Statistical Significance and *p*-Values: A Refresher on Statistical Methodology for the Device Clinical Researcher

Thomas John, Ph.D. (thjohn@syr.edu)

Society of Clinical Research Associates (SOCRA) 24th Annual Conference

Device Research Track, 9/19/2015 at10:50 AM

Abstract & Structure of the Talk

Abstract:
This talk will provide a refresher on statistical inference concepts and terminology with very limited technical/mathematical content. Dr. John will focus on statistical hypothesis testing and associated concepts such as p-values, level of significance, error types, and power. Emphasis will be given to the philosophy, logic, and history of hypothesis testing.

Structure:

- We start with a motivating example from a published research article and from the example context, significance testing terminology is introduced.
- Details are provided regarding where the p-value comes from and what it means.
- Some historical context of quantifying “beyond reasonable doubt” is explored.
- Certain cautions regarding statistical significance are discussed.
- We conclude with a brief discussion of type-I error, type-II error, and power of a test.

Motivating Example

Table 2 from

Lee, Janice S., et al. "Normal vision despite narrowing of the optic canal in fibrous dysplasia." *New England Journal of Medicine* 347.21 (2002): 1670-1676.

Available at:

fibrousdysplasia.org/pdfs/11Normal Vision Despite.pdf

TABLE 2. FINDINGS ON COMPUTED TOMOGRAPHIC EXAMINATION IN THE 38 PATIENTS WITH CRANIOFACIAL FIBROUS DYSPLASIA AND THEIR MATCHED CONTROLS.*

VARIABLE	PATIENTS WITH FIBROUS DYSPLASIA (N=38)	CONTROLS (N=38)	DIFFERENCE	P VALUE‡
No. of optic canals	67 (involved)	67 (not involved)		
Right	32	32		
Left	35	35		
Dimensions of optic canal				
Height — mm	3.7±0.9	4.2±0.6	0.48±1.2	0.001
Right	3.6±0.9	4.2±0.6	0.54±1.2	0.015
Left	3.8±0.9	4.2±0.6	0.42±1.2	0.04
Width — mm	3.3±0.7	3.6±0.5	0.29±0.9	0.007
Right	3.3±0.7	3.6±0.6	0.34±0.9	0.04
Left	3.3±0.7	3.6±0.5	0.26±0.8	0.08
Area — mm^2	9.8±3.7	11.9±2.8	2.09±4.8	<0.001
Right	9.6±3.8	12.0±2.9	2.37±4.8	0.009
Left	9.9±3.6	11.9±2.7	1.83±4.9	0.03
Extent of involvement — no.†				
Circumferential	49	NA		
Partial	18	NA		
Length of optic nerve — mm				
Right	58.4±10.2			
Left	57.7±9.1			
Difference between right and left	3.2±2.2			<0.001

*Plus–minus values are means ±SD. NA denotes not applicable.

†Involvement was considered circumferential if there was 100 percent encasement of the optic canal and partial if there was 25 to 75 percent encasement.

‡P values are for the comparison between patients and their matched controls.

Motivating Example (continued)

- We will focus on the left optic canal height.

Variable	Patients with Fibrous Dysplasia (N=38)	Controls (N=38)	Difference	P Value‡
No. of optic canals	67 (involved)	67 (not involved)		
Right	32	32		
Left	35	35		
Dimensions of optic canal				
Height — mm	3.7±0.9	4.2±0.6	0.48±1.2	0.001
Right	3.6±0.9	4.2±0.6	0.54±1.2	0.015
Left	3.8±0.9	4.2±0.6	0.42±1.2	0.04
Width — mm	3.3±0.7	3.6±0.5	0.29±0.9	0.007
Right	3.3±0.7	3.6±0.6	0.34±0.9	0.04

Statistical Hypotheses

- Research Question: Is the left optic canal height different between the CFD patients and the matched controls?
- Statistical Hypotheses:
  - Null Hypothesis: the average difference among the entire population is ZERO
  - Alternate (research) hypothesis: the population average difference is *Not* ZERO
- Null hypothesis represents the "status quo" (i.e., "nothing new" - what we will hold onto believing, unless evidence convinces us otherwise)
  - Unless the research shows evidence, we have no reason to think that the left optic canal height between CFD patients and matched controls would be different.
  - Analogous to "presumption of innocence"
- Research hypothesis represents "anything new?" (i.e., is the left optic canal height different between the two groups?)
  - This is what we are questioning in the research.

Presumption of Innocence...

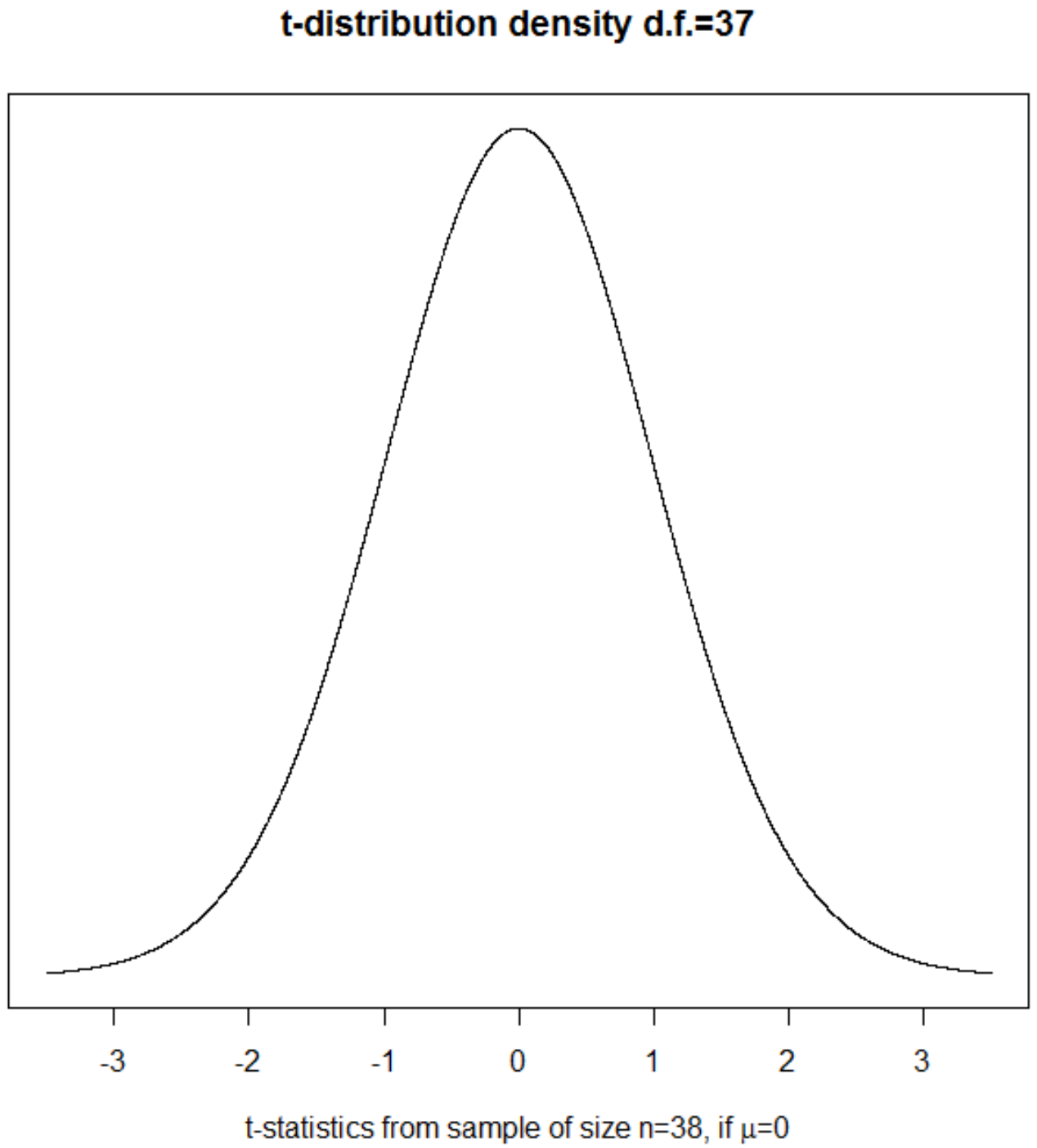
- We proceed similar to a criminal trial setting.
- We will presume innocence (null hypothesis is true - i.e., no difference in the population) unless we see evidence beyond reasonable doubt.

- If there is no difference *in the population* between CFD patients and matched controls, statistical theory tells us that the sample average difference from any sample of size n=38 from a normal distribution scaled "appropriately" will have a t-distribution with degrees of freedom n-1 = 38-1 = 37.

- Sample mean scaled "appropriately" is called the t-statistic:
  - t-statistic: $t = \sqrt{Sample\ Size}\,\frac{Sample\ Mean}{Sample\ Standard\ Deviation} = \sqrt{n}\,\frac{\bar{X}}{S}$

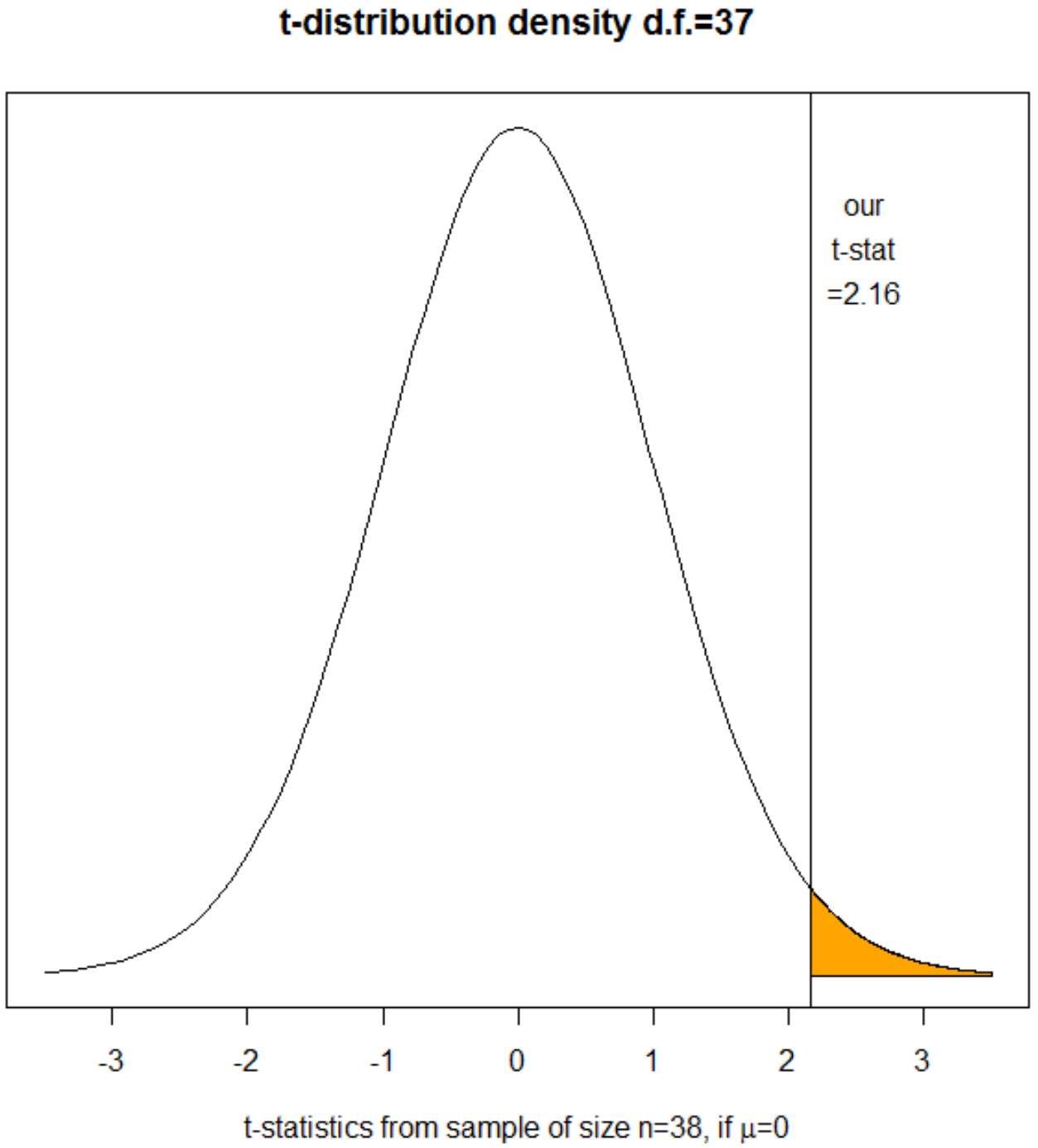
Sampling Distribution

Statistical theory tells us that if we considered *all* possible samples of size n=38, t-statistics computed from those samples would have this shape on the right (t-distribution with degrees of freedom 37), *if the population mean difference is 0.*



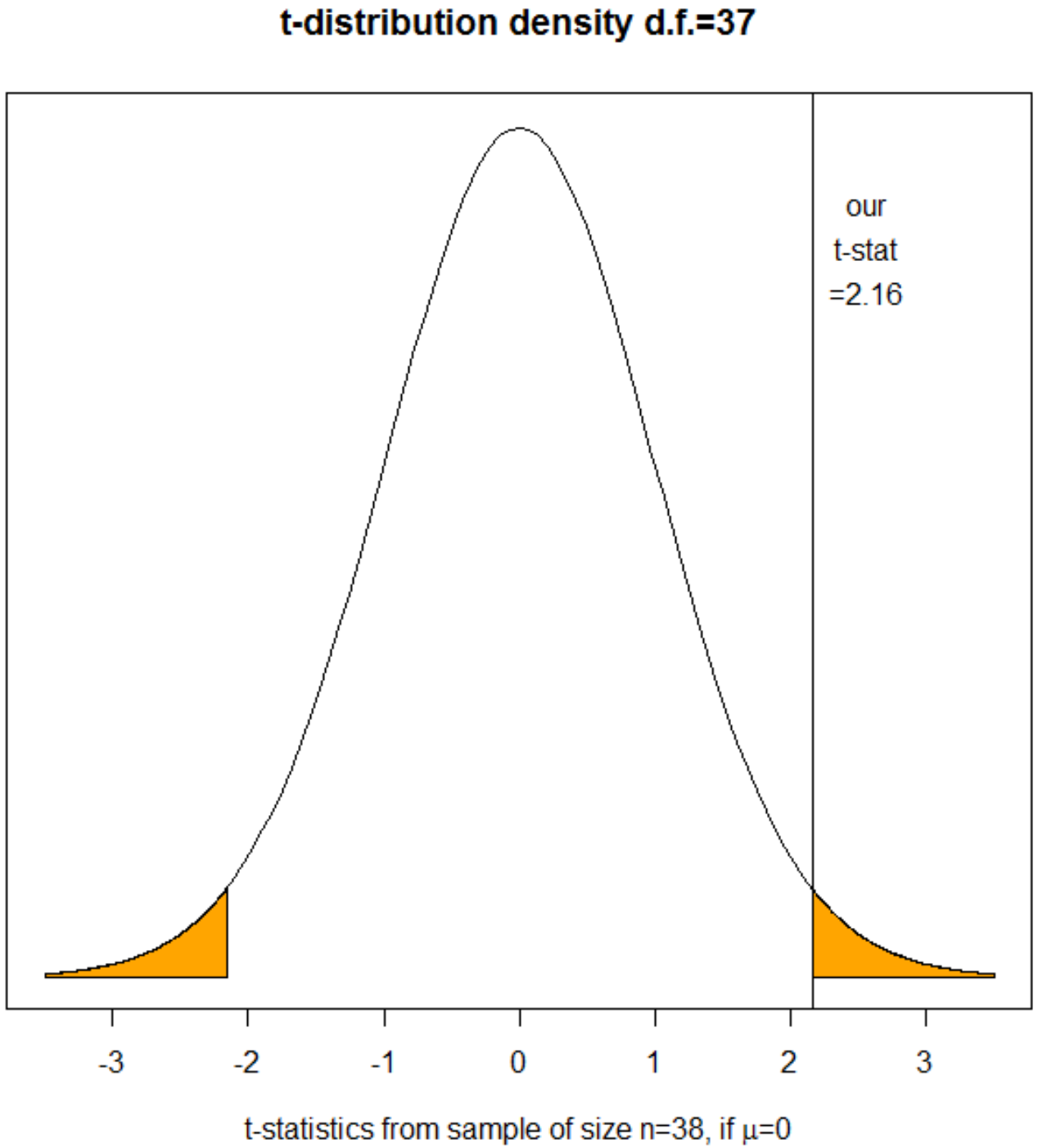
Where does the t-stat from *this study* sample fall?

- From the table out of the NEJM paper:
- Sample mean difference was 0.42 with ample standard deviation of difference 1.2.
- So *t*-stat = $\sqrt{38}\frac{0.42}{1.2}$ = 2.16.
- Tail area: how far away from 0 is the t-stat.



Sum of the two tails

- But our research hypothesis did not pre-suppose that the difference will be positive!
- The orange shaded area (on the previous slide) considers only the positive (right) end.
- So we should add the areas from both tails.



p-Value

- The sum of the two tail probabilities gives us what is called the p-Value for this scenario.
- For example in Excel, T.DIST.2T(2.16, 37) gives 0.03732 which matches with the paper's reported p-Value of 0.04.

- **What is *p*-Value?**: It is the probability of observing any test statistic that is at least as extreme as the one computed from the study sample, assuming the null hypothesis is true.
- That is, we can say "Small" p ⇒ Evidence against the null hypothesis

- **How small is "small"?**

Genesis 18:23-32

Abraham approached Him and said, “Will you sweep away both the righteous and the wicked? Suppose you find fifty righteous people living there in the city will you still sweep it away and not spare it for their sakes?”

...

And The Lord replied, “If I find fifty righteous people in Sodom, I will spare the entire city for their sake.”
Then Abraham spoke again.

...

Finally, Abraham said, “Lord, please don't be angry with me if I speak one more time. Suppose only ten are found there?”
And The Lord replied, “Then I will not destroy it for the sake of the ten.”

Underlines added for emphasis

Jurisprudence

"It is better and more satisfactory to acquit a thousand guilty persons than to put a single innocent one to death."

Rabbeinu Mosheh Ben Maimon ("Moses Maimonides"), 12th Century

"All presumptive evidence of felony should be admitted cautiously; for the law holds it better that ten guilty persons escape, than that one innocent party suffer."

William Blackstone, *Commentaries on the Laws of England*, circa 1760

Underlines added for emphasis

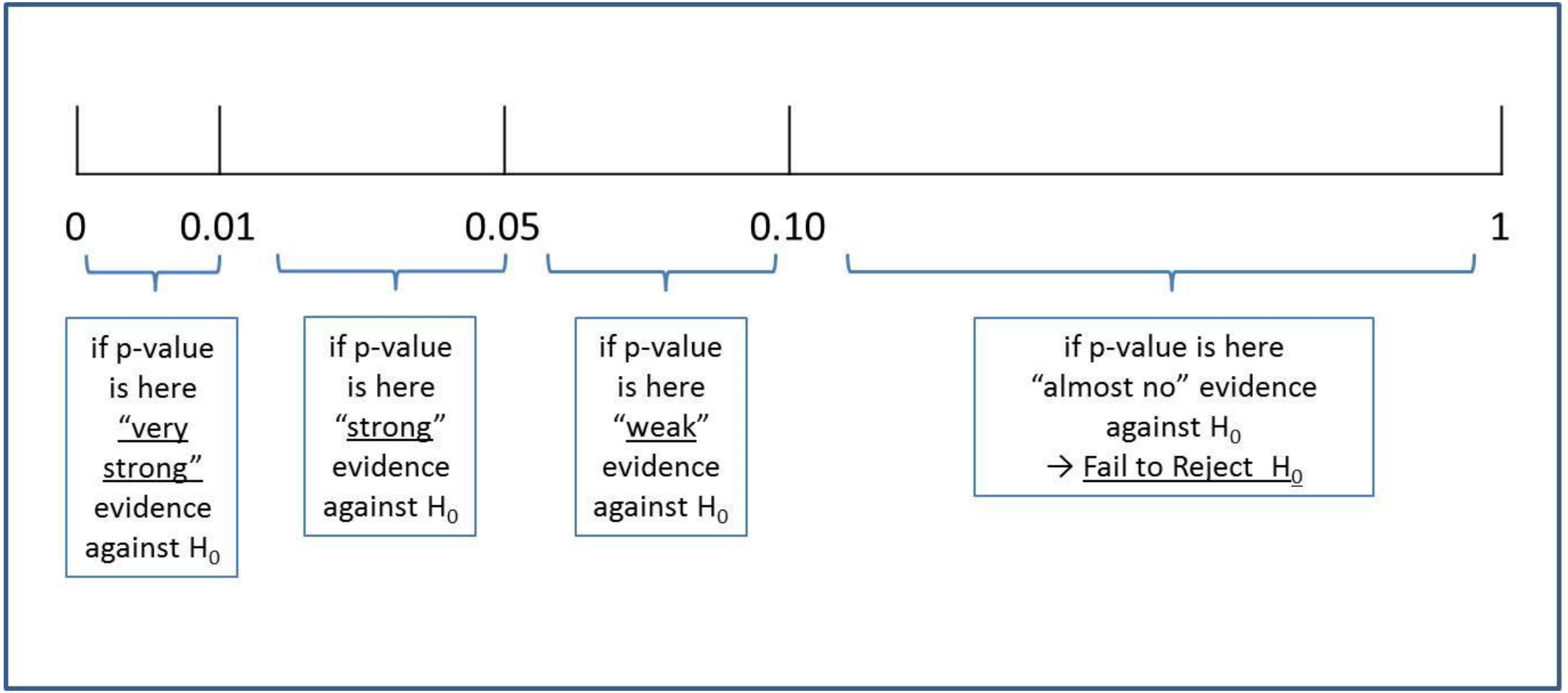
...to the Father of Modern Statistics

"If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 percent point), or one in a hundred (the 1 percent point). Personally, the writer prefers to set a low standard of significance at the 5 percent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance."

Sir Ronald Fisher (1926)

Underlines added for emphasis

p-Value and Significance Levels



Caution: It's Not All About The Significance

- Recent Supreme Court Decision on Statistical Significance:
  - Supreme Court of the United States, No.09-1156, *Matrixx Initiatives, Inc., et al. V. James Siracusano et al.*, March 22, 2011
  - Issue: *"a pharmaceutical company's failure to disclose reports of adverse events associated with a product if the reports do not disclose a statistically significant number of adverse events"*
  - Plaintiff's argument: *"adverse event reports that do not reveal a statistically significant increased risk of adverse events from product use are not material information."*
  - Supreme Court held: *"that the materiality of adverse event reports cannot be reduced to a bright-line rule"*

- "Are the effects of A and B different? They are always different --- for some decimal place."(John Tukey)

Type I and Type II Errors

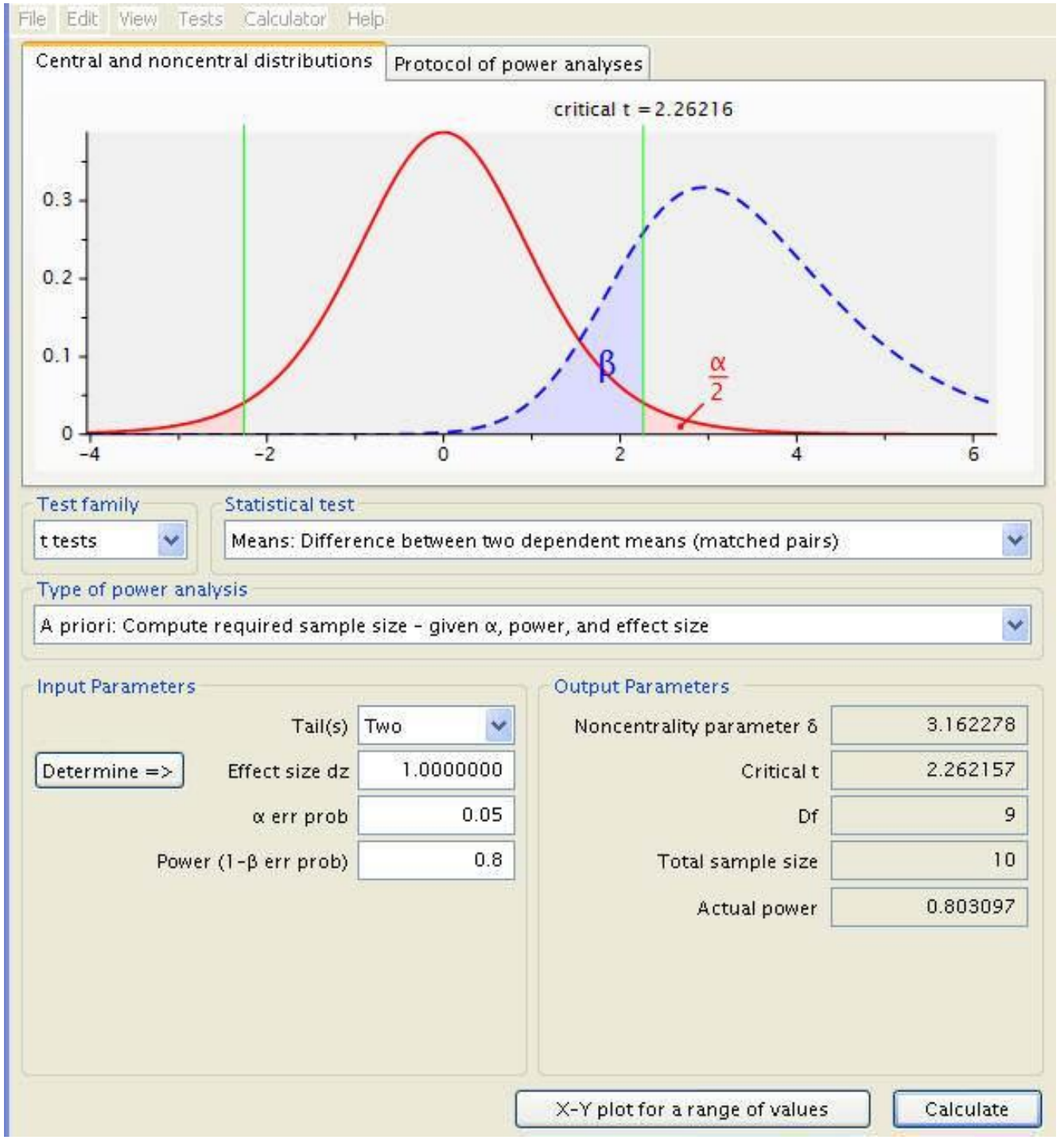
		Truth ("Unknown Reality") About the Population	
		Null Hypothesis is True	**Research Hypothesis is True**
Decision (Research Conclusion) Based on the Sample Data	**Reject Null Hypothesis** (Research Finding is Found Significant)	Type-I Error	Correct Decision
	Fail to Reject Null (Research Finding is Found Not Significant)	Correct Decision	Type-II Error

- Note the "Correct Decision" cell in the upper right corner:
  - Chances of making this correct decision is called the power of the test.
  - Power quantifies the ability of the research design to detect a "difference" away from the "status quo", *if* such a difference truly exists in the population

Power and Sample Size

Picture on the right comes from illustrations done by Stat Consulting Group at UCLA.

See illustrations at: http://www.ats.ucla.edu/stat/gpower/pairedsample.htm



Concluding Remarks

- Be aware of the underlying assumptions:
  - Proper sampling (unbiased/randomized/independent)
  - Normality assumption (or similar assumptions in other scenarios)
- Be conscious of the population and the sampling frame
- Be mindful of the limitations:
  - Statistical significance vs. Practical significance
  - Effect of the sample size
  - Multiple testing problems (e.g., with 5% significance level, 1 out of 20 tests could possibly be incorrect significant finding)
- Consider:
  - Emphasizing estimated difference and its confidence interval
  - Thinking of *p*-Value in the “continuum” instead of arbitrary cut-offs