Part II

Second-generation p-values: equivalence tests, statistical properties, and false discovery rates

Jeffrey D. Blume, PhD
School of Data Science
University of Virginia

and

Megan H. Murray, PhD
Department of Biostatistics
Vanderbilt University

Outline

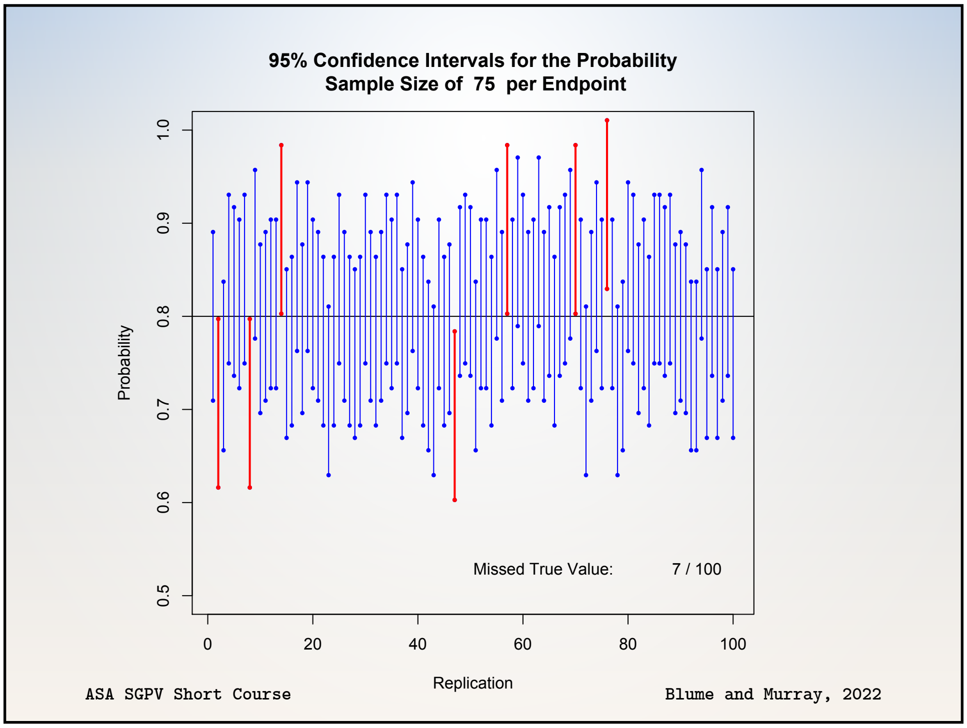
- → Statistical Properties
- → False Discovery Rates
- → Study Planning
- → Comparison to Equivalence Tests
- → Variable Selection References

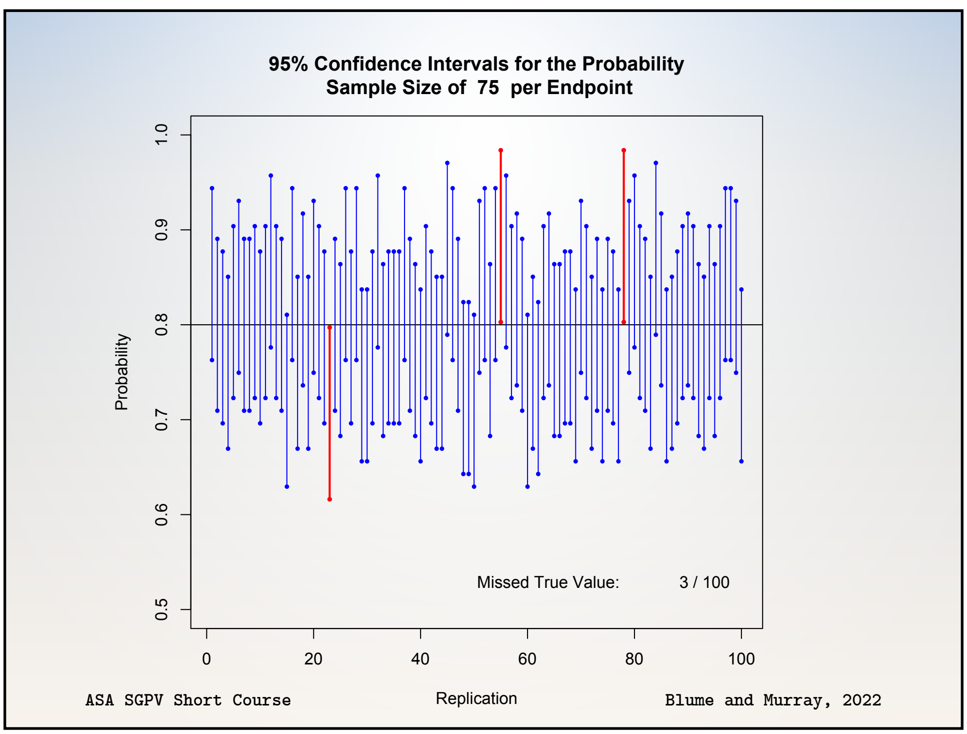
Connecting SGPV and Properties

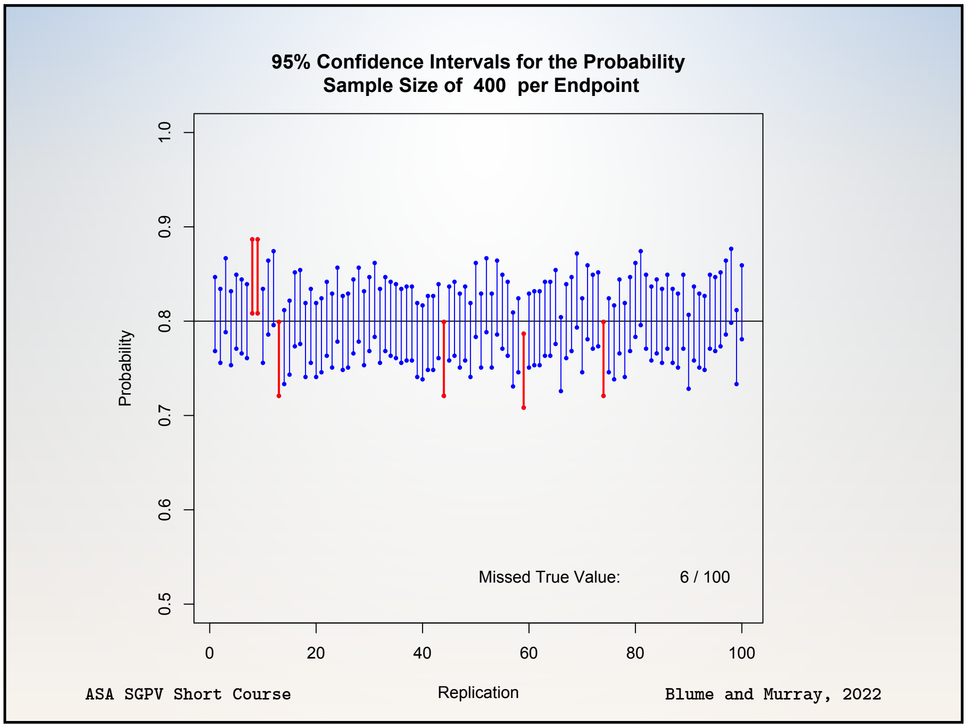
- Three Inferential Categories
  1. $p_\delta = 0 \Rightarrow$ data incompatible with null
  2. $p_\delta = 1 \Rightarrow$ data compatible with null
  3. $0 < p_\delta < 1 \Rightarrow$ data are inconclusive
- Three 'error' rates
  1. $P(p_\delta = 0|H_0)$ when H is null
  2. $P(p_\delta = 1|H_1)$ when H is not null
  3. $P(0 < p_\delta < 1|H)$ when H is either
- Assume H makes statements about a parameter θ
- Large sample setting

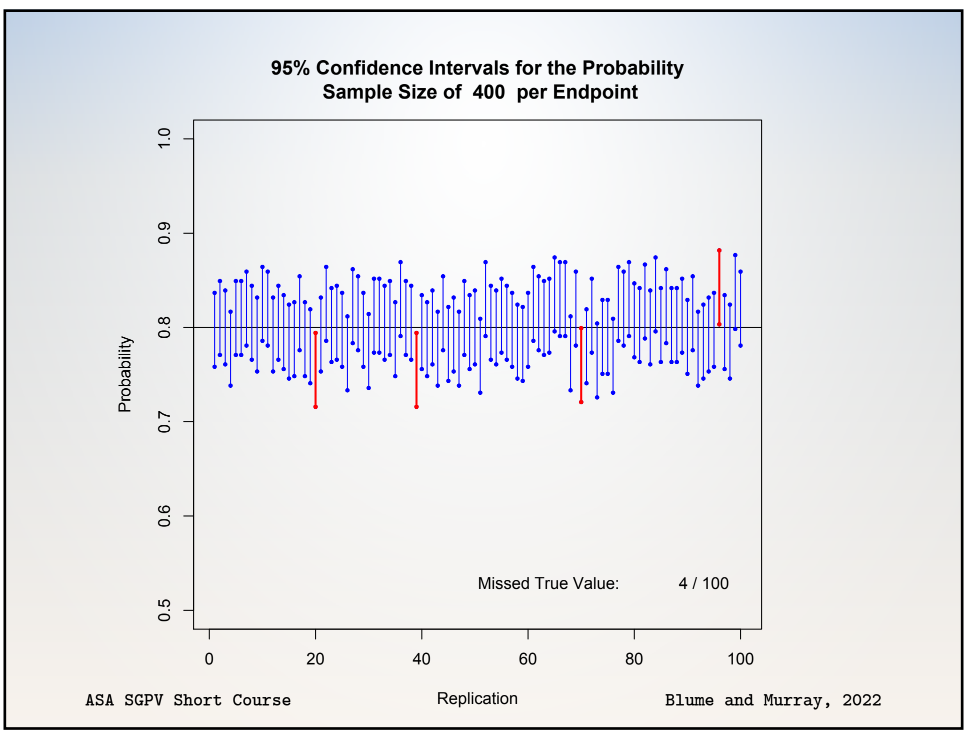
The Intuition

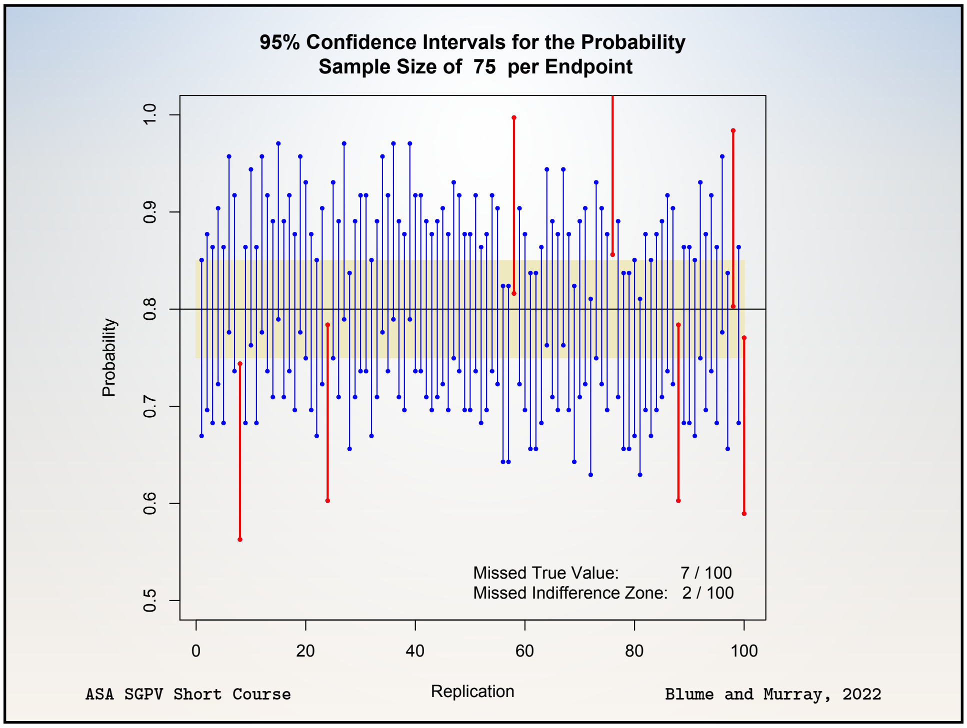
- SGPV error rates converge to zero as the sample size grows.
- Why? The indifference zone is responsible for this.
- Example: 95% Confidence Intervals
- Increasing the sample size…
  - Does not change the miss rate (!)
  - Reduces the width of the CI
  - Reduces the amount by which the CIs miss the truth (!)
- The complement of the coverage probability bounds the SGPV error rates

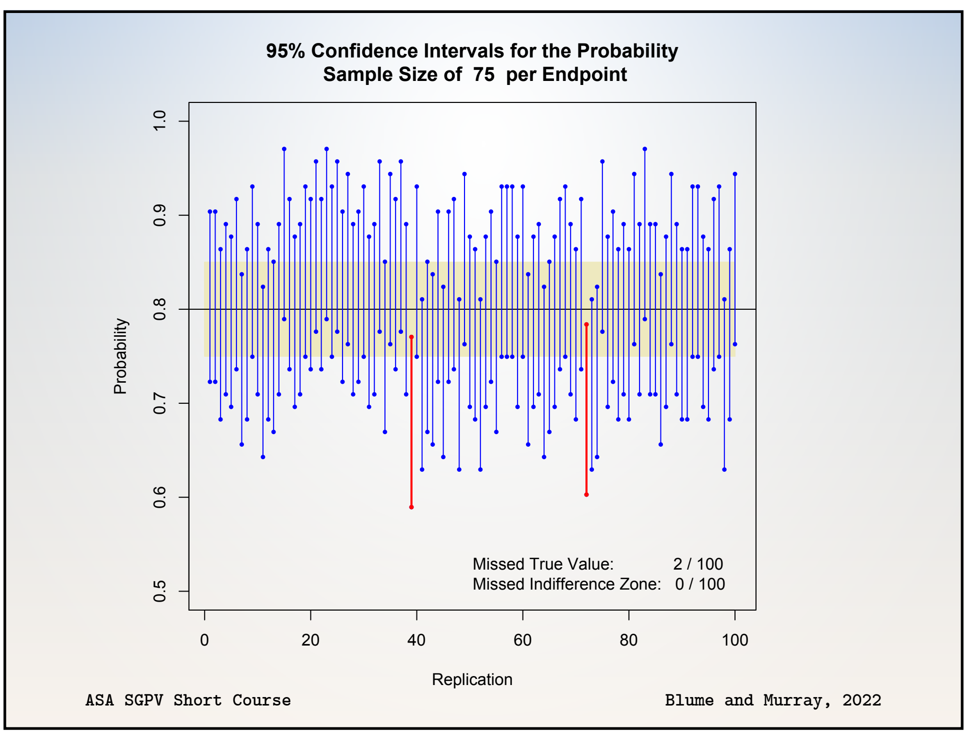

95% Confidence Intervals for the Probability
Sample Size of 75 per Endpoint
Probability
1.0
0.9
0.8
0.7
0.6
0.5
Missed True Value: 7 / 100
0
20
40
60
80
100
Replication
ASA SGPV Short Course
Blume and Murray, 2022

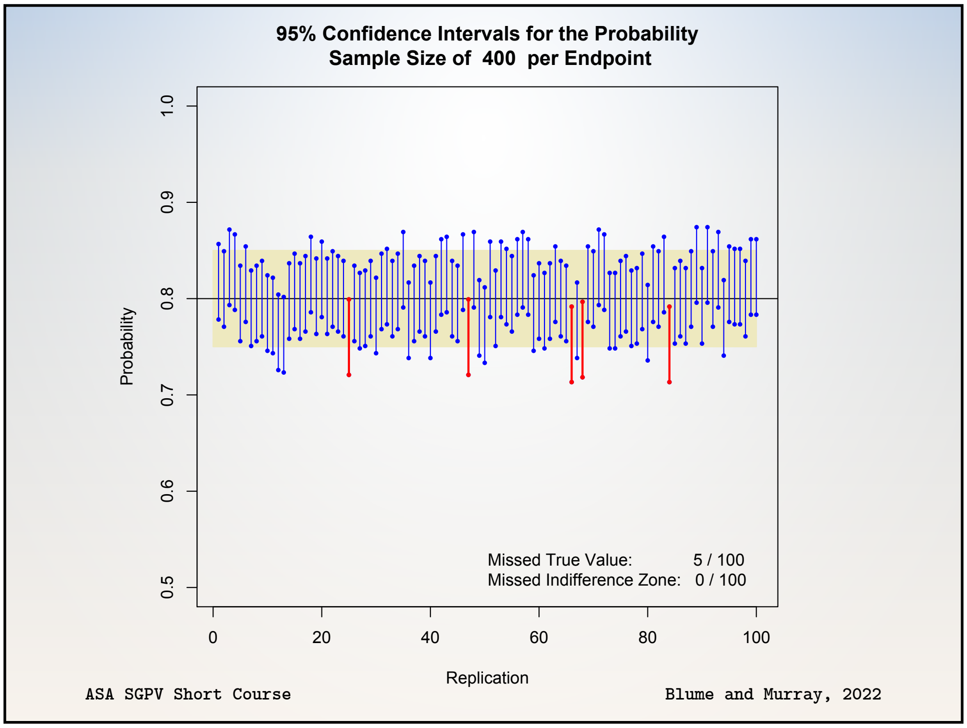

95% Confidence Intervals for the Probability
Sample Size of 75 per Endpoint
Probability
1.0
0.9
0.8
0.7
0.6
0.5
Missed True Value: 3 / 100
0
20
40
60
80
100
Replication
ASA SGPV Short Course
Blume and Murray, 2022

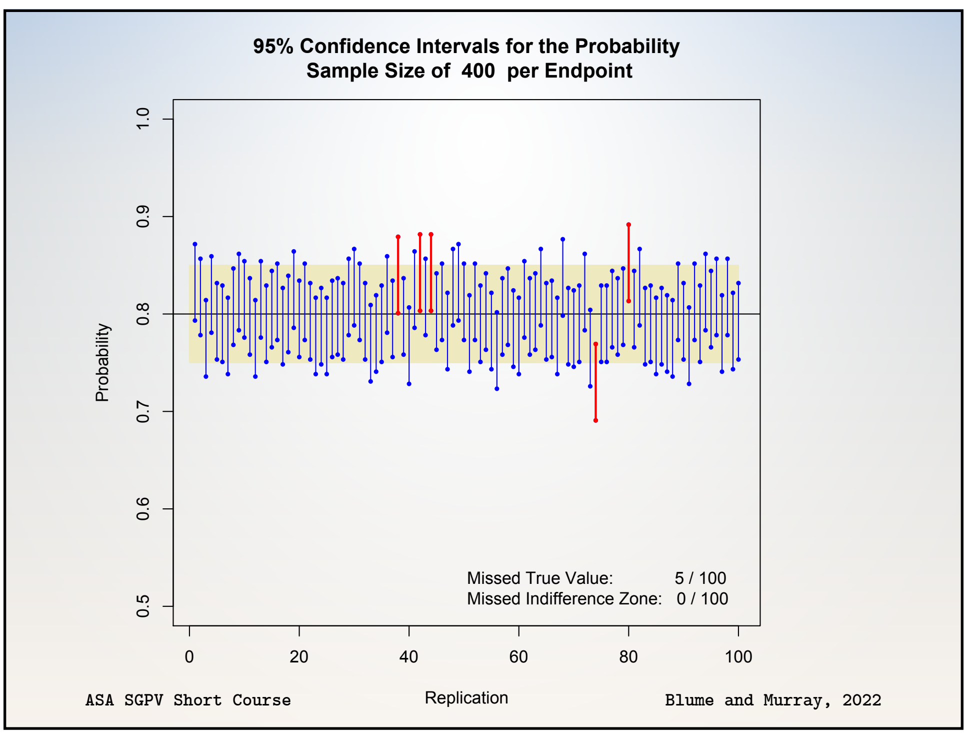

95% Confidence Intervals for the Probability
Sample Size of 400 per Endpoint
Probability
1.0
0.9
0.8
0.7
0.6
0.5
Missed True Value: 6 / 100
0
20
40
60
80
100
ASA SGPV Short Course
Replication
Blume and Murray, 2022


95% Confidence Intervals for the Probability
Sample Size of 400 per Endpoint
Probability
1.0
0.9
0.8
0.7
0.6
0.5
Missed True Value: 4 / 100
0
20
40
60
80
100
ASA SGPV Short Course
Replication
Blume and Murray, 2022


95% Confidence Intervals for the Probability
Sample Size of 75 per Endpoint
Probability
Replication
Missed True Value: 7 / 100
Missed Indifference Zone: 2 / 100
ASA SGPV Short Course
Blume and Murray, 2022


95% Confidence Intervals for the Probability
Sample Size of 75 per Endpoint
Probability
Replication
Missed True Value: 2 / 100
Missed Indifference Zone: 0 / 100
ASA SGPV Short Course
Blume and Murray, 2022


95% Confidence Intervals for the Probability
Sample Size of 400 per Endpoint
Probability
0.5
0.6
0.7
0.8
0.9
1.0
Missed True Value: 5 / 100
Missed Indifference Zone: 0 / 100
0
20
40
60
80
100
Replication
ASA SGPV Short Course
Blume and Murray, 2022


95% Confidence Intervals for the Probability
Sample Size of 400 per Endpoint
Probability
0.5
0.6
0.7
0.8
0.9
1.0
Missed True Value: 5 / 100
Missed Indifference Zone: 0 / 100
0
20
40
60
80
100
Replication
ASA SGPV Short Course
Blume and Murray, 2022

Statistical Properties

Suppose interval I has coverage probability 1-α, then

Three 'Error' Rates

1. $P(p_\delta = 0|H_0) \leq \alpha$ and $\rightarrow 0$ as $n \rightarrow \infty$
2. $P(p_\delta = 1|H_1) \leq \alpha$ and $\rightarrow 0$ as $n \rightarrow \infty$
3. $P(0 < p_\delta < 1|H)$ controlled through sample size

Will examine these first

Two False Discovery Rates

1. $P(H_0 \mid p_\delta = 0)$
2. $P(H_1 \mid p_\delta = 1)$

Will graph to illustrate

Statistical Properties

- How often are the data incompatible with null?
- Examine $P(p_\delta = 0|\theta)$ as θ varies
  - Power function
- This probability
  - converges to one for alternatives not near the edge of interval null
  - converges to zero for null hypotheses not near the edge of the null set
  - converges to alpha for hypotheses approaching or on the edge of the null set

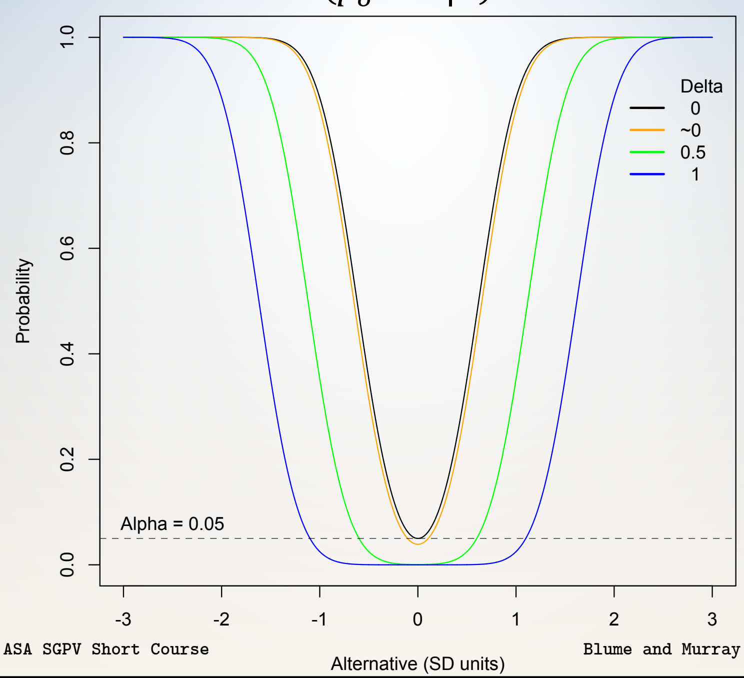
‘Power’ Function

- θ_0: point null, σ: standard deviaiton
- δ: half-width of indifference zone

$$P(p_\delta = 0|\theta) = \Phi\left[\frac{\sqrt{n}(\theta_0-\theta)}{\sigma} - \frac{\sqrt{n}\delta}{\sigma} - Z_{\alpha/2}\right] + \Phi\left[-\frac{\sqrt{n}(\theta_0-\theta)}{\sigma} - \frac{\sqrt{n}\delta}{\sigma} - Z_{\alpha/2}\right]$$

$$P_{\theta_0}(p_\delta = 0|\theta_0) = 2\Phi\left[-\frac{\sqrt{n}\delta}{\sigma} - Z_{\alpha/2}\right]$$

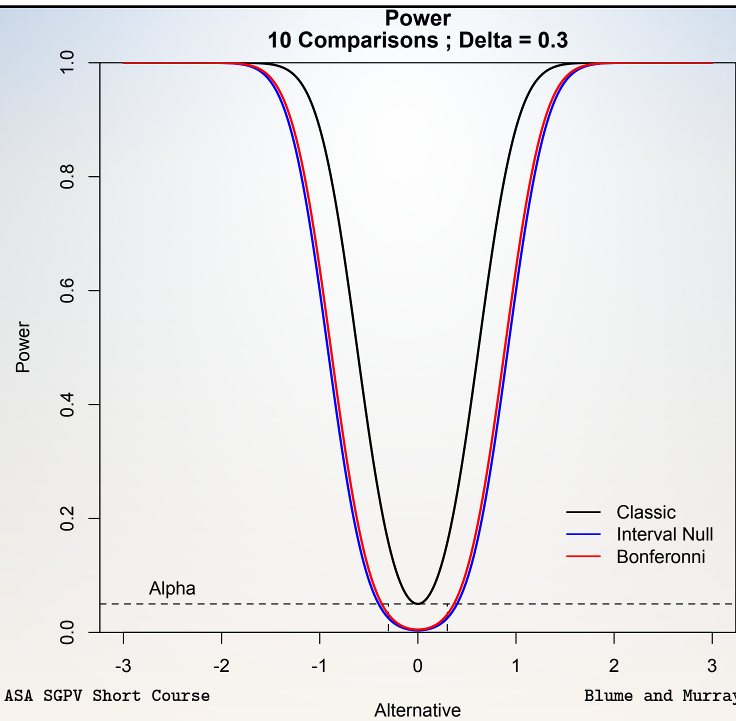
$P(p_\delta = 0|\theta)$

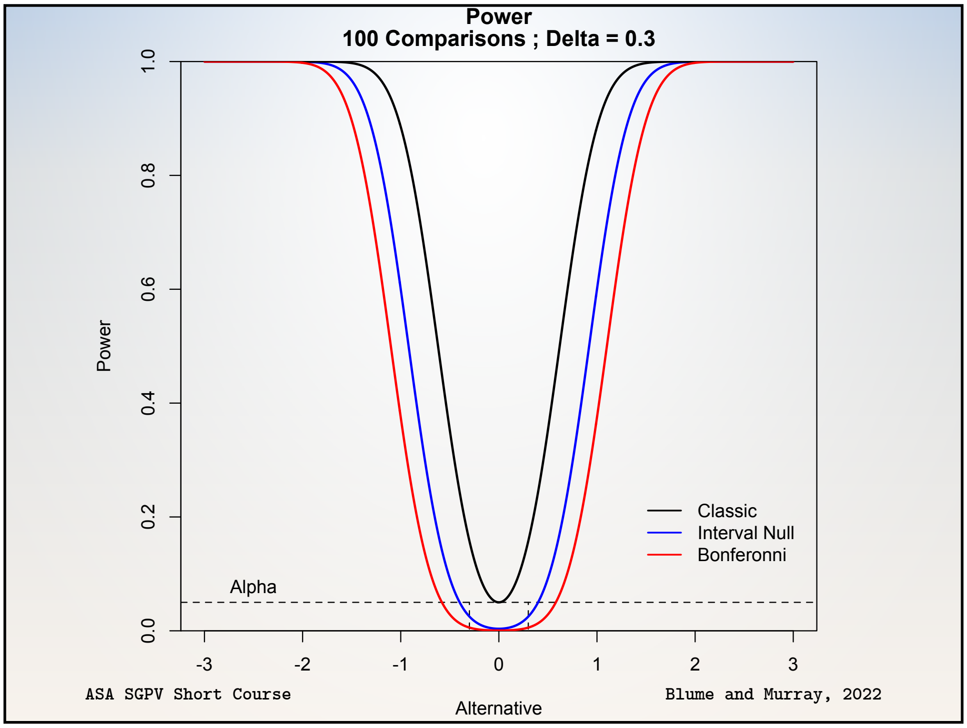


Compare with standard methods

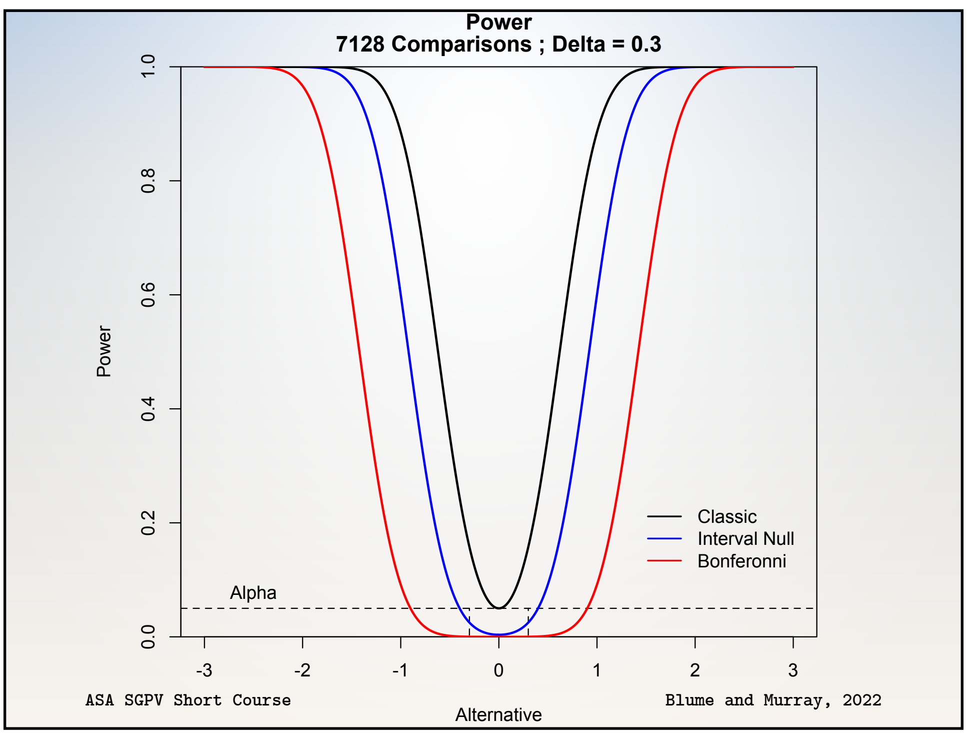
- Second-generation p-value vs. Bonferroni correction
  - Adjusted for {10, 100, 7128} comparisons
  - Leukemia data example

- Remember SGPV are not adjusted for comparisons
- Discuss?





Power
100 Comparisons ; Delta = 0.3
Power
0.0
0.2
0.4
0.6
0.8
1.0
Alpha
Classic
Interval Null
Bonferonni
-3
-2
-1
0
1
2
3
Alternative
ASA SGPV Short Course
Blume and Murray, 2022



Power
7128 Comparisons ; Delta = 0.3
Power
0.0
0.2
0.4
0.6
0.8
1.0
Alpha
Classic
Interval Null
Bonferonni
-3
-2
-1
0
1
2
3
Alternative
ASA SGPV Short Course
Blume and Murray, 2022

Compatible with Null

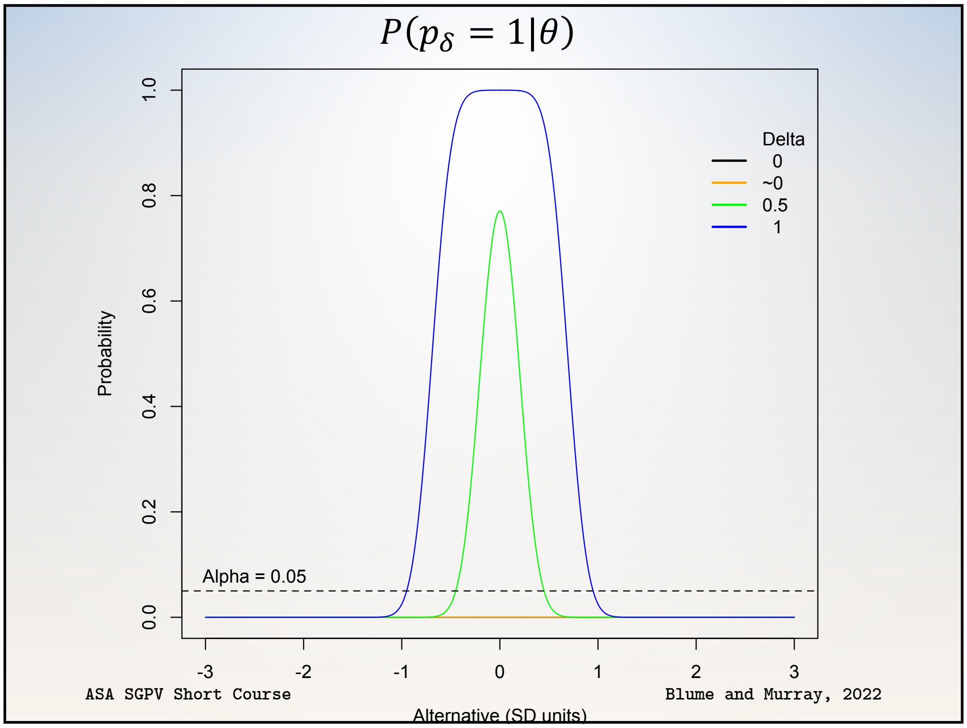
- How often are the data compatible with null?
- Examine $P(p_\delta = 1|\theta)$ when θ is null or practically null
  - Essentially opposite of power function
- Sample size must be large enough to allow the null interval to contain the interval estimate
- This probability
  - converges to zero or one quickly for alternatives not near the edge of interval null

'Null Power' Function

- **How often are the data compatible with null?**
- Sample size must be large enough to allow null interval to contain the interval estimate, so $(\delta > Z_{\alpha/2}/\sqrt{n})$ or $(\sqrt{n} > Z_{\alpha/2}/\delta)$
- This probability converges to 0 or 1 quickly

$$P(p_\delta = 1|\theta) = \Phi\left[\frac{\sqrt{n}(\theta_0+\delta)}{\sigma} - \frac{\sqrt{n}\theta}{\sigma} - Z_{\alpha/2}\right]$$
$$-\Phi\left[\frac{\sqrt{n}(\theta_0-\delta)}{\sigma} - \frac{\sqrt{n}\theta}{\sigma} + Z_{\alpha/2}\right]$$

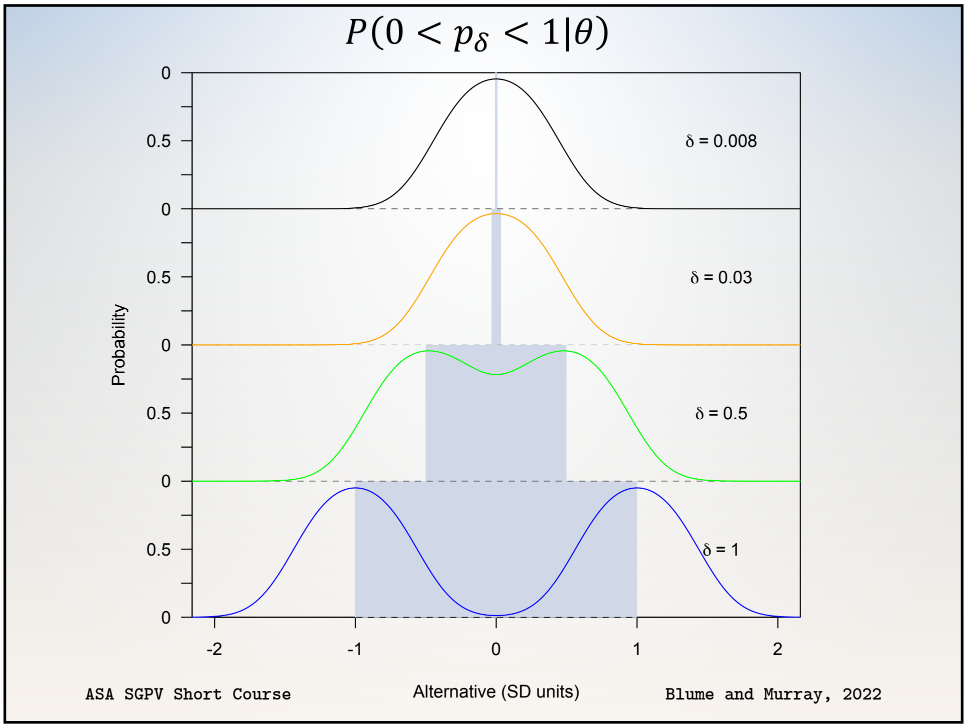
when $\delta > Z_{\alpha/2}/\sqrt{n}$

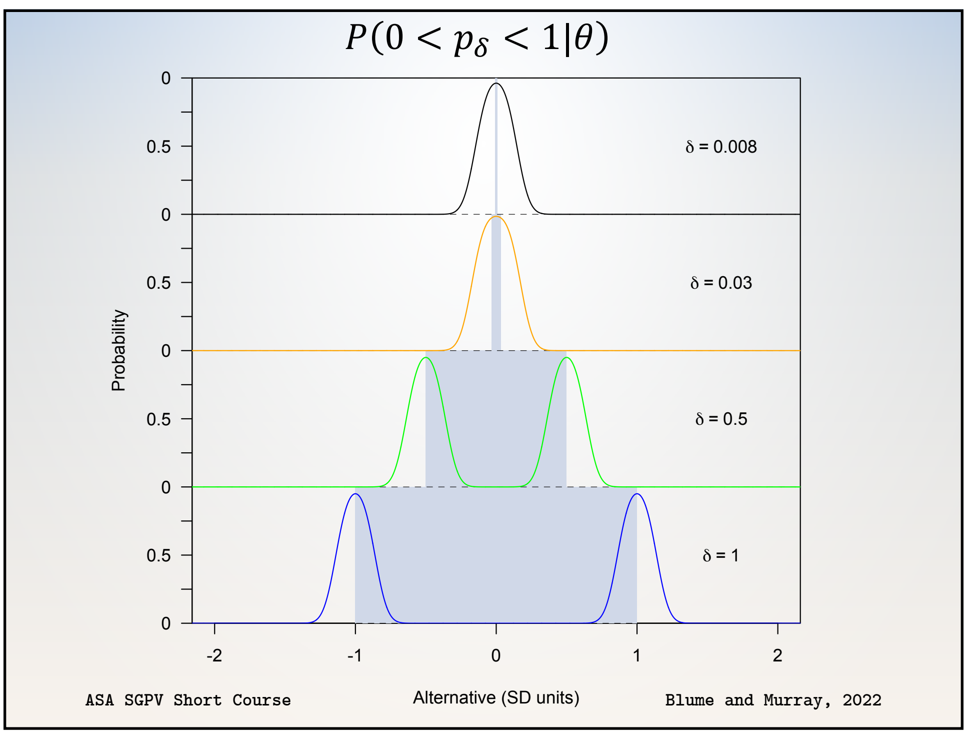


Probability of Inconclusive Data

- How often are the data inconclusive?
- Examine $P(0 < p_\delta < 1|\theta)$ for various θ
- This probability
  - drives sample size projections
  - is maximized when H is near the interval null edge
  - decreases quickly as H moves away from edge of null
- $P(0 < p_\delta < 1|\theta) = 1 - P(p_\delta = 0|\theta) - P(p_\delta = 1|\theta)$

ASA SGPV Short Course — Blume and Murray, 2022


$P(0 < p_\delta < 1|\theta)$
δ = 0.008
δ = 0.03
δ = 0.5
δ = 1
Probability
Alternative (SD units)
ASA SGPV Short Course
Blume and Murray, 2022


$P(0 < p_\delta < 1|\theta)$
δ = 0.008
δ = 0.03
δ = 0.5
δ = 1
Probability
Alternative (SD units)
ASA SGPV Short Course
Blume and Murray, 2022

Remarks

- Second-generation p-values…
  - Has three ''Error'' rates
  - Allows Type I and II rate to converge to zero
  - Control changes of inconclusive results
  - Controls error rate using *science*
  - Reduces the false discovery rate (next section)
- Anchoring the scale of the effect size…
  - Eliminates most Type I Errors
  - Improves scientific translation of statistical model

Statistical Properties

Suppose interval I has coverage probability 1-α, then

Three 'Error' Rates

1. $P(p_\delta = 0|H_0) \leq \alpha$ and $\rightarrow 0$ as $n \rightarrow \infty$
2. $P(p_\delta = 1|H_1) \leq \alpha$ and $\rightarrow 0$ as $n \rightarrow \infty$
3. $P(0 < p_\delta < 1|H)$ controlled through sample size

Two False Discovery Rates

1. $P(H_0 \,|\, p_\delta = 0)$
2. $P(H_1 \,|\, p_\delta = 1)$

False Discovery Rates

- FDR for 5 SGPV=0 findings; computed under various null and alternative configurations (w/ flat prior).

SNP ID	SGPV rank	p-value rank	OR	1/8 SI lower limit	1/8 SI upper limit	FDR_1	FDR_2	FDR_3
kgp4568244_C	1	133	0.10	0.03	0.37	2.9%	17.1%	3.3%
kgp8051290_G	13	2002	15.58	1.95	124.68	4.3%	30.3%	4.9%
kgp4497498_A	28	255	4.37	1.80	10.64	2.5%	8.6%	3.1%
rs3123636_G	423	1	1.39	1.26	1.55	0.01%	0.1%	0.4%
kgp7460928_G	1443	3310	1.78	1.11	2.87	2.4%	2.0%	3.0%

False discovery rates

- Impact of α=0.05 vs α=0.05/7128 (7128 comparisons)

Error rates

- False Discovery Rate (FDR)

$$P(H_0|p<\alpha) = \left[1 + \frac{(1-\beta)}{\alpha} r\right]^{-1}$$

- False Confirmation Rate (FCR)

$$P(H_1|p>\alpha) = \left[1 + \frac{(1-\alpha)}{\beta}\frac{1}{r}\right]^{-1}$$

$r = P(H_1)/P(H_0)$

Error rates

False discovery rates

- Second-generation p-values
- False Discovery Rate (FDR)

Error Rates

$$P(H_0|p_\delta = 0) = \left[1 + \frac{P(p_\delta = 0|H_1)}{P(p_\delta = 0|H_0)} r\right]^{-1}$$

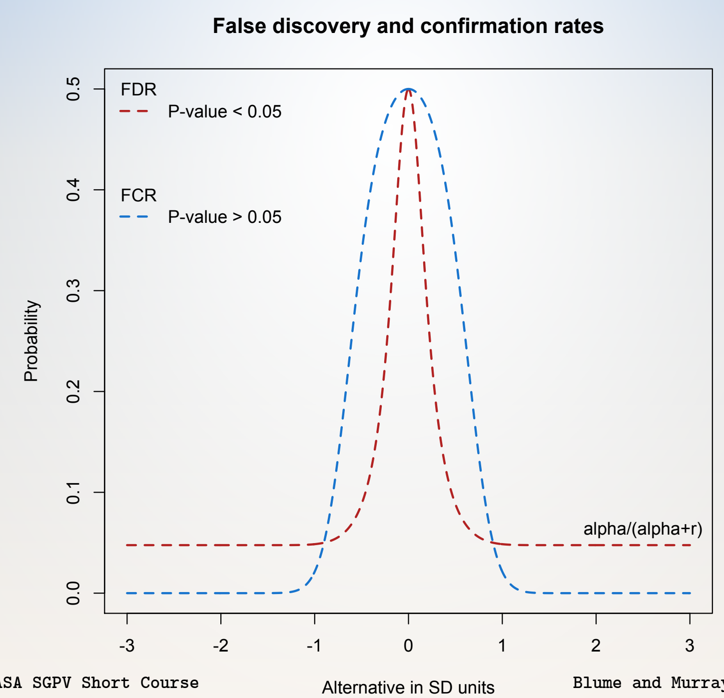
- False Confirmation Rate (FCR)

$$P(H_1|p_\delta = 1) = \left[1 + \frac{P(p_\delta = 1|H_0)}{P(p_\delta = 1|H_1)} \frac{1}{r}\right]^{-1}$$

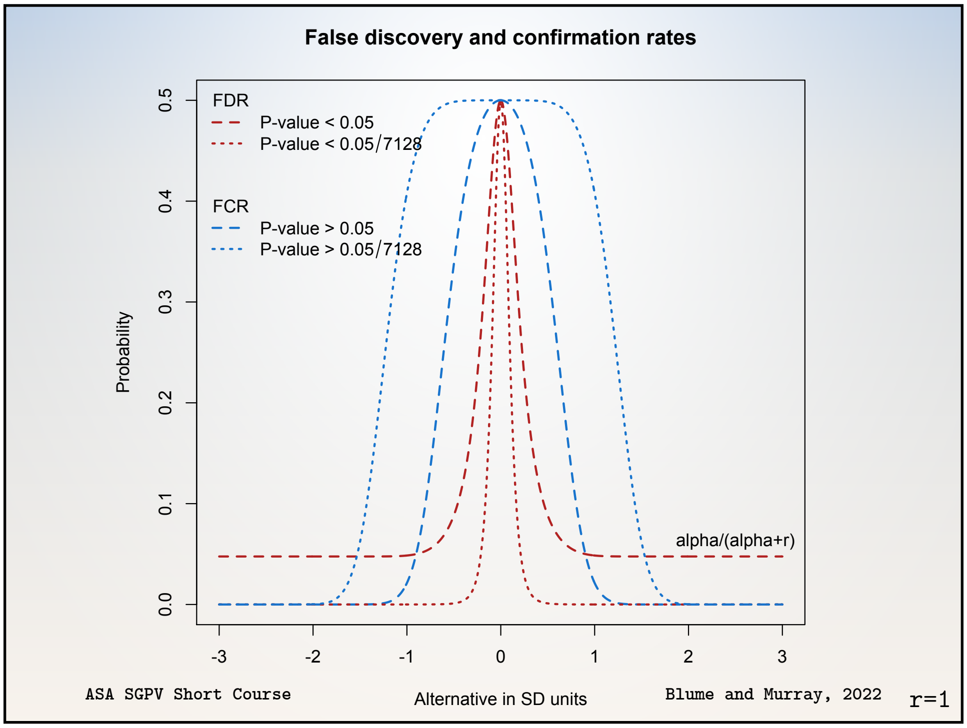
Error Rates

$r = P\,(H_1)/P\,(H_0)$

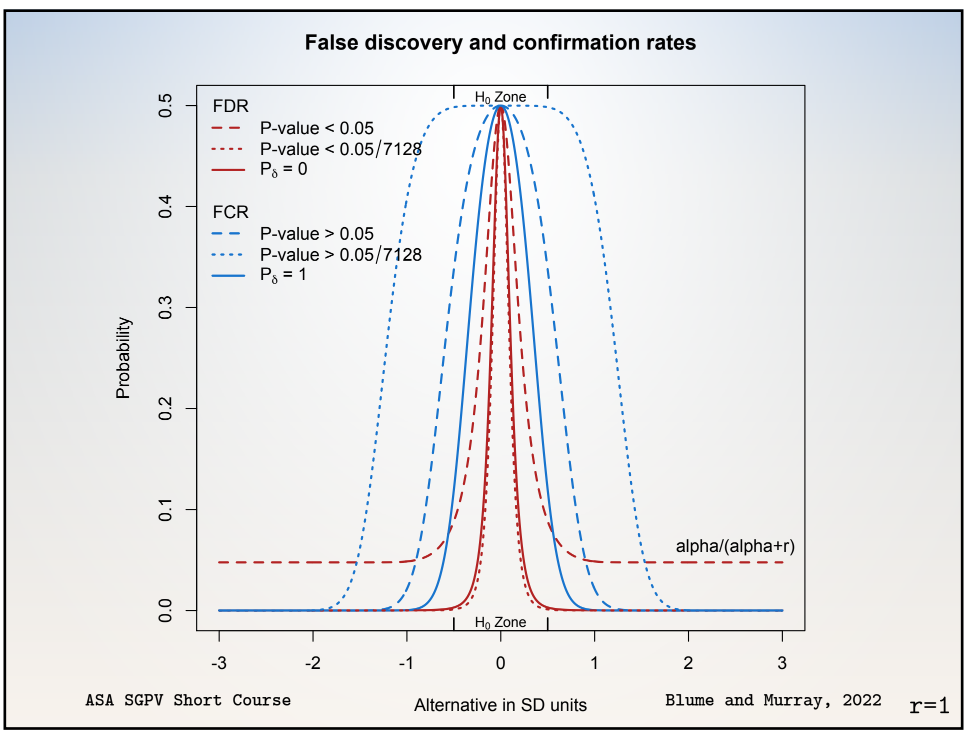
ASA SGPV Short Course Blume and Murray, 2022



ASA SGPV Short Course Blume and Murray, 2022 r=1



False discovery and confirmation rates
FDR
P-value < 0.05
P-value < 0.05/7128
FCR
P-value > 0.05
P-value > 0.05/7128
alpha/(alpha+r)
Probability
0.0
0.1
0.2
0.3
0.4
0.5
-3
-2
-1
0
1
2
3
Alternative in SD units
ASA SGPV Short Course
Blume and Murray, 2022
r=1



False discovery and confirmation rates
H_0 Zone
FDR
P-value < 0.05
P-value < 0.05/7128
P_δ = 0
FCR
P-value > 0.05
P-value > 0.05/7128
P_δ = 1
alpha/(alpha+r)
H_0 Zone
Probability
0.0
0.1
0.2
0.3
0.4
0.5
-3
-2
-1
0
1
2
3
Alternative in SD units
ASA SGPV Short Course
Blume and Murray, 2022
r=1

Study Planning for SGPVs

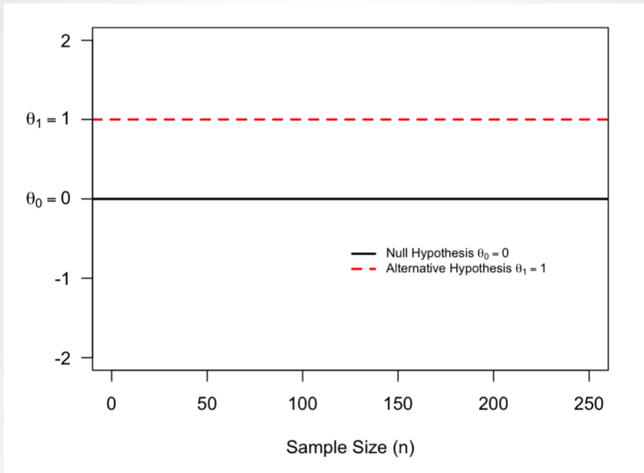
- Purpose: Evaluate how different techniques of setting the indifference zone effect the SGPV study properties.

- What options does a collaborator have when they are uncertain of the indifference zone?

Traditional p-value vs. SGPV

		Null is true		Alternative is true	
		Traditional *p*-value	SGPV	Traditional *p*-value	SGPV
Inference	Data are consistent with the alternative	Type I error $P_{\theta_0}(p \le 0.05)$	β_0 $P_{\theta_0}(p_\delta = 0)$	Power or (1 - Type II error) $P_{\theta_1}(p \le 0.05)$	β_1 $P_{\theta_1}(p_\delta = 0)$
	Data are inconclusive	1 - Type I error $P_{\theta_0}(p > 0.05)$	γ_0 $P_{\theta_0}(0 < p_\delta < 1)$	Type II error $P_{\theta_1}(p > 0.05)$	γ_1 $P_{\theta_1}(0 < p_\delta < 1)$
	Data are consistent with the null	NA	ω_0 $P_{\theta_0}(p_\delta = 1)$	NA	ω_1 $P_{\theta_1}(p_\delta = 1)$

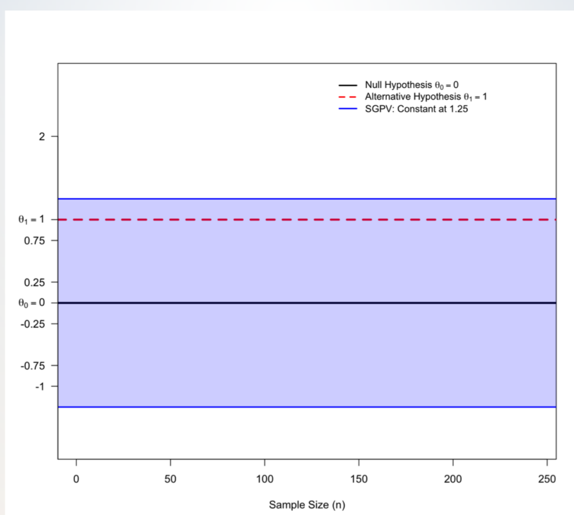
Standard Assumptions



ASA SGPV Short Course

Blume and Murray, 2022

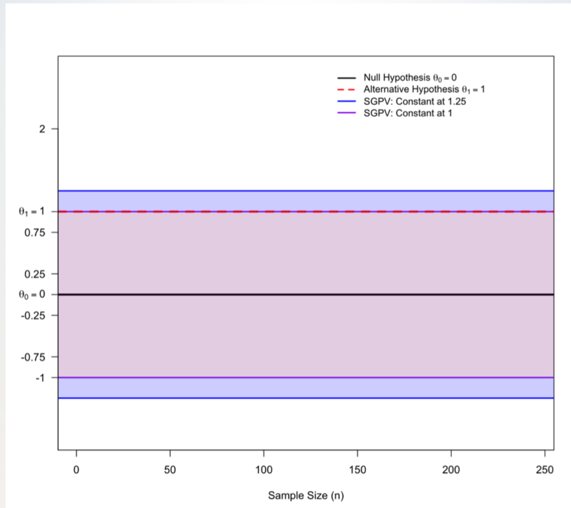
Example Indifference Zones



ASA SGPV Short Course

Blume and Murray, 2022

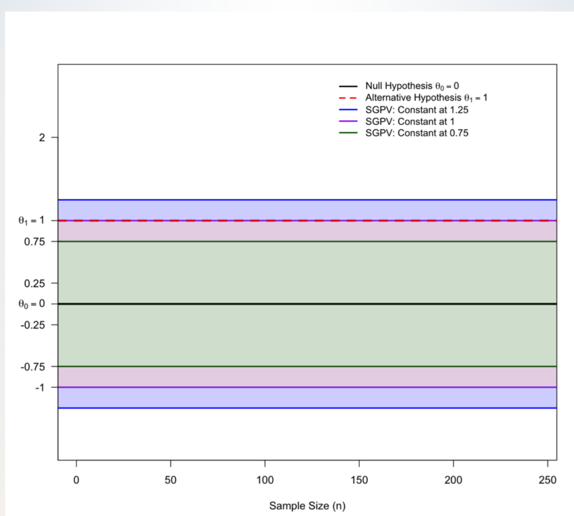
Example Indifference Zones



ASA SGPV Short Course

Blume and Murray, 2022

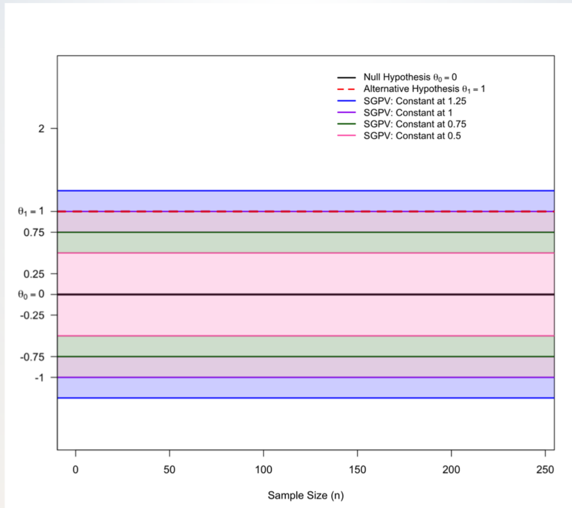
Example Indifference Zones



ASA SGPV Short Course

Blume and Murray, 2022

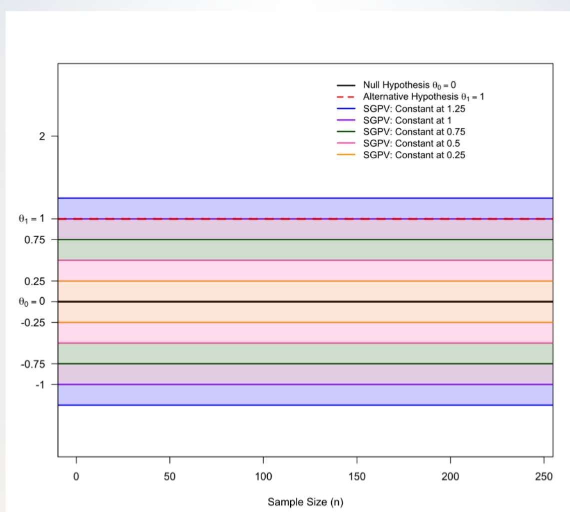
Example Indifference Zones



ASA SGPV Short Course

Blume and Murray, 2022

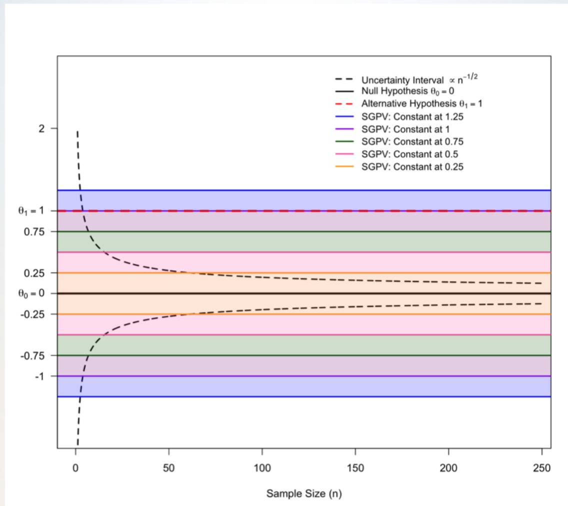
Example Indifference Zones



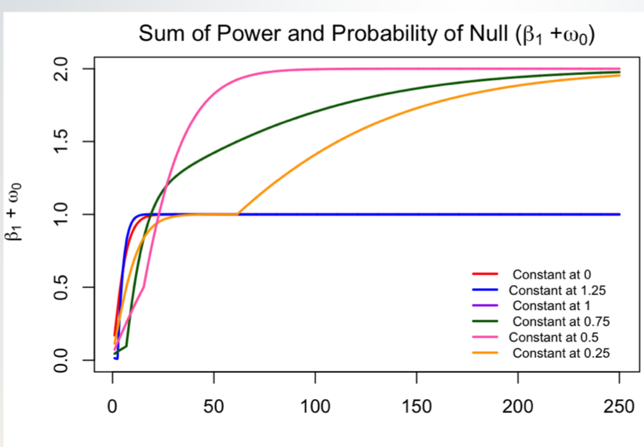
ASA SGPV Short Course

Blume and Murray, 2022

Now with CI centered at null



Indifference Zone Properties



- Maximize power and probability of true nulls
- Plot $\beta_1 + \omega_0$ over sample size
- The half a standard deviation between $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_0}$ gives the highest sum for reasonable sample sizes

Recommendations for Collaborator Uncertainty

Collaborator Hypothesis	Suggested SGPV Analysis	Outcome
Confident in a null zone.	Use it!	Ideal
Confident in the alternative point closest to null.	Use the halfway point as the null zone.	Great
Uncertain of null zone.	Use the null zone that shrinks.	Good
Uncertain of alternative point.	Use the null zone that shrinks.	Ok
Cannot identify.	Do NOT use SGPV! (but can't ever measure evidence for null)	Poor

Equivalence Tests

- Establish bioequivalence between data and an established equivalence range or interval null
- Example: A pharmaceutical company tests for drug approval by comparing new drug's performance to an approved drug's performance
- Uses an interval null or equivalence range
  - $H_0 = [\theta^-, \theta^+]$

TOST Test

- Most popular frequentist test is the Two One-Sided t-tests (TOST) (Schuirmann 1987)
  - Flips the null and alternative (be careful)
  - Uses the $(1-2\alpha)\%$ confidence interval (be careful)
- Tests are ordinary, one-sided, α-level t-tests
- If *both* one-sided tests reject, then conclude the evidence is contained in the equivalence range

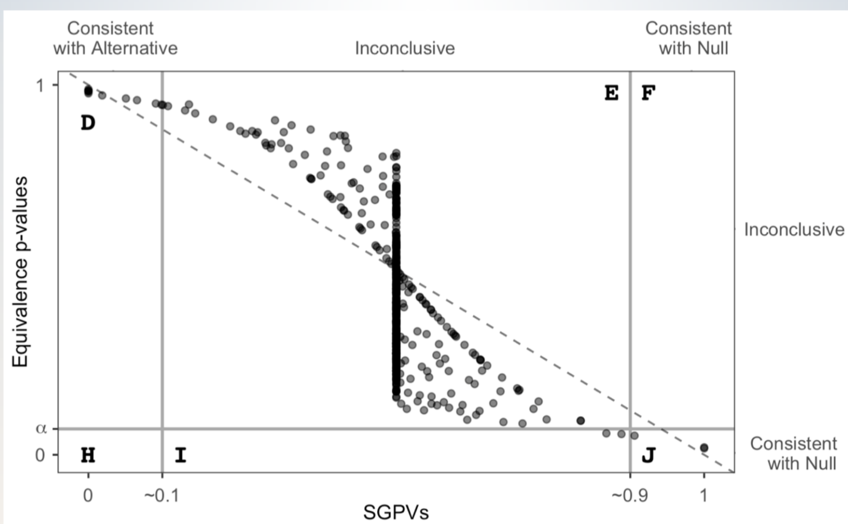
TOST vs. SGPV comparison

		SGPV Outcomes		
		Consistent with the alternative (SGPV near 0)	**Inconclusive** (SGPV near ½)	**Consistent with the null** (SGPV near 1)
Equivalence Tests Outcomes	**Consistent with the alternative** (*p*-value is unable to indicate this)	Not applicable **A**	Not applicable **B**	Not applicable **C**
	Inconclusive (*p*-value is non-significant)	Can occur **D**	Can occur **E**	Never occurs **F**
	Consistent with the null (*p*-value is significant)	Never occurs **H**	Can occur in small samples **I**	Can occur **J**

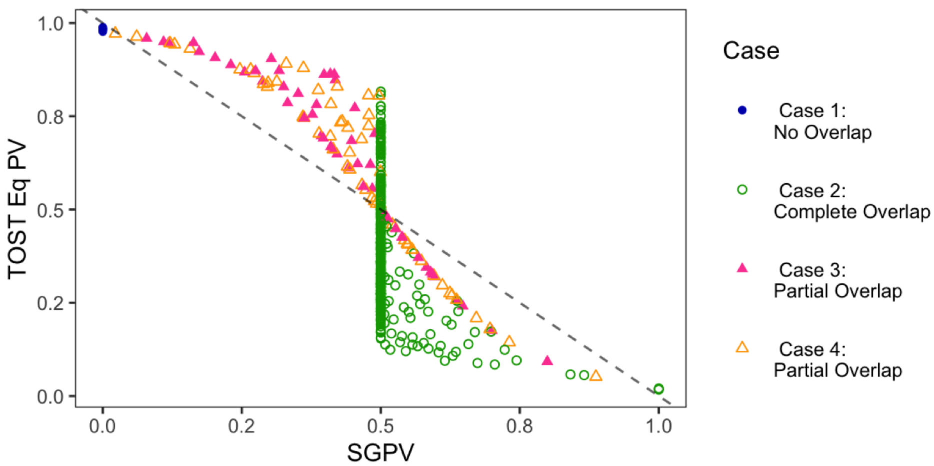
TOST v SGPV Simulation

- Simulate TOST and SGPV reported p-values
  - Data generated under the null, $N(0,1)$
  - Sample size of n=6
  - Yields 70% power for $\Delta = 1$ with 5% type 1 error
  - Indifference zone is $[\theta^-, \theta^+] = [-0.375, 0.375]$
  - Uncertainty interval is 95% confidence interval
  - 500 iterations (for illustration)
  - TOSTER R package (Link)

TOST vs. SGPV comparison



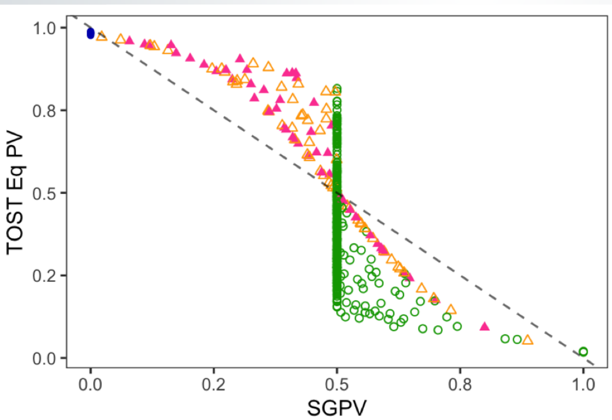
TOST vs. SGPV comparison

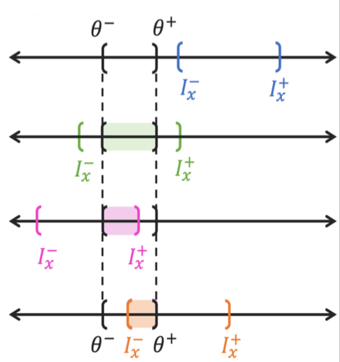


ASA SGPV Short Course

Blume and Murray, 2022

TOST vs. SGPV comparison





ASA SGPV Short Course

Blume and Murray, 2022

Relationship

- Derived the mathematical relationship in all cases of overlap
- Case 3 and 4 of partial overlap:

$$p_\delta = \left[\left(\frac{1}{2c_\alpha}\right)\mathrm{F}_\mathrm{n}^{-1}(1-p_T) + \frac{1}{2}\right] \times max\left\{\frac{\frac{c_\alpha S}{\sqrt{n}}}{(\theta^+ - \theta^-)}, 1\right\}$$

- Limiting behavior under the null

$$P(Case\ 2) \to 1$$

$p_\delta \to 1$ (right)

$p_T \to 0$ (bottom)

TOST vs. SGPV comparison

TOST	SGPV
• 2 inference outcomes	• 3 inference outcomes
• Conclusions only about $(1-2\alpha)$% confidence interval	• Any uncertainty data interval can be used
• Type I Error is ultra-conservative (distribution of p_T is non-uniform)	• Type I error is accurately assessed (Tied to width of data interval)
• Not uniformly most powerful	• Indicates when data agree with null or alternative without additional testing
• No measure of overlap included in computation	• Includes overlap in reported p-value

Variable Selection with SGPVs

- ProSGPV GitHub
  - https://github.com/zuoyi93/ProSGPV
- Papers
  - https://www.tandfonline.com/doi/full/10.1080/00031305.2021.1946150
  - f1000research.com/articles/11-58
- Linear ProSGPV
  - https://cran.r-project.org/web/packages/ProSGPV/vignettes/linear-vignette.html
- GLM and Cox ProSGPV
  - https://cran.r-project.org/web/packages/ProSGPV/vignettes/glm-cox-vignette.html

Review of Topics

→ Second-generation p-value framework, definition, and examples

→ The SGPV achieves the inferential properties that many scientists hope, or believe, are attributes of the classic p-value.

→ Statistical Properties of SGPVs

→ False Discovery Rates

→ Study Planning

→ Comparison to Equivalence Tests: Two One-Sided Tests (TOST)

→ SGPV Variable Selection

→ Coding Examples

Acknowledgements

- Collaborators
  - William D. Dupont
  - Robert A. Greevy
  - Lucy D’Agostino McGowan
  - Jonathan Chipman
  - Valerie Welty
  - Lisa Lin
  - Jeffrey R. Smith
  - Yi Zuo
  - Thomas G. Stewart
  - Vanderbilt SEDS Lab

- Website
  - www.statisticalevidence.com

Questions?