What p values really mean (and why I should care)

Francis C. Dane, PhD

Jefferson College of Health Sciences at CarilionClinic

Session Objectives

- Understand the statistical decision process
- Appreciate the limitations of interpreting p values
- Value the use of statistics regarding effect size

Inferential Statistics

- Test of fit between data and model
- Usually, model = null hypothesis
  - There is no difference
  - There is no relationship
  - It's all measurement error and individual differences
- Usually, we hope to reject null hypothesis
  - But sometimes we want to accept it

Statistical Decision Making

- Unknowable State of the World
  - We (almost) never get to measure the entire population

State of the World	
Model is Correct	Model is Wrong

Statistical Decision Making

- Decision about Fit between Data and Model
  - We decide whether or not our data fit the chosen model

Decision about Fit	Accept Model (Data fit)
	Reject Model (Data do not fit)

Statistical Decision Making

- Four Possible Decision Outcomes
  - Two ways to be correct, and two to be wrong

		State of the World	
		Model is Correct	Model is Wrong
Decision about Fit	Accept Model (Data fit)		
	Reject Model (Data do not fit)		

Statistical Decision Making

- We might incorrectly reject the model
  - Observe an effect in what is really random variation

		State of the World	
		Model is Correct	Model is Wrong
Decision about Fit	Accept Model (Data fit)		
	Reject Model (Data do not fit)	Type I Error α p value	

Statistical Decision Making

- We might incorrectly accept the model
  - Miss an effect hidden within random variation

		State of the World	
		Model is Correct	Model is Wrong
Decision about Fit	Accept Model (Data fit)		Type II Error β
	Reject Model (Data do not fit)	Type I Error α p value	

Statistical Decision Making

- We might correctly reject the model
  - Observe an effect that exists

		State of the World	
		Model is Correct	Model is Wrong
Decision about Fit	Accept Model (Data fit)		Type II Error β
	Reject Model (Data do not fit)	Type I Error α p value	(1-β) Power

Statistical Decision Making

- We might correctly accept the model

		State of the World	
		Model is Correct	Model is Wrong
Decision about Fit	Accept Model (Data fit)	(1-α)	Type II Error β
	Reject Model (Data do not fit)	Type I Error α p value	(1-β) Power

Statistical Decision Making

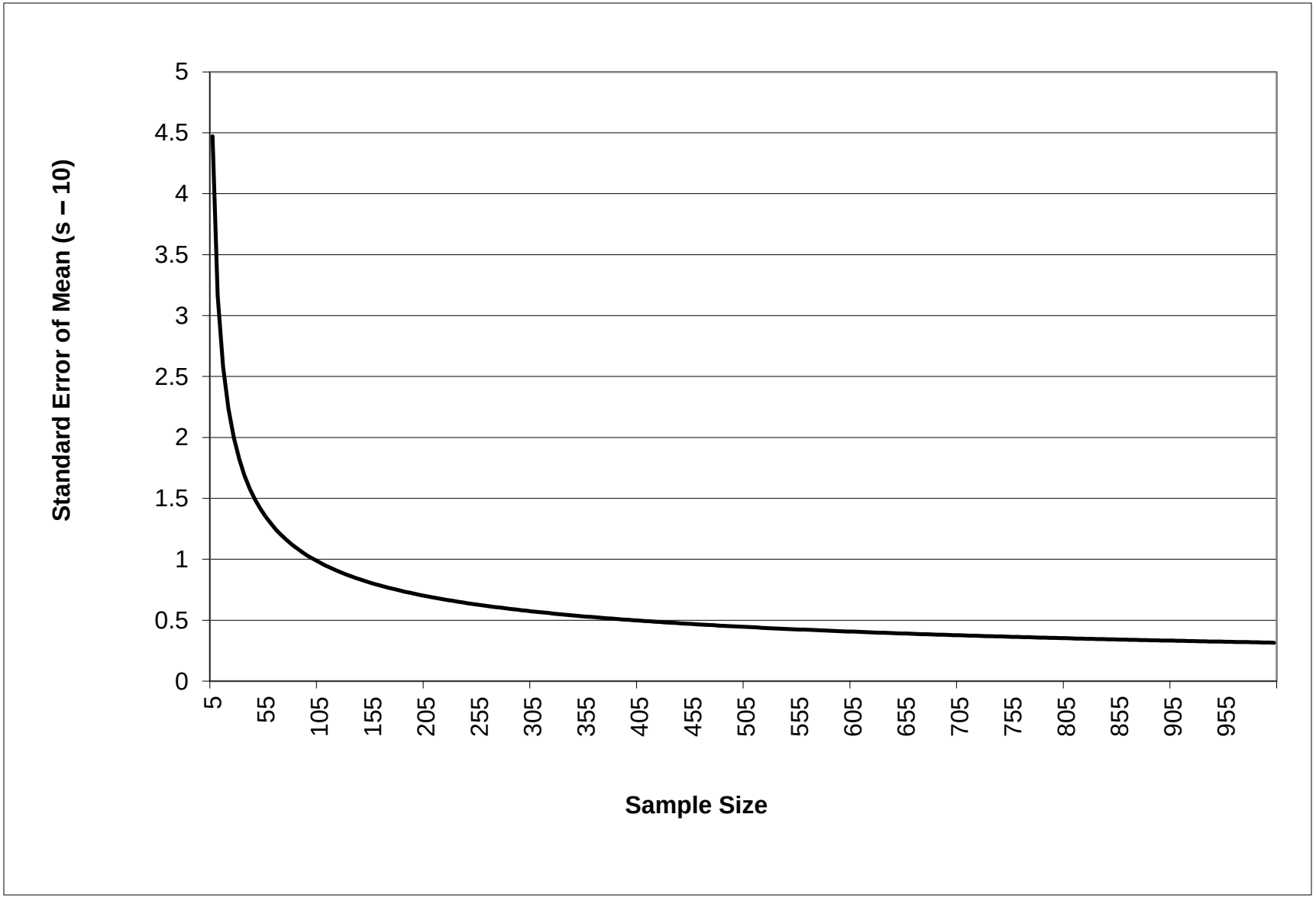
		State of the World	
		Model is Correct	Model is Wrong
Decision about Fit	Accept Model (Data fit)	(1-α)	Type II Error β
	Reject Model (Data do not fit)	Type I Error α p value	(1-β) Power

Sample Size Dependence

- Both α and β are inversely dependent upon sample size
  - Larger samples → less error
  - Larger samples → more power
- Function of Standard Error
- $$S_{\bar{x}} = \frac{s}{\sqrt{n}}$$

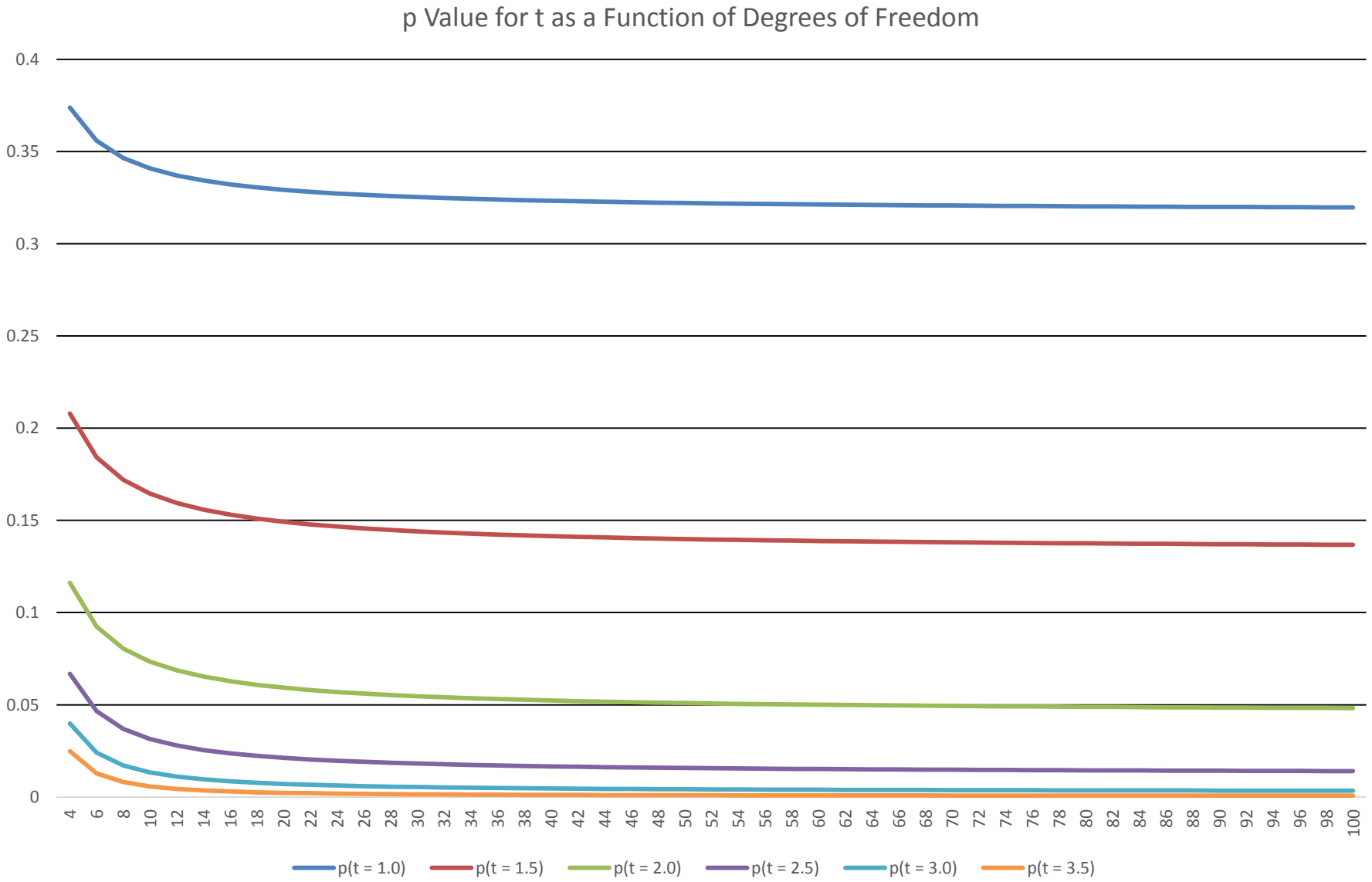
  s = sample standard deviation

  n = sample size


Standard Error of Mean (s – 10)
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
5
55
105
155
205
255
305
355
405
455
505
555
605
655
705
755
805
855
905
955
Sample Size

Degrees of Freedom

- All model-fitting statistics are based on standard error
  - Or an approximation (nonparametric statistics)
- Usually represented by degrees of freedom
  - Number of estimable data points before last data points are determined by the model
- Student's t as example
  - df = n-1 (per group)


p Value for t as a Function of Degrees of Freedom
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48 50 52 54 56 58 60 62 64 66 68 70 72 74 76 78 80 82 84 86 88 90 92 94 96 98 100
p(t = 1.0)
p(t = 1.5)
p(t = 2.0)
p(t = 2.5)
p(t = 3.0)
p(t = 3.5)

Limitations of $p < .05$ (or $< .0001$)

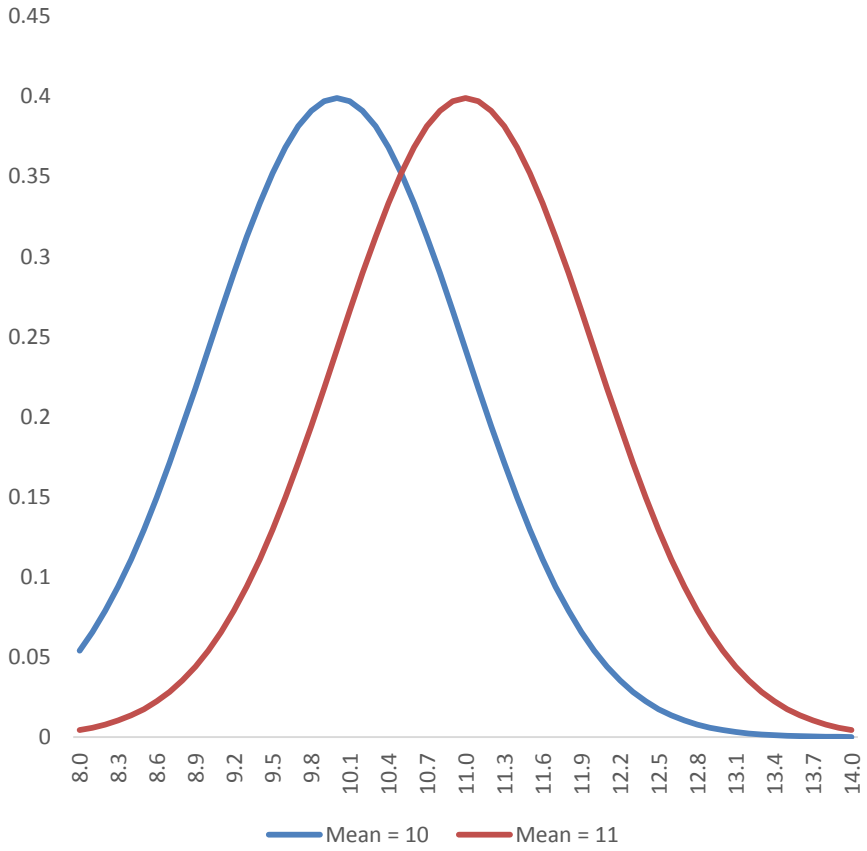
- Smaller p value does NOT mean larger effect
  - It is an indication of lack of fit between model and data
- p value is not relevant to the truth of the conclusion
- $p < .05$ (or any other number) is not a magic limen
  - "significance" should not be used as sole basis for a conclusion, policy, treatment, etc.
- $p < .05$ is basis for further consideration

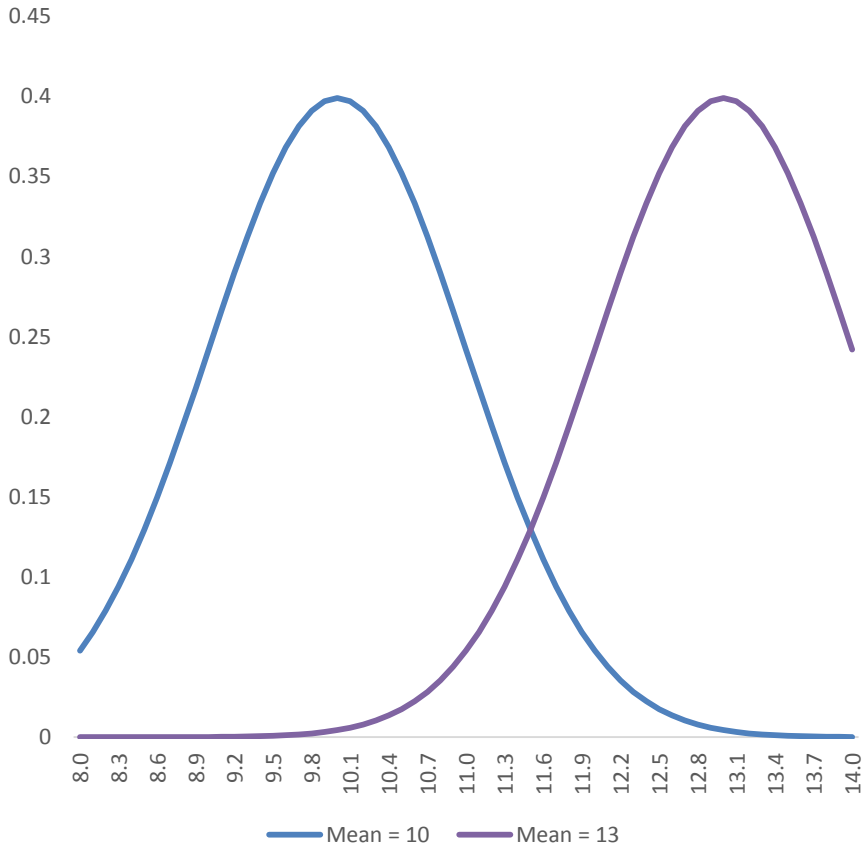
Effect Size

- Measures how much data deviate from model
  - How large is the mean difference?
  - How strong is the relationship between variables?
  - How well does X predict Y?
  - How much variance is accounted for?

Which is the Larger Effect?

(Both are “significant”)





Measures of Effect Size

- Mean Differences
  - Cohen's d
  - Partial η^2
- Relationships
  - r^2 or R^2
  - β for regression
  - OR, RR, etc.

Summary

- p value is probability of incorrectly rejecting the notion that the data fit the selected model
- p value is not an indication of the truth or validity of conclusions
- p value is an inadequate standard for deciding on importance
- Despite common usage, there is no mathematically justified p value for “significant”
- Effect size is an important part of interpreting results

Additional Reading

- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2 ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Dane, F. C. (2011). *Evaluating Research: Methodology for People Who Need to Read Research*. Thousand Oaks, CA: Sage Publications.
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic & Applied Social Psychology, 37*(1), 1-2. doi: 10.1080/01973533.2015.1012991
- Wasserstein, R. L. & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician.* DOI: 10.1080/00031305.2016.1154108
- Weisberg, H. I., Carver, R. P., Chachkin, N. J., & Cohen, D. K. (1979). Correspondence. *Harvard Educational Review, 49*, 280-283.
- https://en.wikipedia.org/wiki/Effect_size

Questions?

- fcdane@jchs.edu
- 540-224-4515 (office)
- 540-588-9404 (mobile)