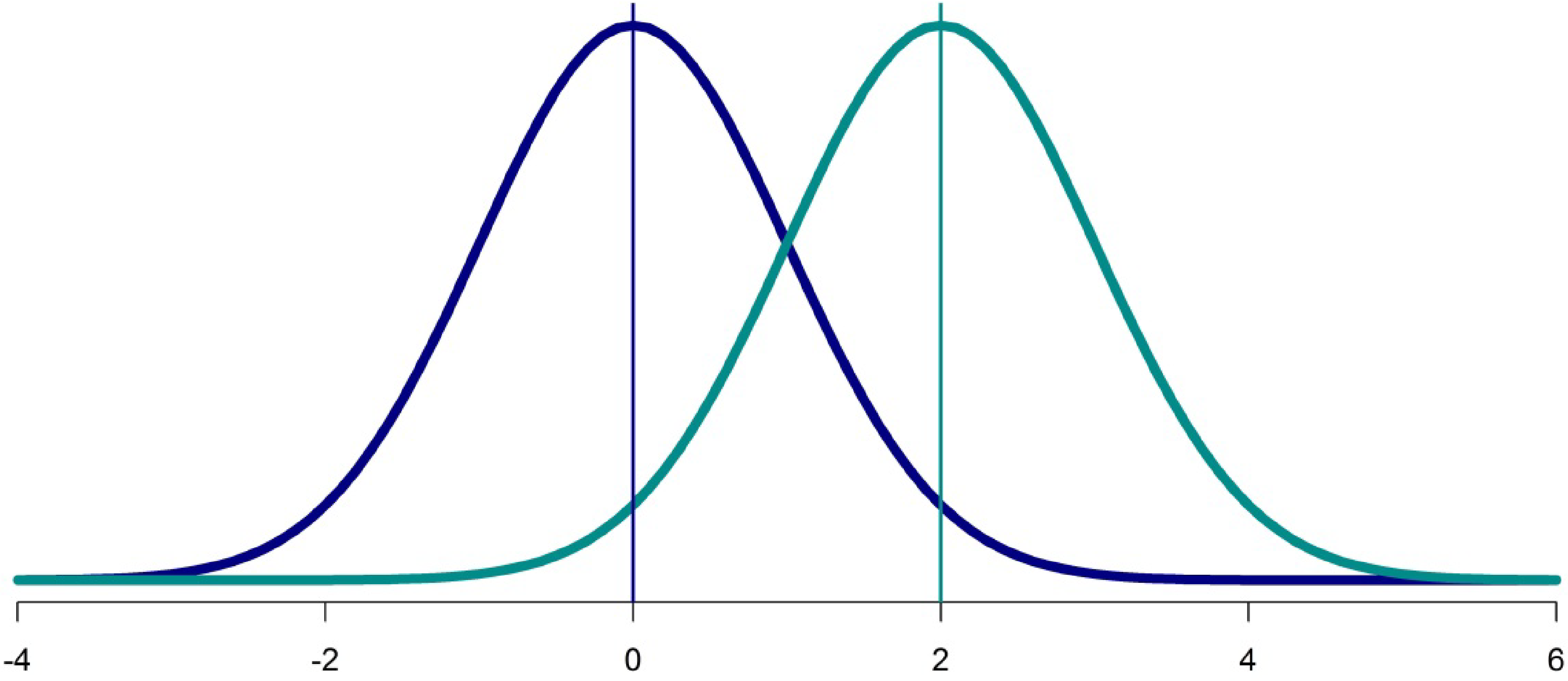
- 600 Subjects
- Presence/absence of coronary artery disease
- Demographics – age, sex, race, BMI
- Inflammatory biomarkers – CRP, LLPLA2, SAA, PTX3, FIBRIN, and HOMA

I will use this dataset to illustrate various points.

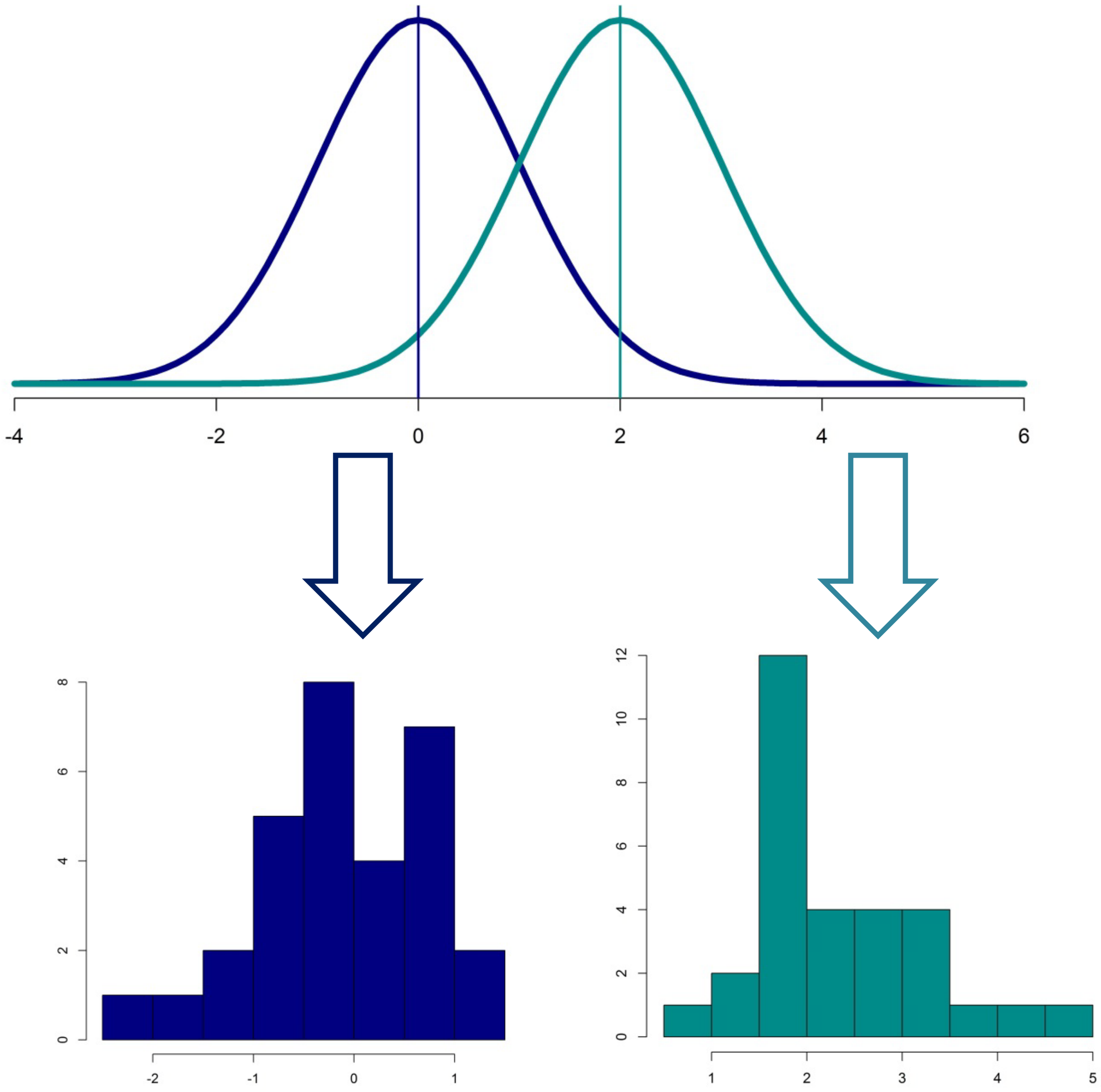
Primary and Secondary Aims

- **Primary Aim: Do HOMA levels differ between CAD(+) and CAD(-) subjects?**
  - Does the mean of HOMA levels differ between CAD(+) and CAD(-) subjects?
- **Secondary Aims: Do CRP, LLPLA2, SAA, PTX3, and FIBRIN levels differ between CAD(+) and CAD(-) subjects?**

The truth is out there.



If we had data from every person in our population we would know with certainty the difference in the group means.



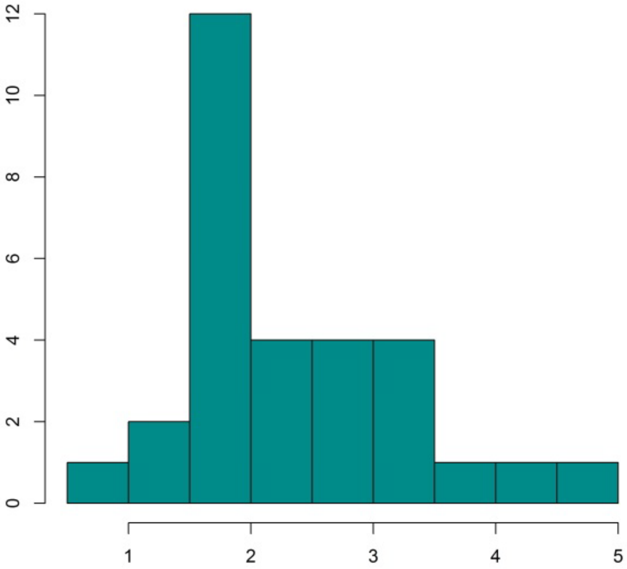
- Since we can’t observe every individual in a population, we collect a sample from the population.

- We seek to make inferences (i.e., make decision regarding our hypothesis) about the entire population based on the sample.

Sampling yields variability

Between Subject Variability

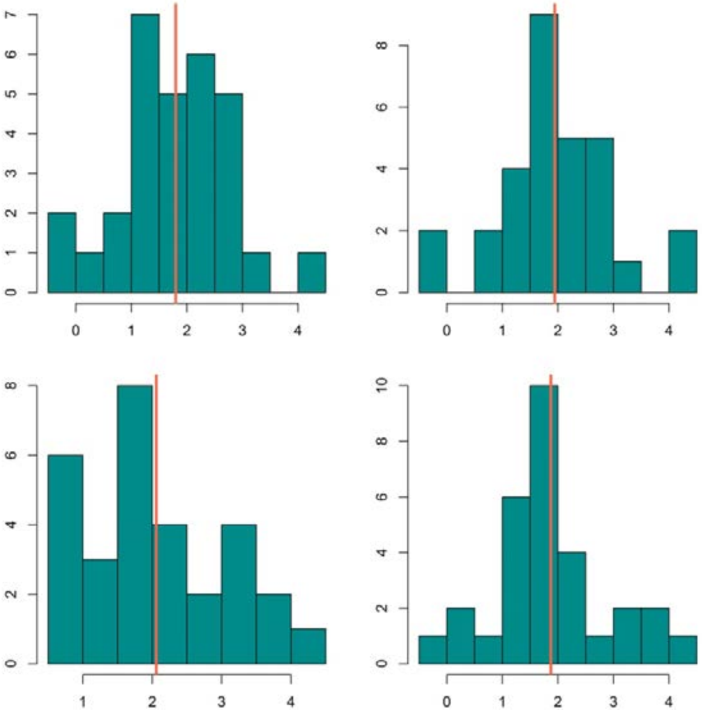
- Values differ between subjects



- Standard deviation

Between Sample Variability

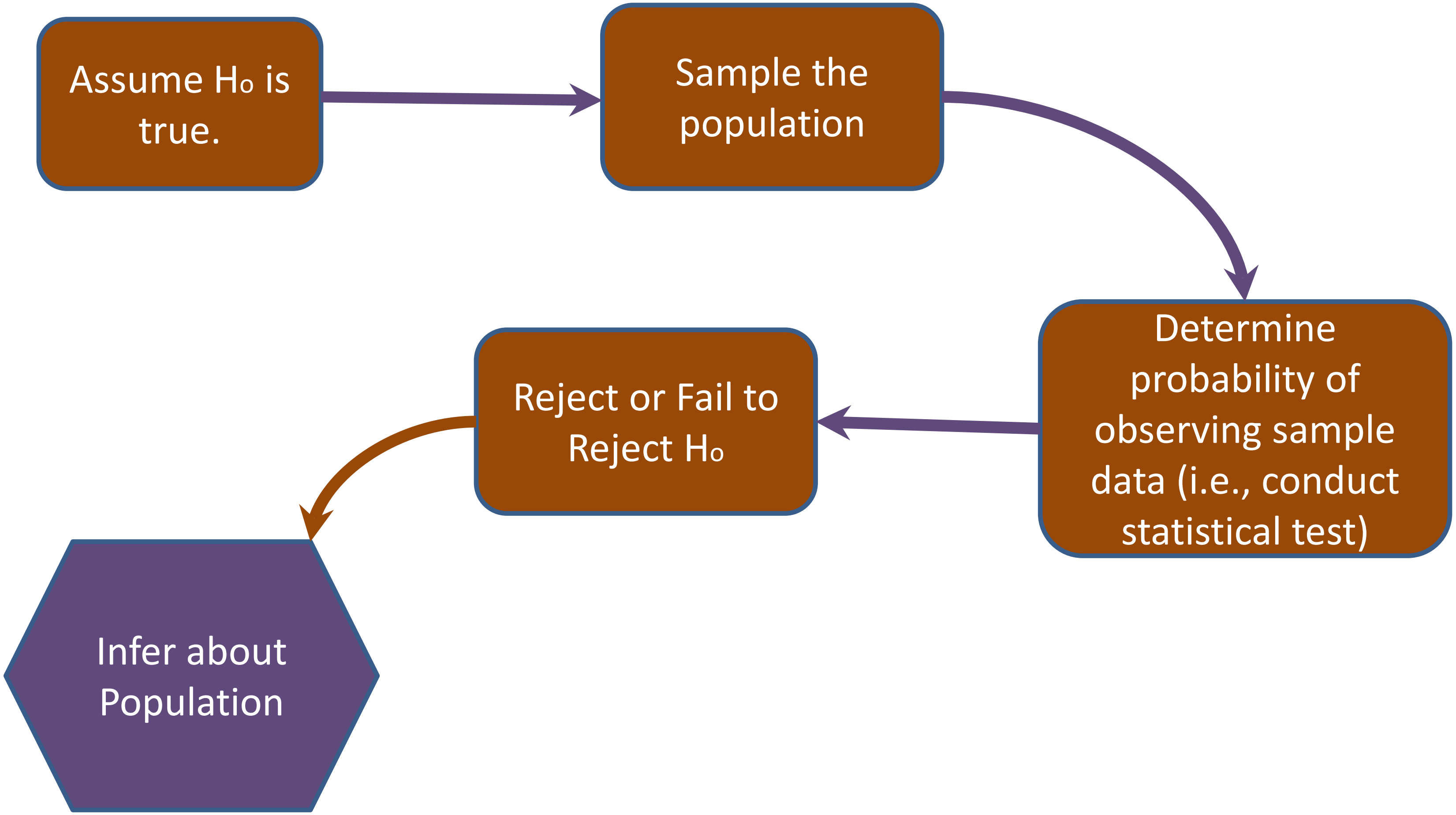
- Estimates differ between studies



- Standard error

Illustration of between study variability

How do we go from a sample to a decision? – Statistics!



Assume H_o is true.

Sample the population

Determine probability of observing sample data (i.e., conduct statistical test)

Reject or Fail to Reject H_o

Infer about Population

Null Hypothesis Significance Testing Framework

- **In null hypothesis significance testing, we posit a null hypothesis**
  - H_o: Mean CAD(+) = Mean CAD(-)
- **We seek to reject the null hypothesis in favor of an alternative hypothesis.**
  - H_a: Mean CAD(+) ≠ Mean CAD(-)
- **Notice the simplicity of H_a**
  - It's just that they aren't equal. No info on magnitude

Hypothesis Testing: Ideas on Trial

Courtroom

- **Presume innocent**
- **Present and evaluate evidence**
- **Jury verdict**
  - Guilty – ‘beyond a reasonable doubt’ standard avoids incorrect conviction
  - Acquittal – not proof of innocent
- **Incorrect guilty verdict worse than incorrect acquittal**

Hypothesis Testing

- **Assume null hypothesis is true**
- **Gather and evaluate evidence**
- **Statistical test result**
  - Reject H_0 – significance level (α) controls incorrect rejection
  - Fail to Reject H_0 – not unlikely to observe data
  - Does not prove H_0 is true
- **False positive worse than false negative**

Absence of evidence is NOT evidence of absence!

Courtroom

Conviction: Beyond a reasonable doubt

Acquittal: Reasonable doubt – evidence insufficient

Hypothesis Testing

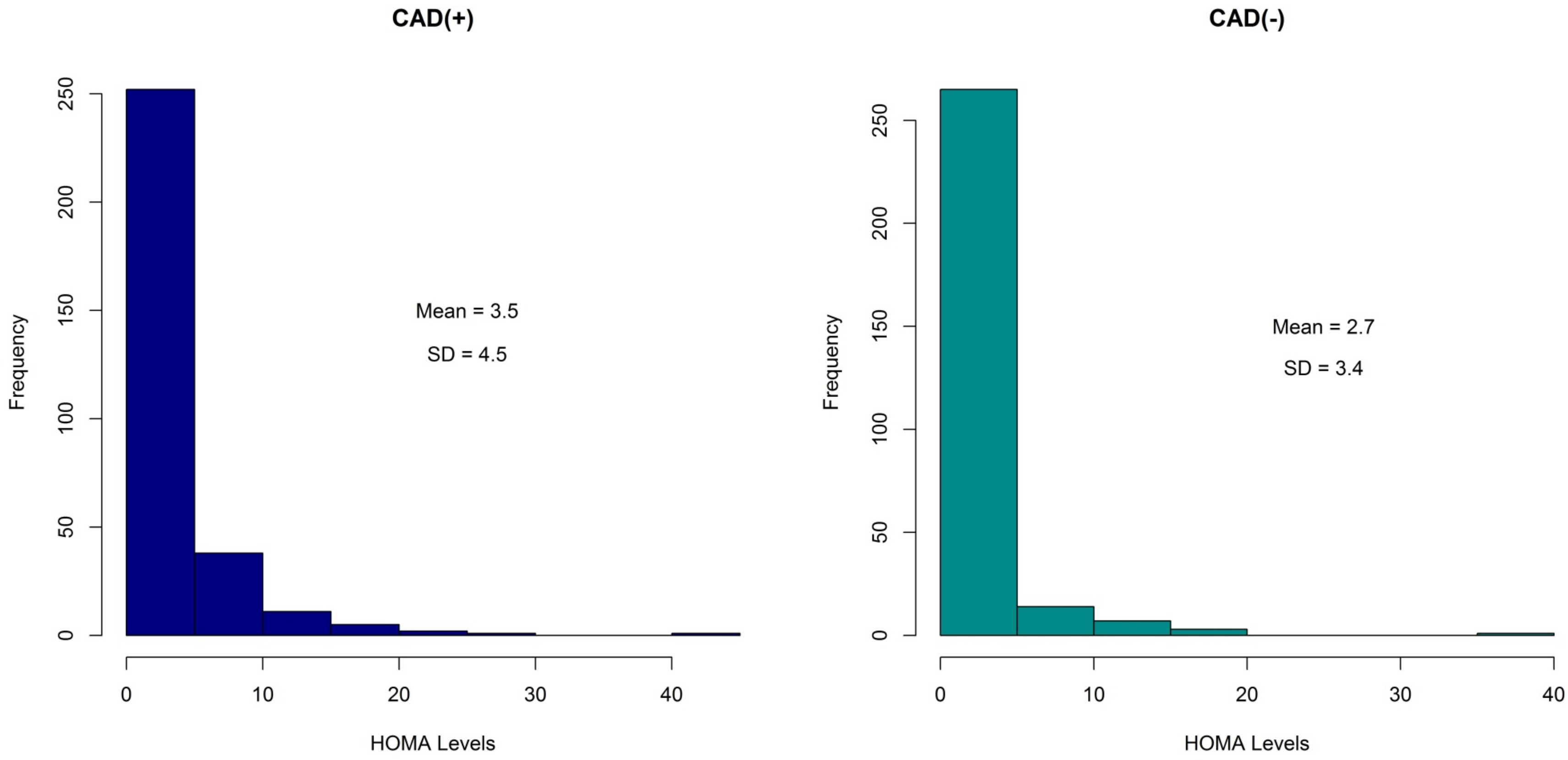
Reject H_o: Probability of observing data if null hypothesis is true is unlikely

Fail to Reject H_o: Probability of observing data if null hypothesis is true is not unlikely

Hypothesis Testing: Ideas on Trial

	H_o False (Defendant is Guilty)	H_o True (Defendant is Innocent)
Reject H_o (Guilty Verdict)	Correct decision	Type I error (α)
Fail to Reject H_o (Acquittal)	Type II error (β)	Correct decision

Return to CAD Example



Does HOMA differ between CAD(+) and CAD(-) Groups?

CAD(+)
mean = 0.84, sd = 0.83, n = 310

CAD(-)
mean = 0.67, sd = 0.73, n = 290

- Define the Null (H_o) and Alternative (Ha) Hypotheses

H_o: Mean HOMA levels do not differ between CAD(+) and CAD(-)

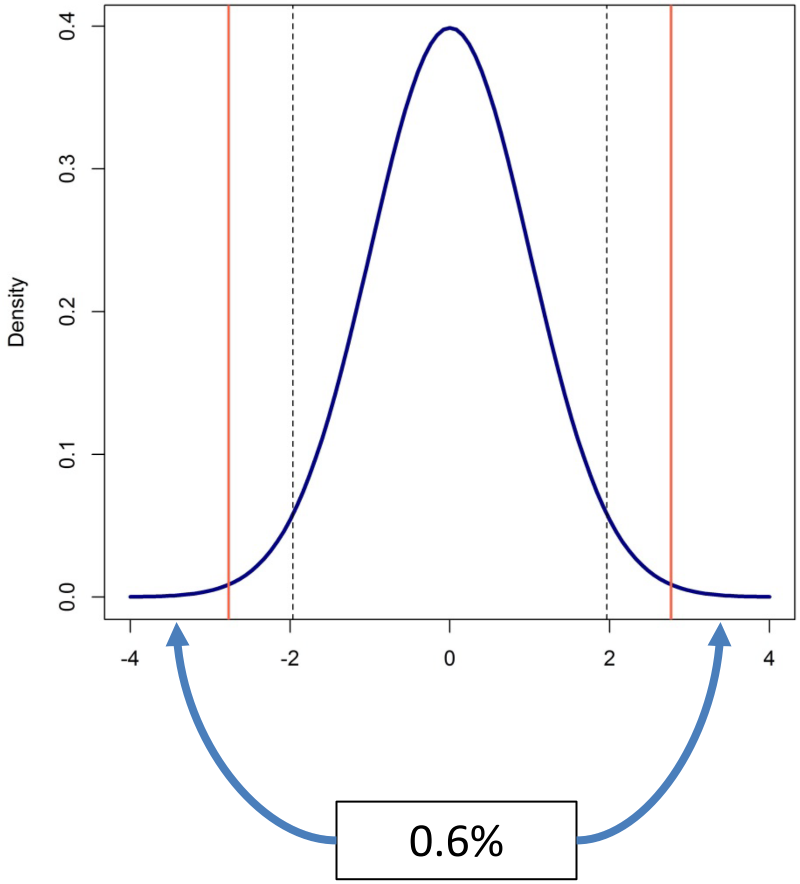
Ha: Mean HOMA levels differ between CAD(+) and CAD(-)

- Calculate test statistic
  - *t = 2.77*

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{{s_x}^2}{n_x} + \frac{{s_y}^2}{n_y}}}$$

- Calculate the probability of observing a *t* ≥ ± 2.77 *if the null hypothesis was true!*
- p-value = 0.006

What exactly are p-values?



0.6%

- **Probability that you would observe a test statistic at least extreme as you did *if the null hypothesis is true***
  - We know the distributions test statistics under H_o which allows us to calculate p-values
- **P = 0.006 – small probability so reject null hypothesis**
- **Did not *prove* alternative hypothesis**

What's so special about 0.05?

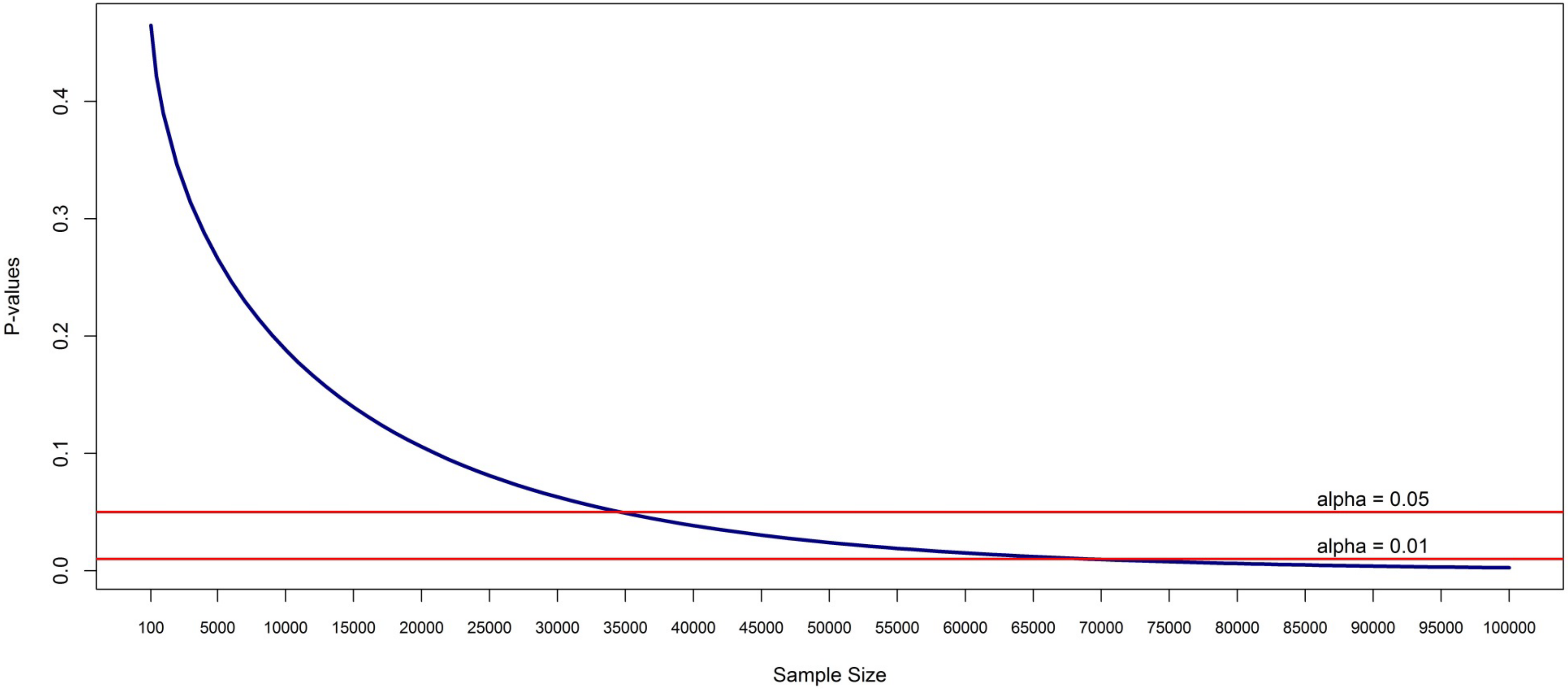
- Origin attributed to Ronald Fisher (1890-1962)
- English statistical evolutionary biologist
- Authored *Statistical Methods for Research Workers*
  - Very influential text
  - Provided probabilities between coarse bounds rather than very detailed tables – these were widely copied

"The value for which P=0.05 or 1 in 20; it is convenient to take this point as a limit in judging whether a deviation ought to be considered significant."

What if we had a different sample?

Statistical vs. Clinical Significance

- Statistically significant is not necessarily clinically significant
- Not statistically significant is not necessarily not clinically significant



Point estimates and confidence intervals more informative

- P-values help in decision-making about the null but provide no additional useful information
- Point estimates – size and direction of differences/relationships
- Confidence intervals – precision of estimates

What are confidence intervals and what do they tell us?

- **Define a range that includes the true value with a high degree of confidence, typically 95%.**
- **The confidence interval is NOT the probability that the true value is within the confidence limits.**
  - The true value is either in the limits or not with probability 1 or 0.
- **Repeated sampling and construction of confidence limits will encompass the true value 95% of the time**

Illustration of confidence intervals

Type II Errors and Power

- Significance level (α) limits type I error
  - Set fairly low to minimize false positives (e.g., wrongly convicting an innocent person)
- Type II errors (β) are false negatives – failing to reject the null hypothesis when it is false
- Power is probability of rejecting Ho when it is false
- Power = 1 - β

What determines the power of a test?

- **Size of the effect, e.g., difference between groups**
  - Larger effect → more power
- **Variability of the data**
  - Greater variability → less power
- **Sample size**
  - Larger sample → more power
- **Significance level (α)**
  - Smaller significance level → less power

How does sample size affect power?

- Assumes difference in means of 0.6 with SD = 1. So the two groups truly differ.

Sample Size (Per group)	Number of Rejections (Power)
10	18.0%
30	60.0%
50	86.0%
100	99.0%

- If you only have 10 samples per group, you will reject the null hypothesis about 18% of the time if the true difference in 0.6.

Hypothesis Testing: Summary

- Significance level controls type I error (false positives)
- Power controls type II error (false negatives)
- P-values aid in decision making about H_0
- Point estimates and confidence intervals are more informative than p-values
- Keep in mind between sample/study variation
- Keep in mind the sample size

Multiple Hypothesis Testing

- What is it?
- What does it mean to me?
- What do I do about it?

What is Multiple Testing?

- **Conducting many hypothesis tests simultaneously**
- **Examples:**
  - Comparing heart rate, respiratory rate, blood pressure, SOFA scores, mean arterial pressure, and additional laboratory values
  - Comparing multiple patient outcomes, e.g., 28-day mortality, in-hospital mortality, LOS, ICU LOS, ventilator days, readmissions
  - Evaluating scores from a battery of behavioral assessments

What does it mean to me?

- **Type I error not controlled at 0.05**
  - Recall Type I error = probability of rejecting the null hypothesis when it is actually true
- **Prob(at least 1 significant result) =**

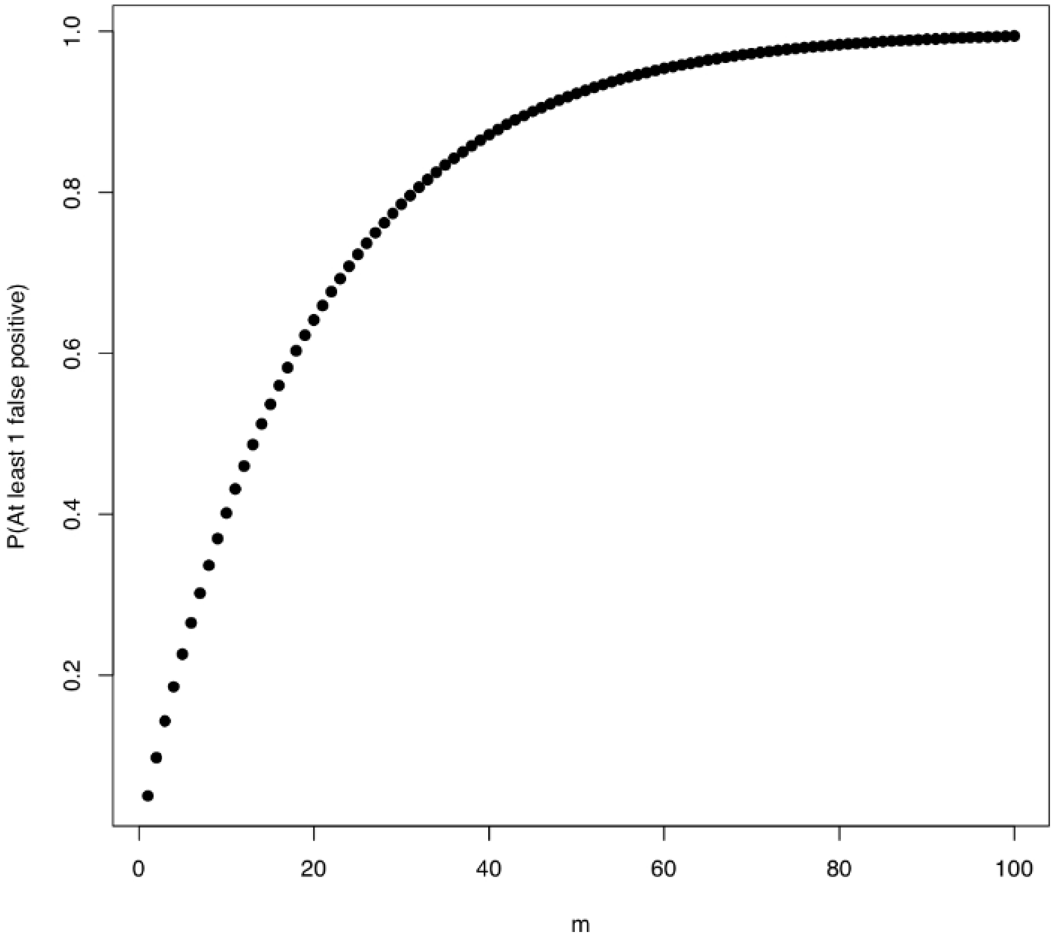
  1 – Prob(no significant result)^n =

  1- $(1-0.05)^n$

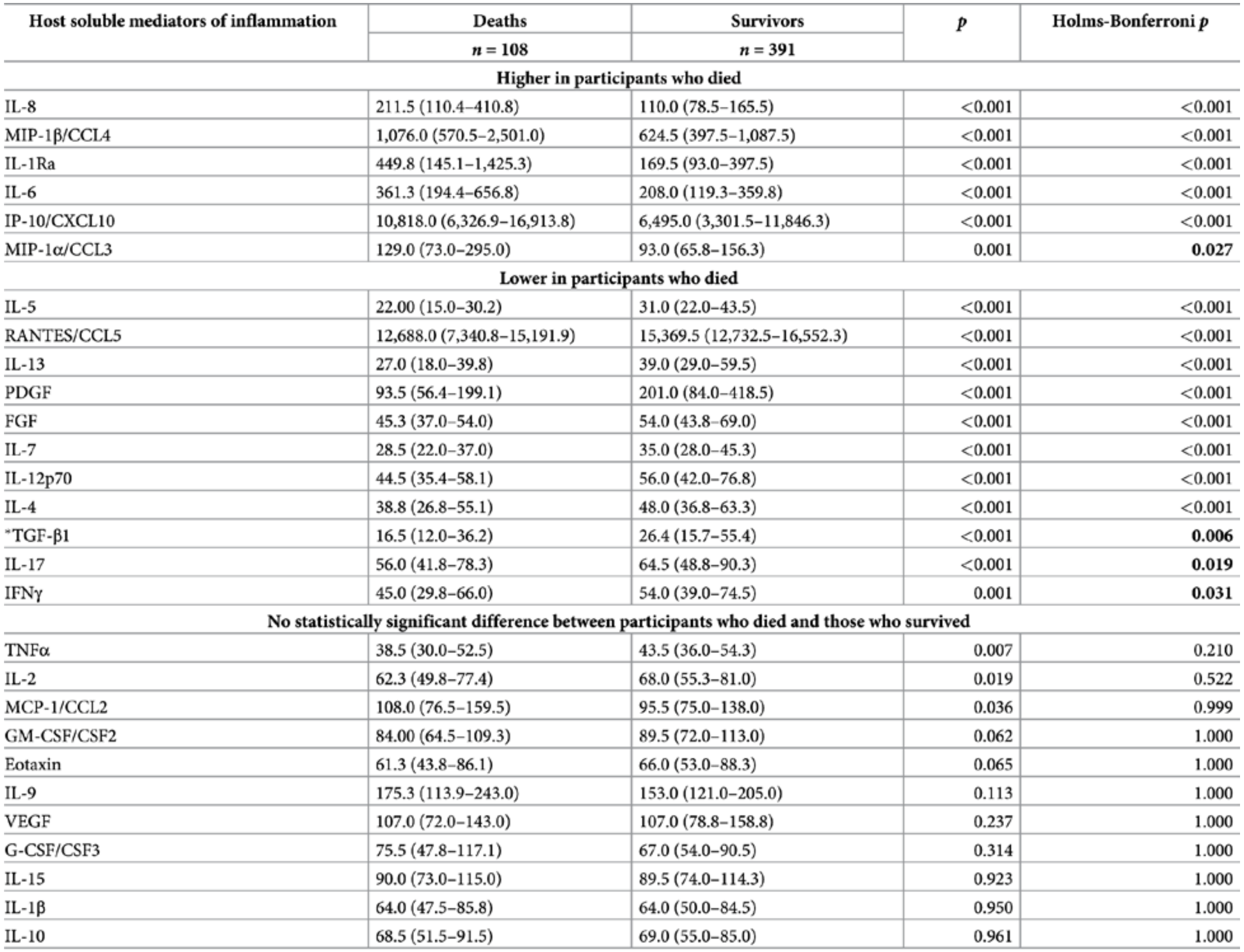
For 10 tests, Prob = $1-(1-0.05)^{10}$ = 0.40

40% probability of at least 1 false positive across 10 tests

Probability of at least 1 false positive



What do I do about it?



Host soluble mediators of inflammation	Deaths	Survivors	*p*	Holms-Bonferroni *p*
	n = 108	*n* = 391		
Higher in participants who died				
IL-8	211.5 (110.4–410.8)	110.0 (78.5–165.5)	<0.001	<0.001
MIP-1β/CCL4	1,076.0 (570.5–2,501.0)	624.5 (397.5–1,087.5)	<0.001	<0.001
IL-1Ra	449.8 (145.1–1,425.3)	169.5 (93.0–397.5)	<0.001	<0.001
IL-6	361.3 (194.4–656.8)	208.0 (119.3–359.8)	<0.001	<0.001
IP-10/CXCL10	10,818.0 (6,326.9–16,913.8)	6,495.0 (3,301.5–11,846.3)	<0.001	<0.001
MIP-1α/CCL3	129.0 (73.0–295.0)	93.0 (65.8–156.3)	0.001	**0.027**
Lower in participants who died				
IL-5	22.00 (15.0–30.2)	31.0 (22.0–43.5)	<0.001	<0.001
RANTES/CCL5	12,688.0 (7,340.8–15,191.9)	15,369.5 (12,732.5–16,552.3)	<0.001	<0.001
IL-13	27.0 (18.0–39.8)	39.0 (29.0–59.5)	<0.001	<0.001
PDGF	93.5 (56.4–199.1)	201.0 (84.0–418.5)	<0.001	<0.001
FGF	45.3 (37.0–54.0)	54.0 (43.8–69.0)	<0.001	<0.001
IL-7	28.5 (22.0–37.0)	35.0 (28.0–45.3)	<0.001	<0.001
IL-12p70	44.5 (35.4–58.1)	56.0 (42.0–76.8)	<0.001	<0.001
IL-4	38.8 (26.8–55.1)	48.0 (36.8–63.3)	<0.001	<0.001
*TGF-β1	16.5 (12.0–36.2)	26.4 (15.7–55.4)	<0.001	**0.006**
IL-17	56.0 (41.8–78.3)	64.5 (48.8–90.3)	<0.001	**0.019**
IFNγ	45.0 (29.8–66.0)	54.0 (39.0–74.5)	0.001	**0.031**
No statistically significant difference between participants who died and those who survived				
TNFα	38.5 (30.0–52.5)	43.5 (36.0–54.3)	0.007	0.210
IL-2	62.3 (49.8–77.4)	68.0 (55.3–81.0)	0.019	0.522
MCP-1/CCL2	108.0 (76.5–159.5)	95.5 (75.0–138.0)	0.036	0.999
GM-CSF/CSF2	84.00 (64.5–109.3)	89.5 (72.0–113.0)	0.062	1.000
Eotaxin	61.3 (43.8–86.1)	66.0 (53.0–88.3)	0.065	1.000
IL-9	175.3 (113.9–243.0)	153.0 (121.0–205.0)	0.113	1.000
VEGF	107.0 (72.0–143.0)	107.0 (78.8–158.8)	0.237	1.000
G-CSF/CSF3	75.5 (47.8–117.1)	67.0 (54.0–90.5)	0.314	1.000
IL-15	90.0 (73.0–115.0)	89.5 (74.0–114.3)	0.923	1.000
IL-1β	64.0 (47.5–85.8)	64.0 (50.0–84.5)	0.950	1.000
IL-10	68.5 (51.5–91.5)	69.0 (55.0–85.0)	0.961	1.000

Adjust p-values to control the overall error rate at desired level rather than controlling the error rate for just one hypothesis

Source: Schutz et al. 2019. PLoS Med 16(7): e1002840. https://doi.org/10.1371/journal.pmed.1002840

Multiple Testing Adjustment

- **Control Family-wise Type I Error**
  - Bonferroni adjustment
    - Use $\alpha' = {}^{\alpha}/_{n}$ where n = number of tests
    - Simple, applicable anywhere, most conservative
  - Sequential procedures
    - Less conservative than Bonferroni
    - Holm's step-down procedure
- **Control False Discovery Rate (FDR)**
  - Controls proportion of false positives out of all rejected hypotheses
  - Benjaminin & Hochburg procedure

Secondary Objectives:
CRP, LPPLA2, SAA, PTX3, FIBRIN

Biomarker	Raw P-value	Bonferroni	Holm’s	FDR
CRP	0.0557	0.279	0.194	0.093
LLPLA2	0.0855	0.428	0.194	0.107
SAA	0.0486	0.243	0.194	0.093
PTX3	0.8117	1.000	0.812	0.812
Fibrin	0.0361	0.180	0.181	0.093

Interpretation & Reporting

P-value Points to Remember

- **Probability of observing data more extreme than you did *if the null hypothesis is true***
- **NOT the probability that the null hypothesis *is* true**
- **Absence of evidence is NOT evidence of absence**
  - Particularly important for small studies
  - Non-significant P values do not distinguish between group differences that are truly negligible and group differences that are non-informative because of large standard errors.
- **P-values provide no information about the magnitude of differences.**

Reporting & Interpretation

Suppose p = 0.006

- We could state, “Mean HOMA levels were significantly higher in subjects with CAD (p = 0.006). Log transformed mean [95% CI] values were 0.84 [0.75, 0.93] and 0.67 [0.59, 0.72] for CAD(+) and CAD(-) groups respectively.”
- Also report sample sizes: n = 310 and 290, for CAD(+) and CAD(-)

Now suppose p = 0.32

- Would not want to say “CAD status had no effect on HOMA levels” or “HOMA levels did not differ by CAD status.”
- We could state, “Evidence was not sufficient to reject the null hypothesis of no difference in mean HOMA levels by CAD status (p = 0.32). Log transformed mean [95% CI] values were 0.84 [0.75, 0.92] and 0.79 [0.65, 0.85] for CAD(+) and CAD(-) groups respectively.”
- Again, report sample sizes.

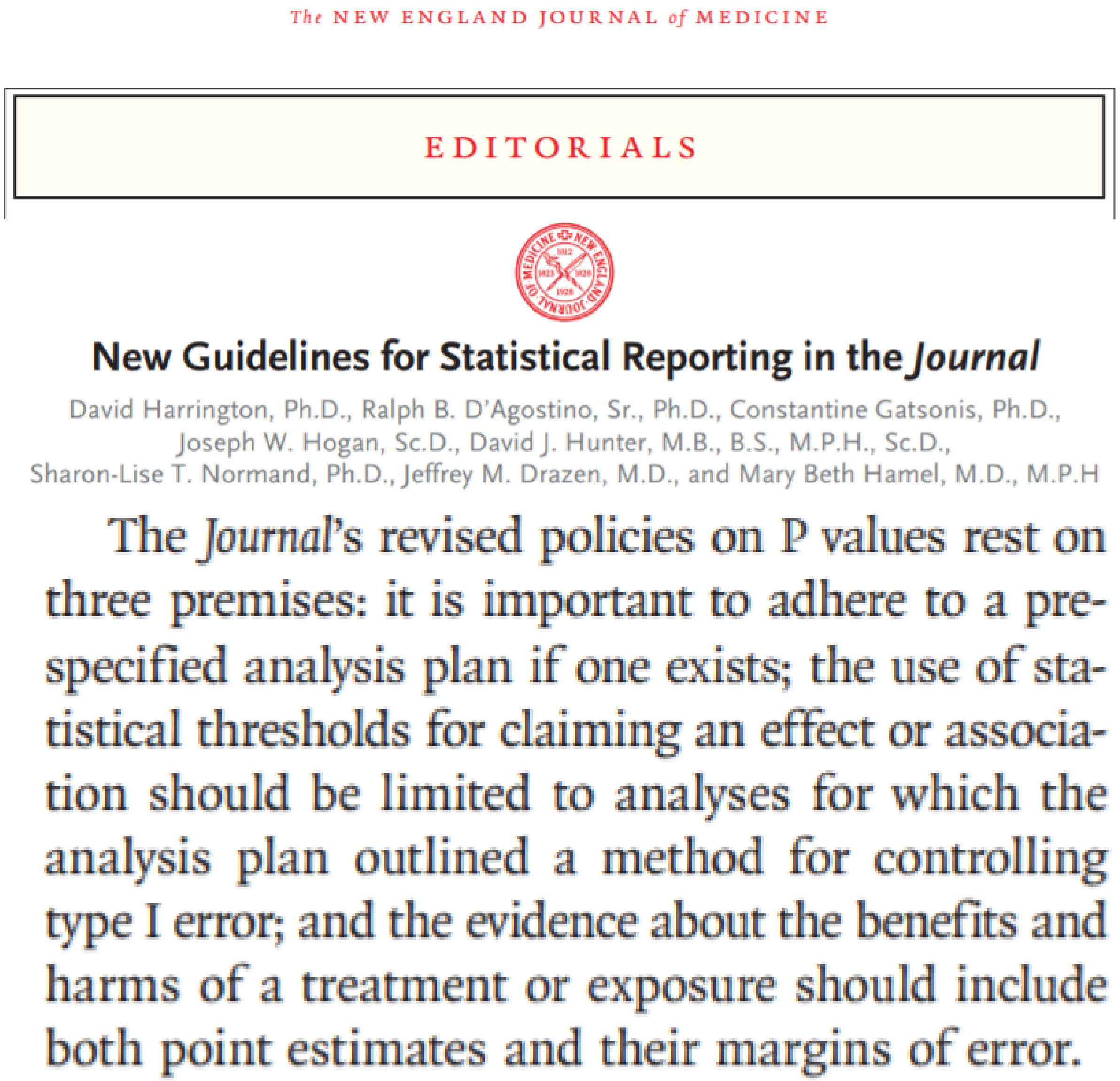
What if we see…

Scenario 1

- CAD(+): 0.84 [0.54, 1.14], n = 20
- CAD(-): 0.42 [0.12, 0.72], n = 18

Scenario 2

- CAD(+): 0.85 [0.83, 0.88], n = 2000
- CAD(-): 0.80 [0.78, 0.82], n = 1800



The NEW ENGLAND JOURNAL of MEDICINE

EDITORIALS

New Guidelines for Statistical Reporting in the *Journal*

David Harrington, Ph.D., Ralph B. D'Agostino, Sr., Ph.D., Constantine Gatsonis, Ph.D.,
Joseph W. Hogan, Sc.D., David J. Hunter, M.B., B.S., M.P.H., Sc.D.,
Sharon-Lise T. Normand, Ph.D., Jeffrey M. Drazen, M.D., and Mary Beth Hamel, M.D., M.P.H

The *Journal*'s revised policies on P values rest on three premises: it is important to adhere to a prespecified analysis plan if one exists; the use of statistical thresholds for claiming an effect or association should be limited to analyses for which the analysis plan outlined a method for controlling type I error; and the evidence about the benefits and harms of a treatment or exposure should include both point estimates and their margins of error.

NEJM 381:3 July 18, 2019

NEJM Statistical Reporting Guidelines

- Significance tests should be accompanied by confidence intervals for estimated effect sizes, measures of association, or other parameters of interest.
- P values adjusted for multiplicity should be reported when appropriate and labeled as such in the manuscript
- When appropriate, observational studies should use pre-specified accepted methods for controlling family-wise error rate or false discovery rate when multiple tests are conducted.

Help is Available

- **CTSC Biostatistics Office Hours**
  - Every Tuesday from 12 – 1:30 in Sacramento
  - Sign-up through the CTSC Biostatistics Website
- **EHS Biostatistics Office Hours**
  - Every Monday from 2-4 in Davis
- **Request Biostatistics Consultations**
  - CTSC - www.ucdmc.ucdavis.edu/ctsc/
  - MIND IDDRC - www.ucdmc.ucdavis.edu/mindinstitute/centers/iddrc/cores/bbrd.html
  - Cancer Center and EHS Center

Selected References

- Nuzzo. 2014. Statistical errors. *Nature* 506: 150
- Kim and Bang. 2016. Three common misuses of P values. *Dent Hypotheses* 7: 73
- Ioannidis 2005. Why most published research findings are false *PLoS Medicine* 2(8) e124
- Wasserstein and Lazar. 2016. The ASA’s statement on p-Values: Context, process, and purpose. *The American Statistician* 70(2): 129