Lecture XI:
Power Analysis and Sample Size Calculations

Stanislao Maldonado

Universidad del Rosario

Impact Evaluation

May 2nd, 2017

1. Motivation

- Unless data collection is cheap, a sample is required to identify the causal effect of interest
- Two key questions:

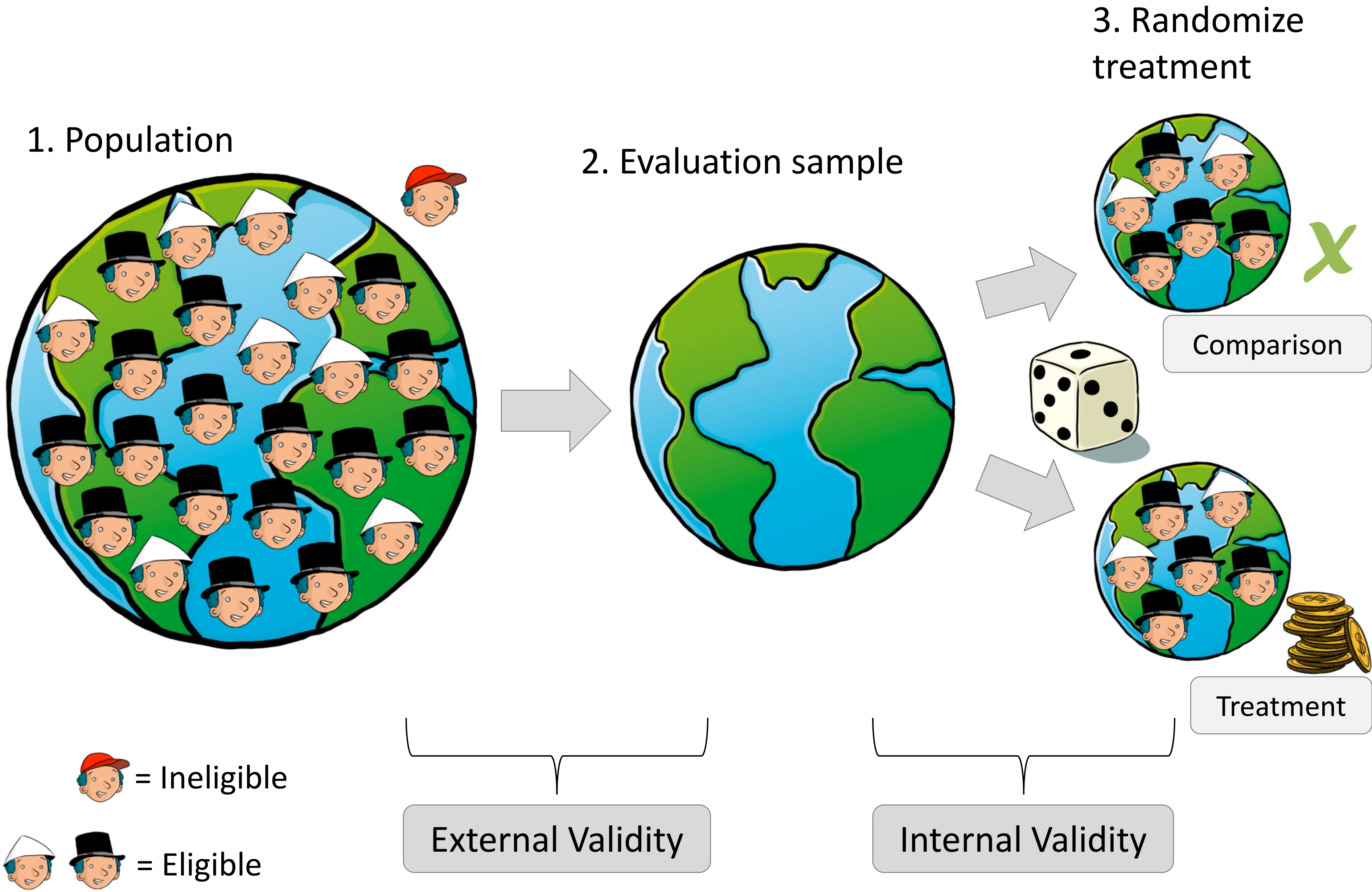
How large does the sample need to be to credibly detect a given effect size?

How will the sample be chosen?

- What does “credibly” mean here?

- Randomization (when possible) removes bias, but it does not remove noise: it works because of the law of large numbers… how “large” much large be?

- At the end of an experiment, we will compare the outcome of interest in the treatment and the comparison groups.
- But we do not observe the entire population, just a **sample**
- Sampling matters for external validity



1. Population

2. Evaluation sample

3. Randomize treatment

Comparison

Treatment

= Ineligible

= Eligible

External Validity

Internal Validity

Power Analysis for Individual-level Randomized Designs

2. Intro to Power Analysis

- We have shown that statistical significance is designed to be consistent with statistical regularity

- Signal/noise ratio becomes more distinct as the sample size increases

- Sample size calculations are performed to insure we take a large enough sample to be able to detect a difference if it exists

- Power analysis is performed to discover the probability of detecting a difference (if exists) for a planned sample size

- **Significance level**: *the probability* that we will reject the null hypothesis even though it is true
- **Type I error**: rejecting the null hypothesis even though it is true (false positive)
- **Type II Error**: failing to reject the null hypothesis (concluding there is no difference), when indeed the null hypothesis is false.
- **Power**: If there is a measureable effect of our intervention (the null hypothesis is false), the probability that we will detect an effect (reject the null hypothesis)

Statistical power is the ability to reject the hypothesis that program doesn’t work when it really does

Ho = program does not increase test scores

			True state of the world		
			Program does not change test scores	Program increases test scores	

This column represents Program working

Statistical power is the ability to reject the hypothesis that program doesn’t work when it really does

Ho = program does not increase test scores

		True state of the world	
		Program does not change test scores	Program increases test scores
Estimate	Program does not change test scores		
	Program increases test scores		

This row represents the evaluation finding that program is working

Statistical power is the ability to reject the hypothesis that program doesn't work when it really does

Ho = program does not increase test scores

		True state of the world	
		Program does not change test scores	Program increases test scores
Estimate	Program does not change test scores		
	Program increases test scores		If program works, we want to maximize the chance that evaluation says it works

We believe program works

We hope evaluation will confirm that it works

There are two types of error that we want to avoid

Ho = program does not increase test scores

		True state of the world	
		Program does not change test scores	Program increases test scores
Estimate	Program does not change test scores	No evidence to reject Ho	
	Program increases test scores	Evaluation says program works when it doesn’t	Correct rejection of Ho

There are two types of error that we want to avoid

Ho = program does not increase test scores

		True state of the world	
		Program does not change test scores	Program increases test scores
Estimate	Program does not change test scores	No evidence to reject Ho	Evaluation says program doesn't work when it really does
	Program increases test scores	**Type I Error** **"False positive"**	Correct rejection of Ho

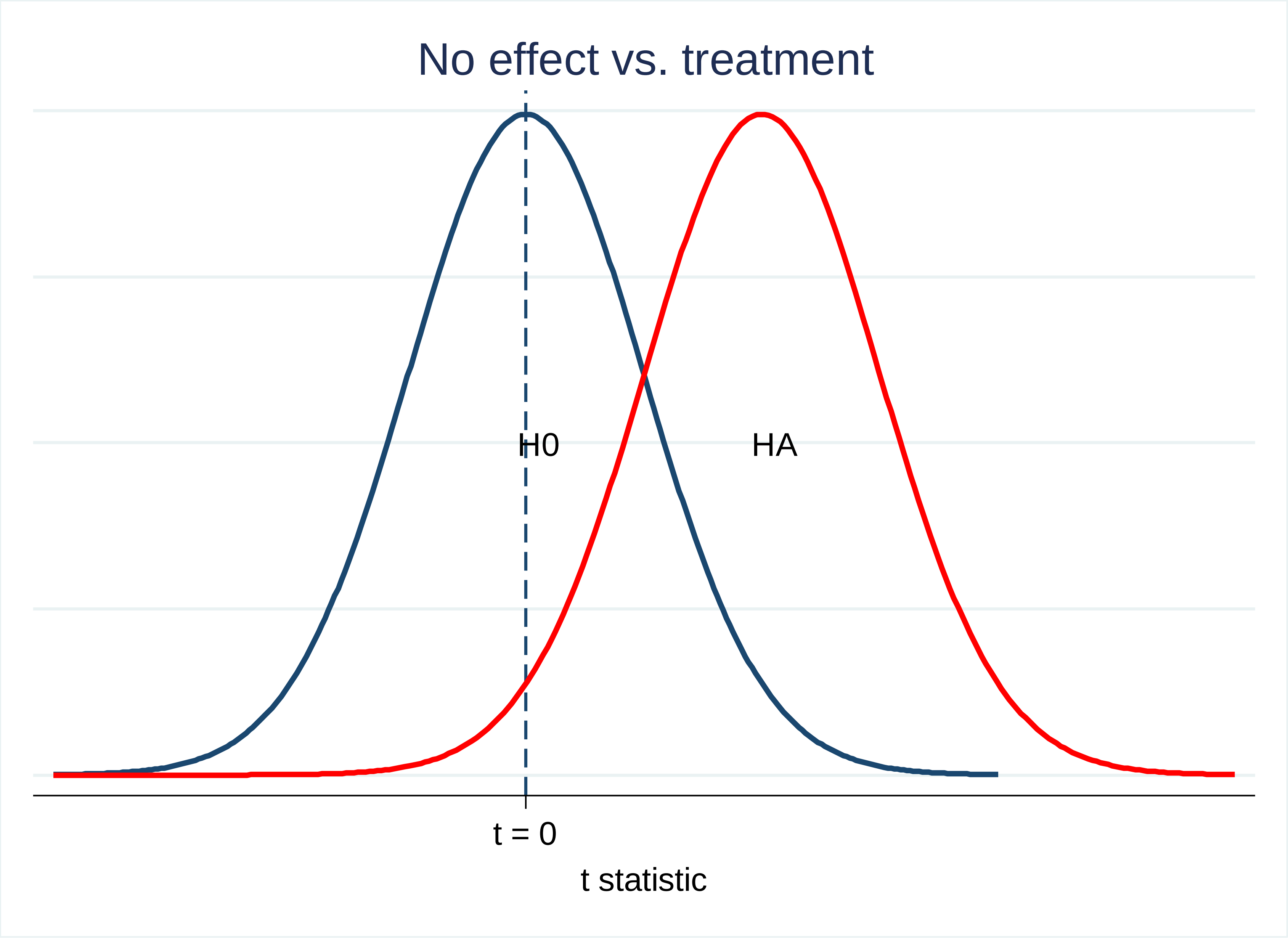
Statistical power is the chance that we reject the null hypothesis when it is false

Ho = program does not increase test scores

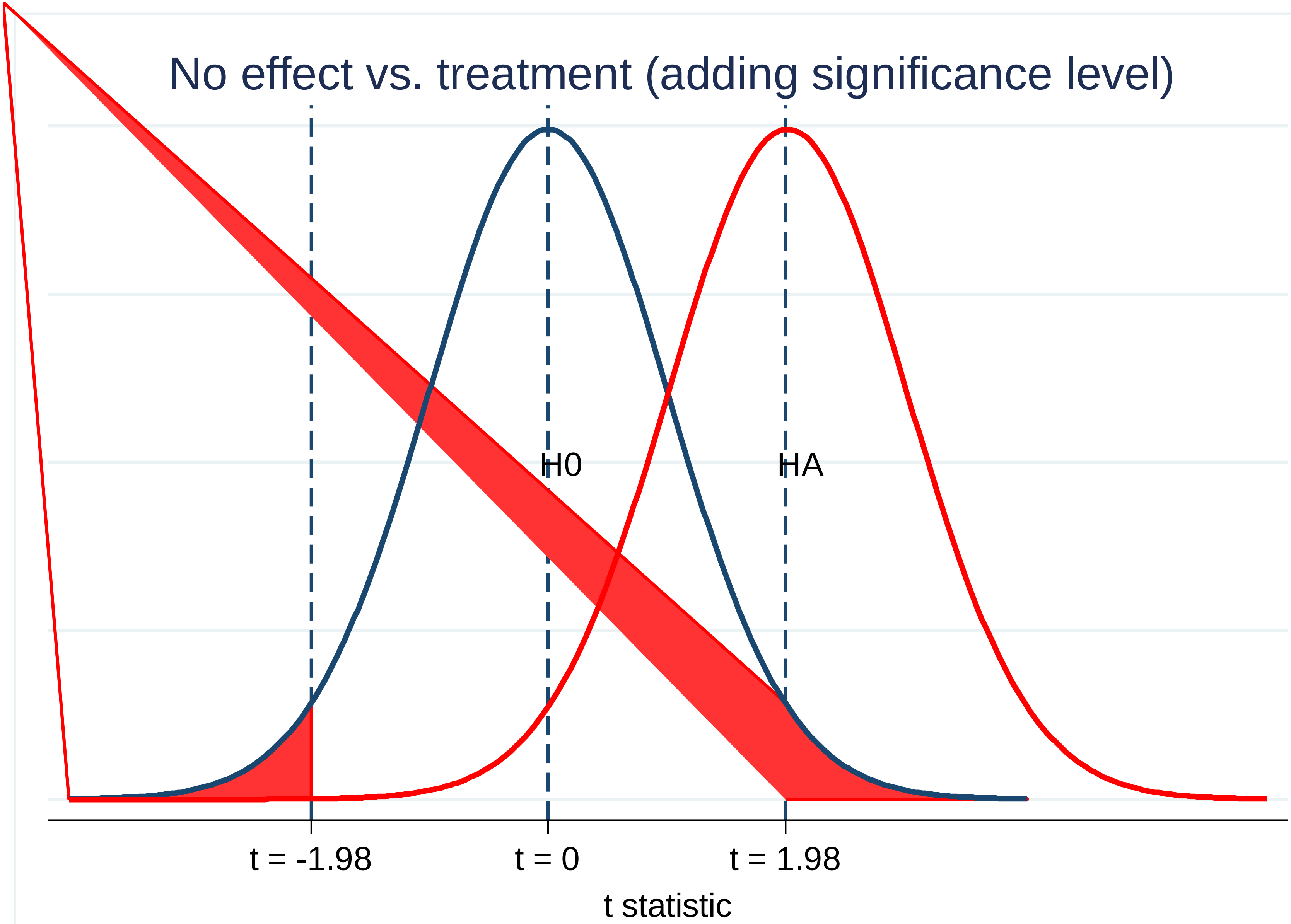
		True state of the world	
		Program does not change test scores	Program increases test scores
Estimate	Program does not change test scores	No evidence to reject Ho	**Type II Error** **“False Negative”**
	Program increases test scores	**Type I Error** **“False positive”**	Correct rejection of Ho

Power: probability that you reject “no impact” when there really is impact

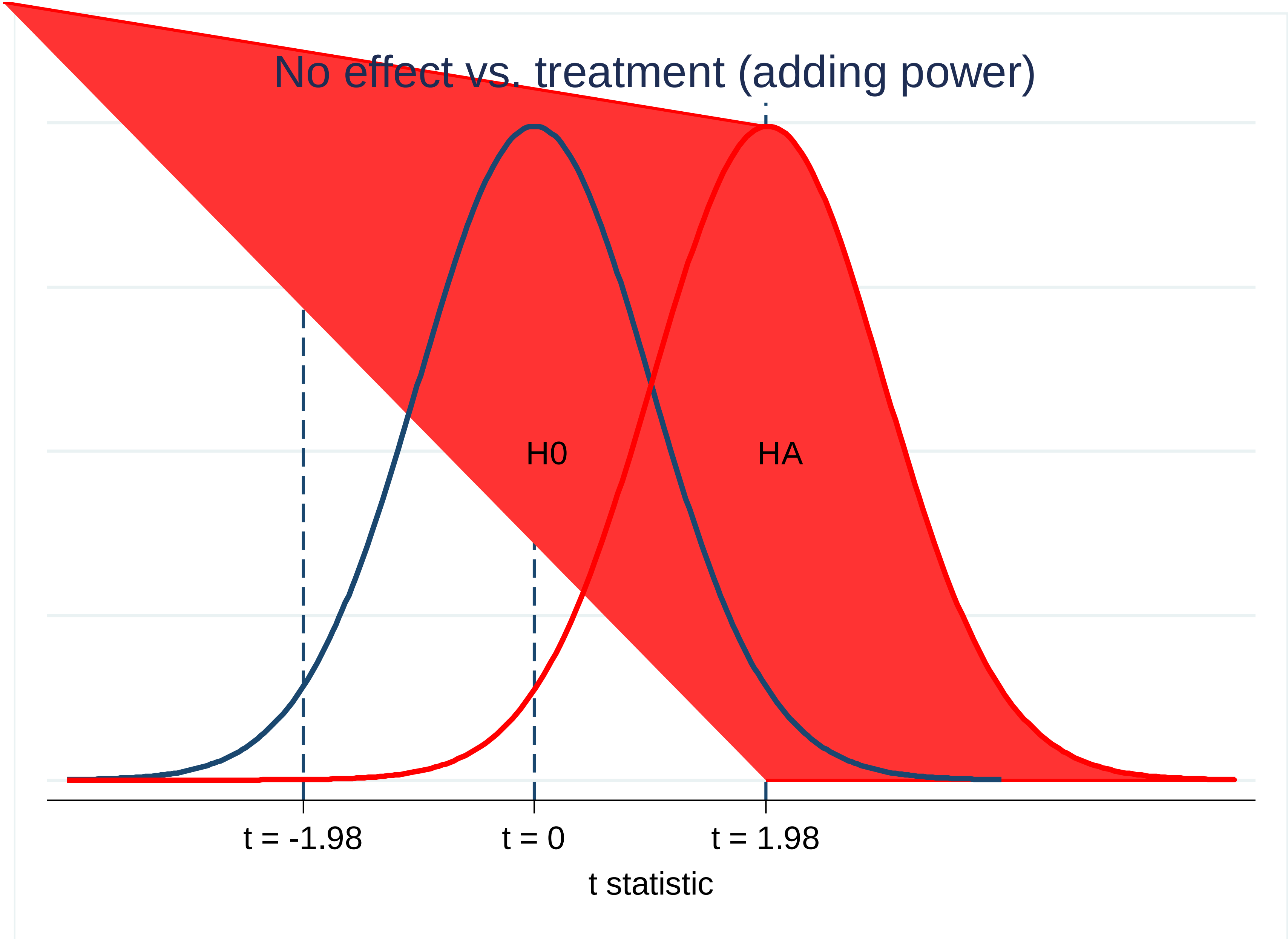
2.1 Power: Graphical treatment



No effect vs. treatment (adding significance level)



No effect vs. treatment (adding power)



2.2 Power: a formal treatment

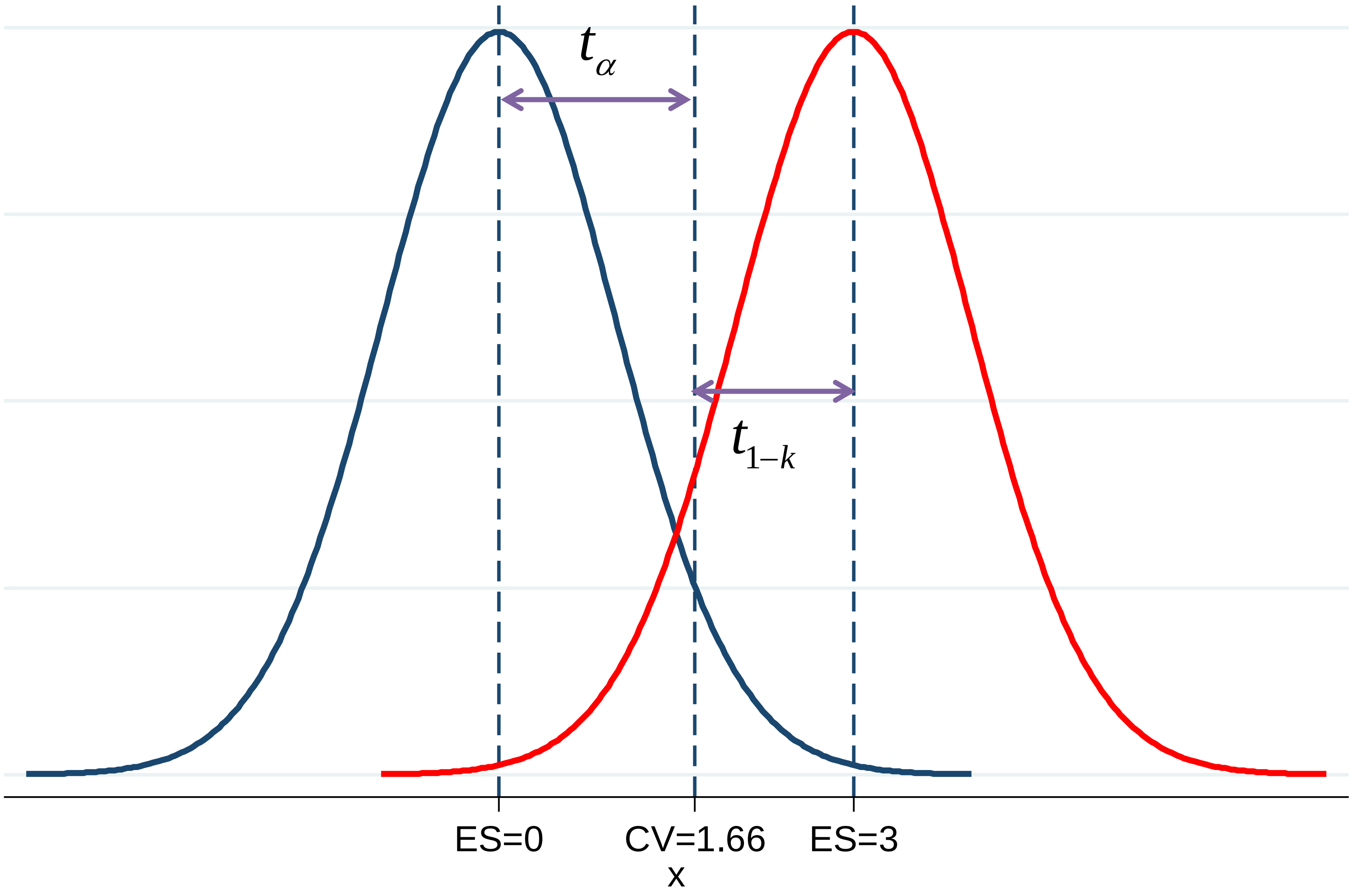
- Let’s consider a simple program with one possible individual level treatment. ATE can be computed by a simple OLS regression:

$$Y_i = \alpha + \beta T + \varepsilon_i$$

- Assume that each individual was randomly sampled from an identical population (observations are i.i.d), the variance of the treatment effect is given by:

$$Var(\beta) = \frac{1}{P(1-P)} \frac{\sigma^2}{N}$$

Power Analysis



- To achieve power k:

$$\beta > (t_{1-k} + t_{\alpha})SE(\beta)$$

- Example: $t_{1-k} = 0.84 \text{ for } k = 80\%$

- Using the graph, we can identify different approaches to power analysis:
  - ***Minimum Detectable Effect***
  - ***Power Calculation***
  - ***Sample Size Determination***

A. Minimum Detectable Effect (MDE)

- We define the **Minimum Detectable Effect** (MDE):

$$MDE(k,\alpha,N,P) = (t_{(1-k)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

- This formula already suggest many interesting features of power analysis:
  - If we increase sample size, what happens with MDE?
  - If we increase power, what happens with the MDE?

B. Sample Size Estimation

- We define N (sample size) in the following way:

$$N(k,\alpha,\beta_E,P)=\left[\frac{\sigma*(t_{(1-k)}+t_\alpha)*\sqrt{\dfrac{1}{P(1-P)}}}{\beta_E}\right]^2$$

- Again, you may analyze what happens with sample size if we change level of power and the effect size

C. Power

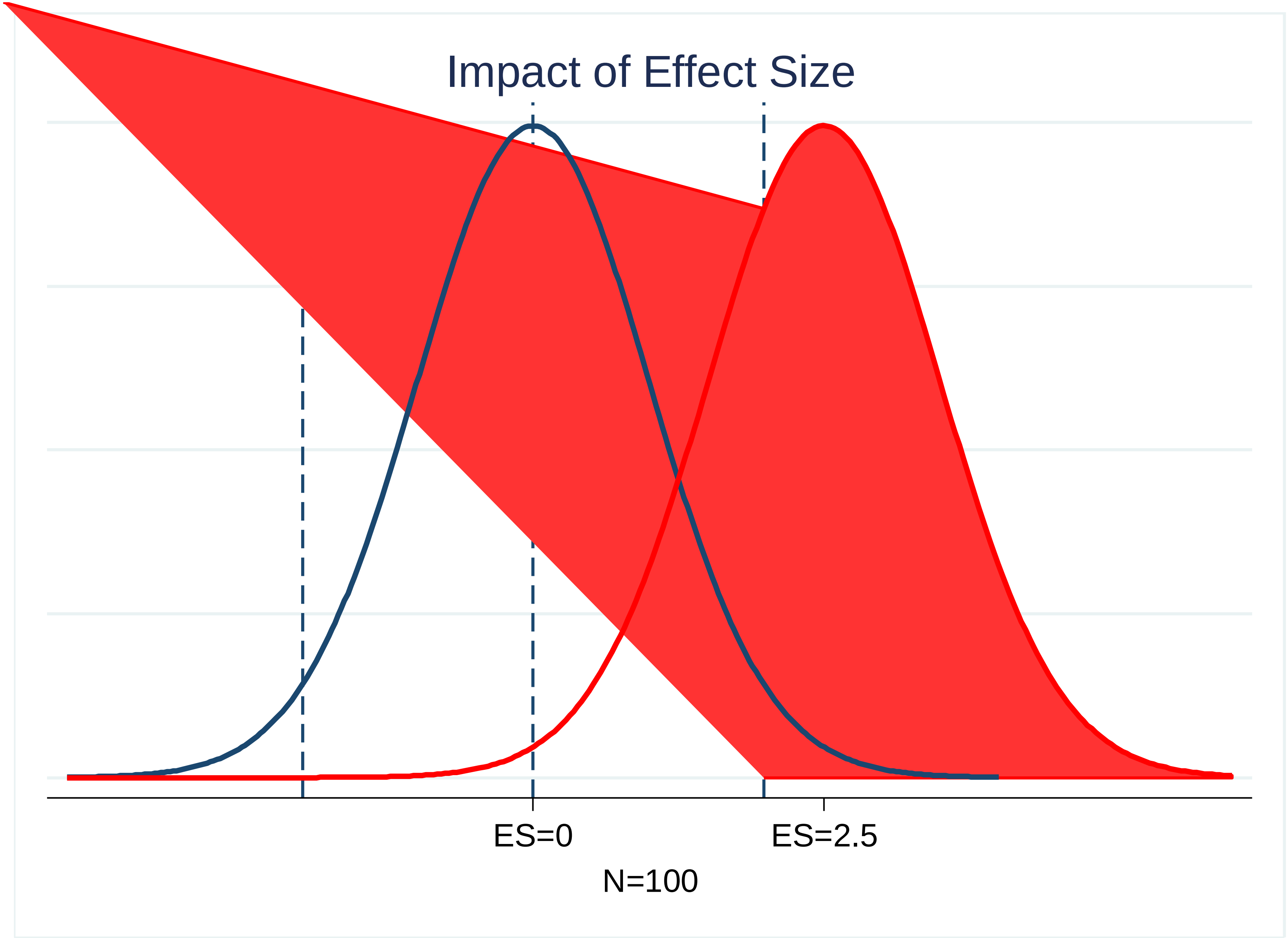
- We define the power level in the following way:

$$t_{(1-k)}(N,\alpha,\beta_E,P)=\frac{\beta_E}{\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}}-\tau_\alpha$$

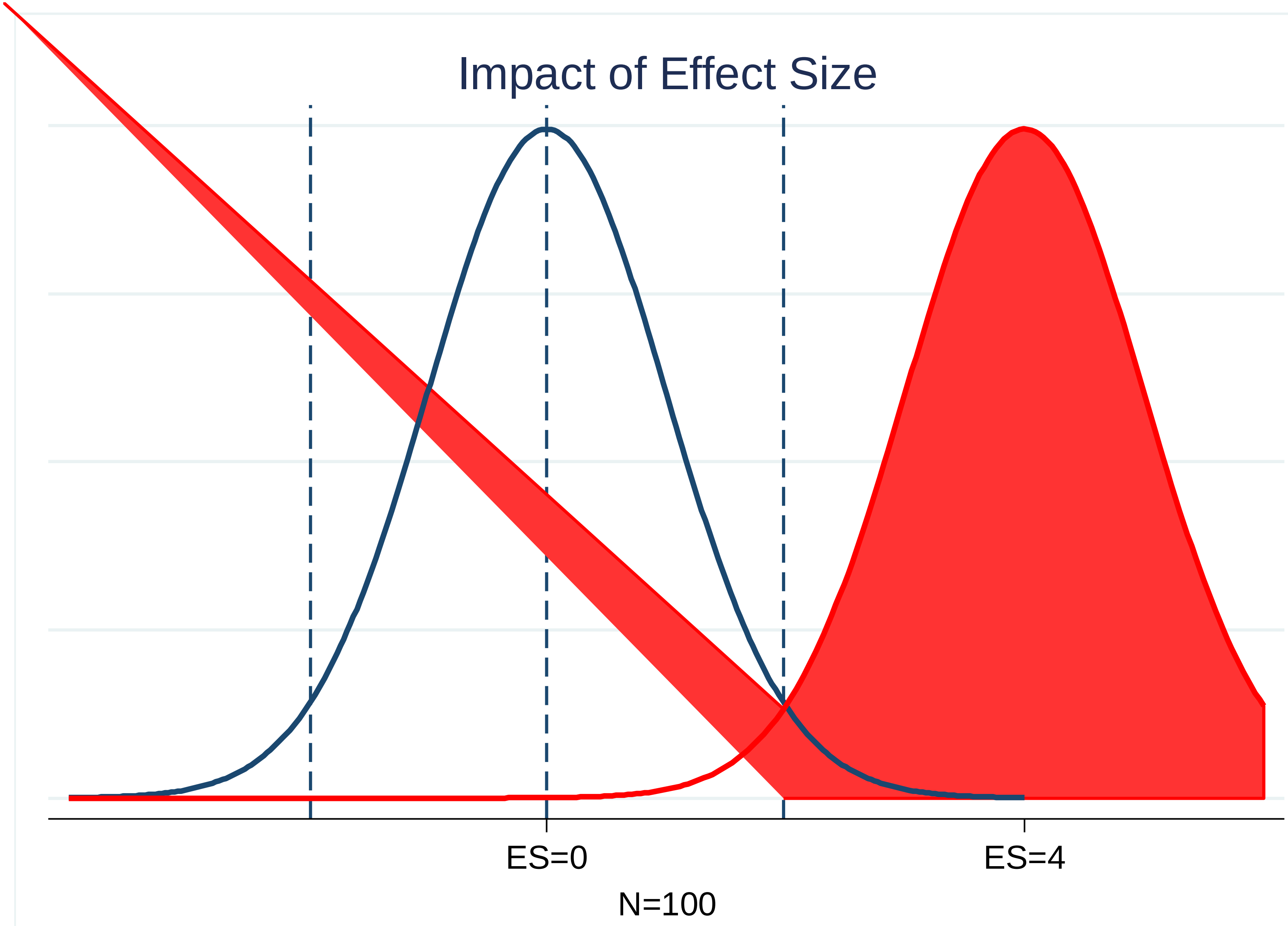
- Again, you may analyze what happens with power if we change sample size and the effect size

What influences Power?

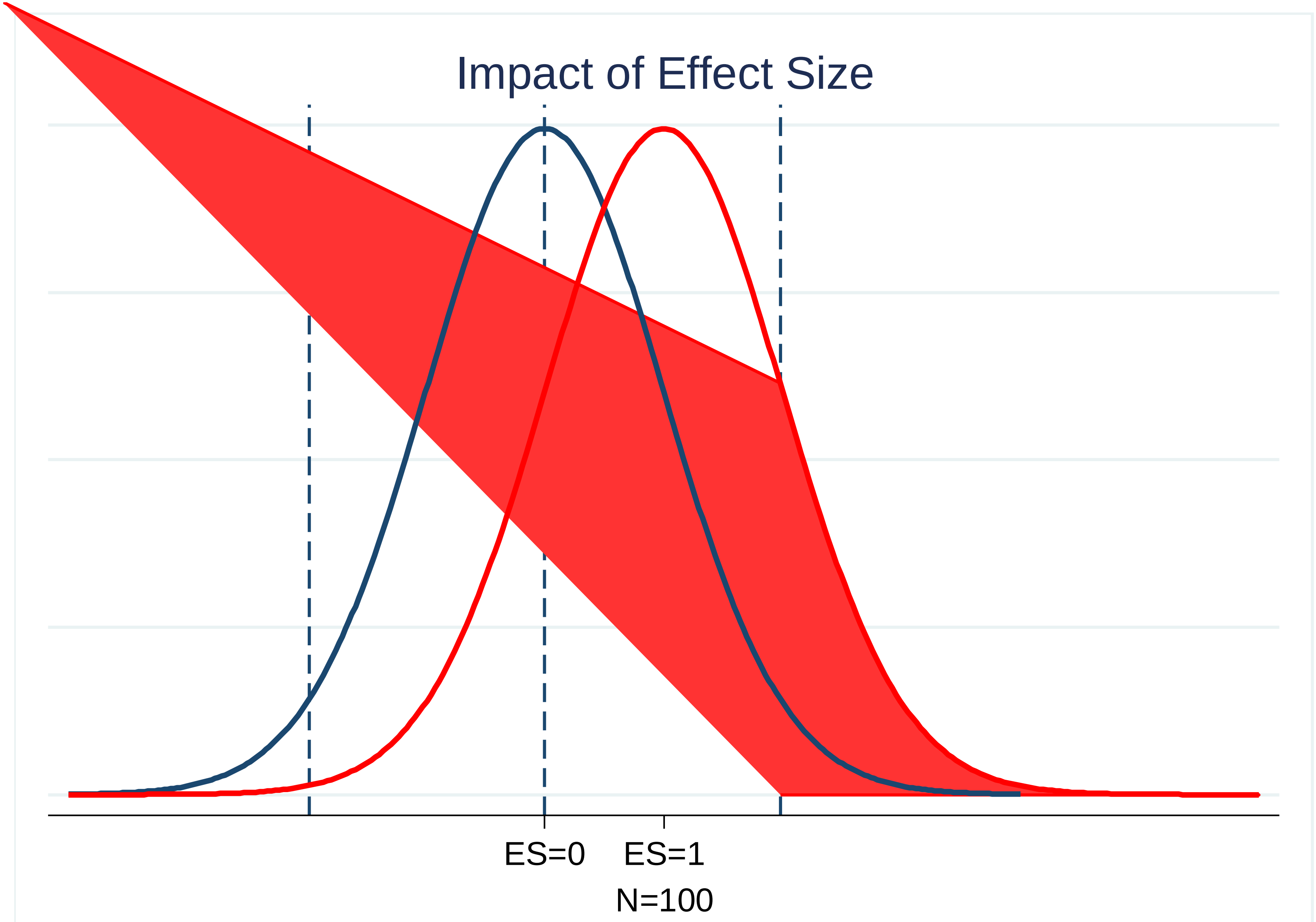
1. Effect Size
2. Sample Size
3. Variance
4. One-sided versus two-sided test
5. Proportion of sample in Treatment vs. Control groups

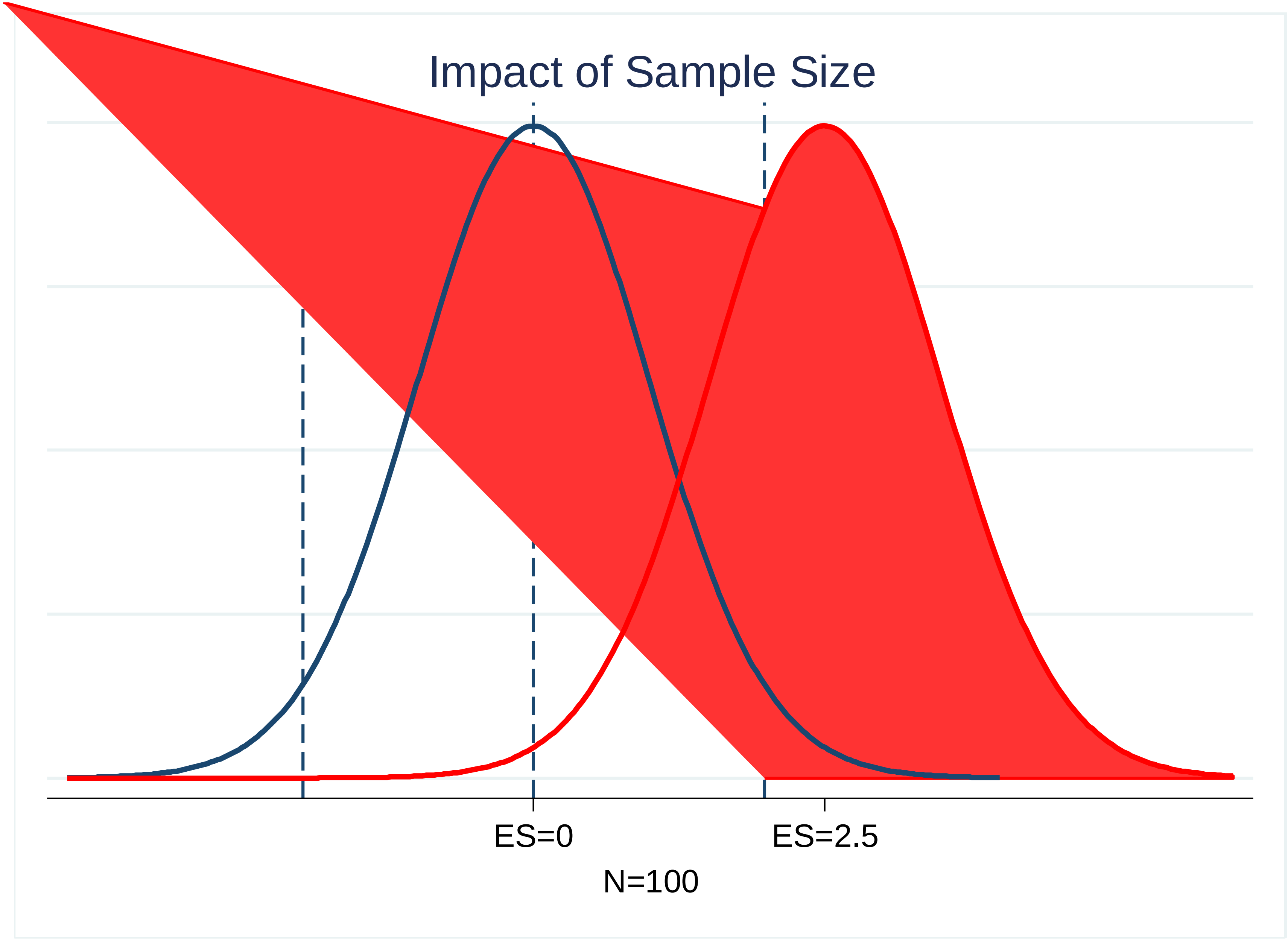

Impact of Effect Size
ES=0
ES=2.5
N=100

Impact of Effect Size

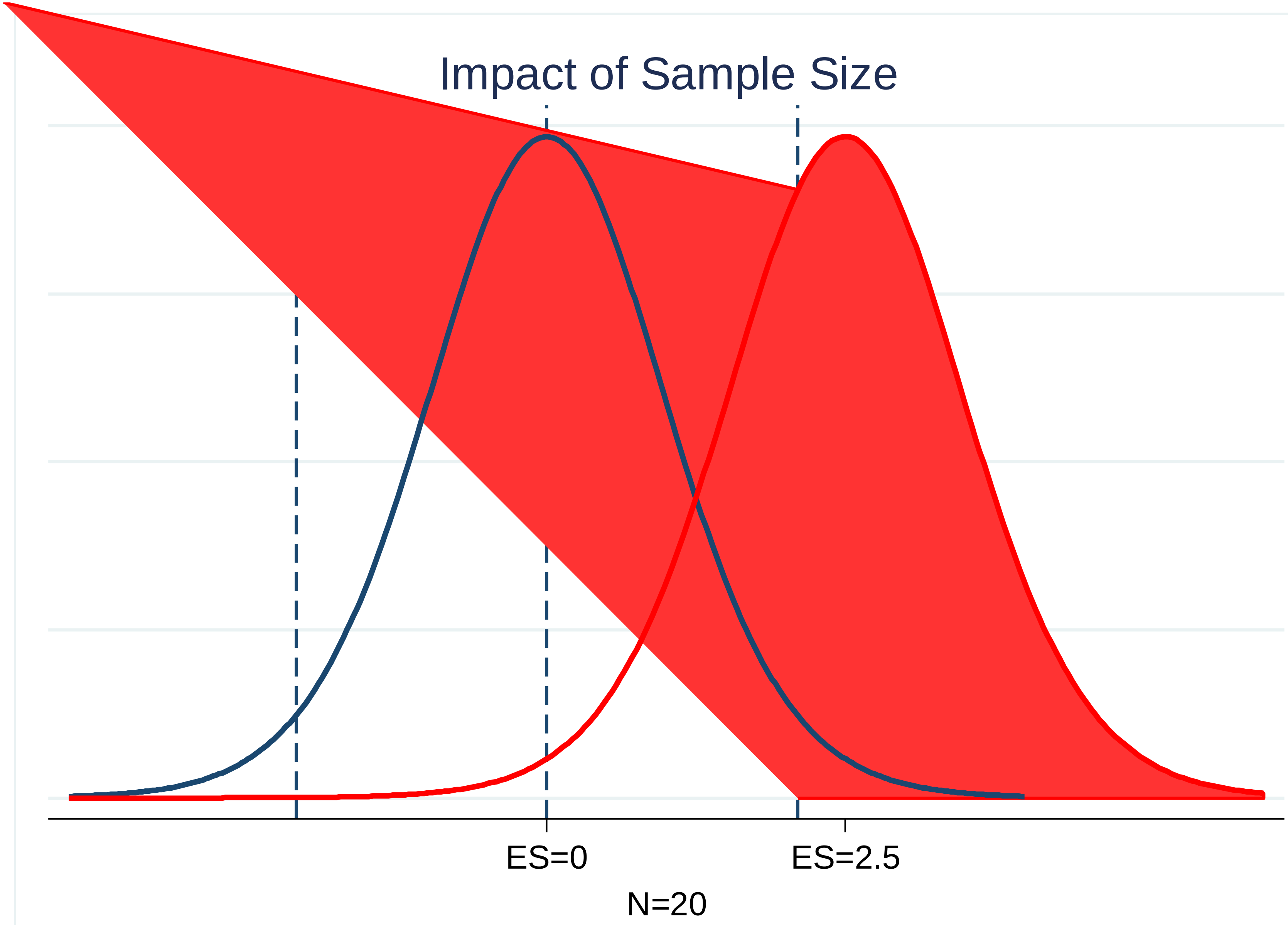


Impact of Effect Size



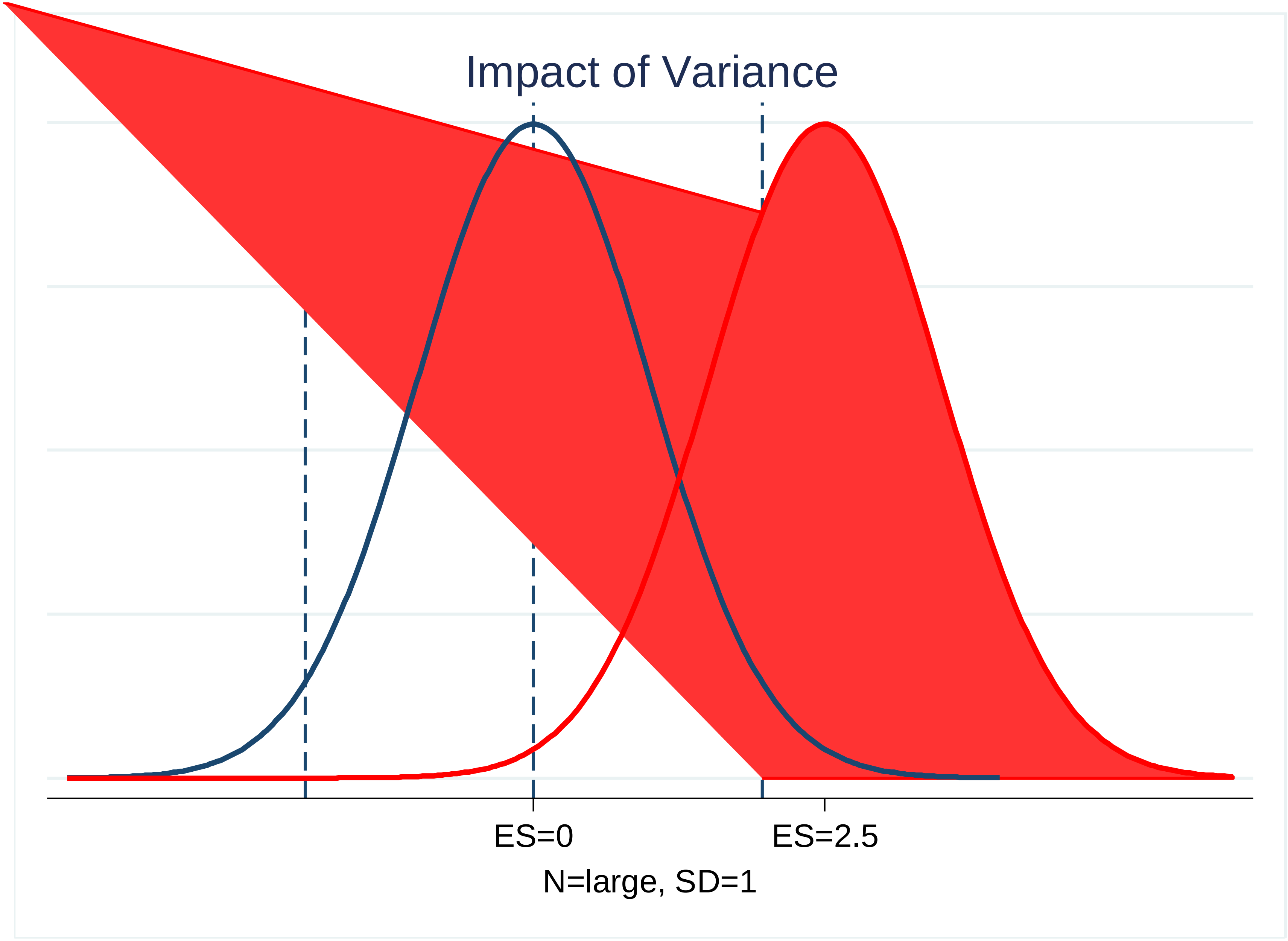

Impact of Sample Size
ES=0
ES=2.5
N=100

Impact of Sample Size

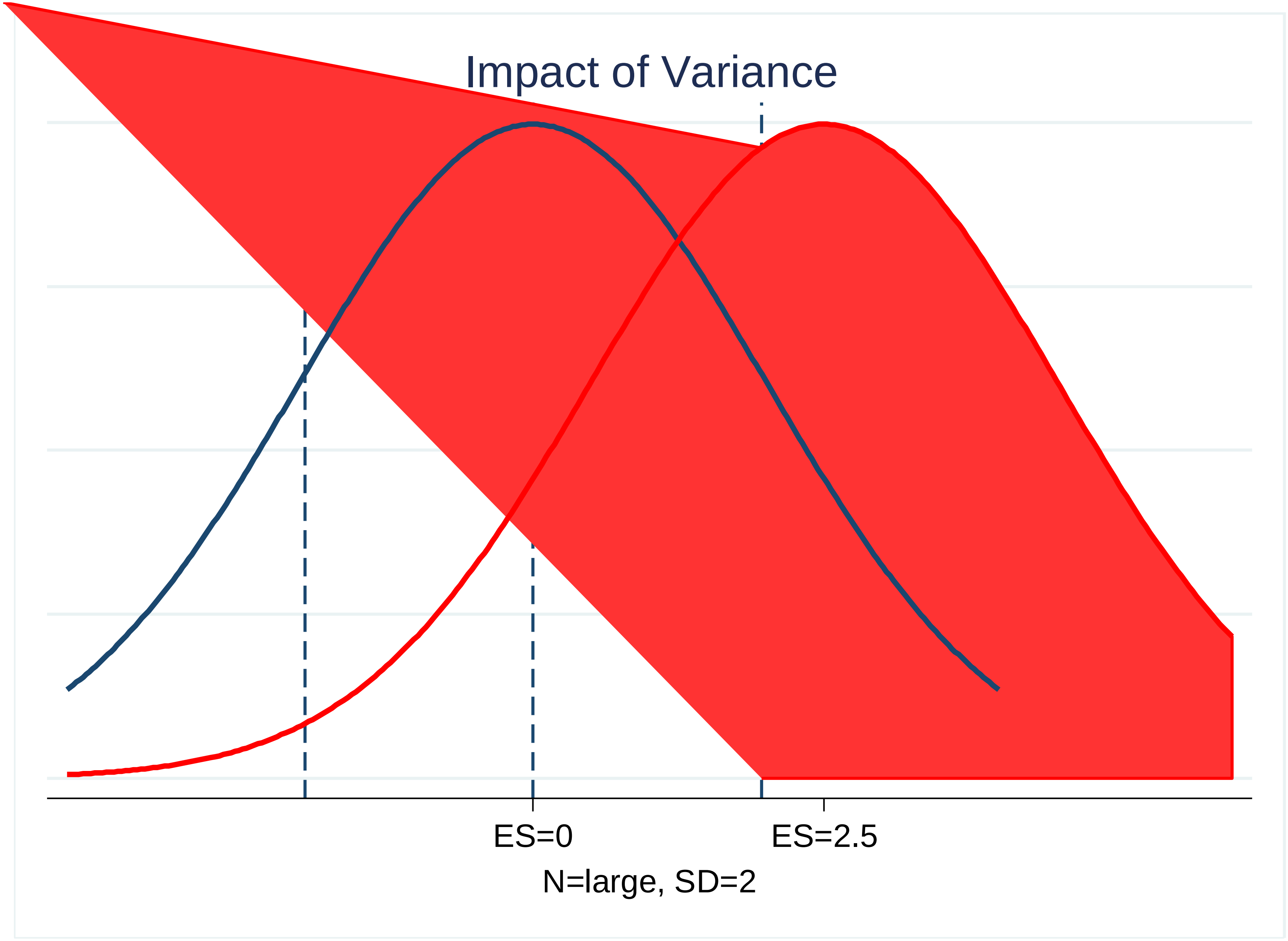


Variance

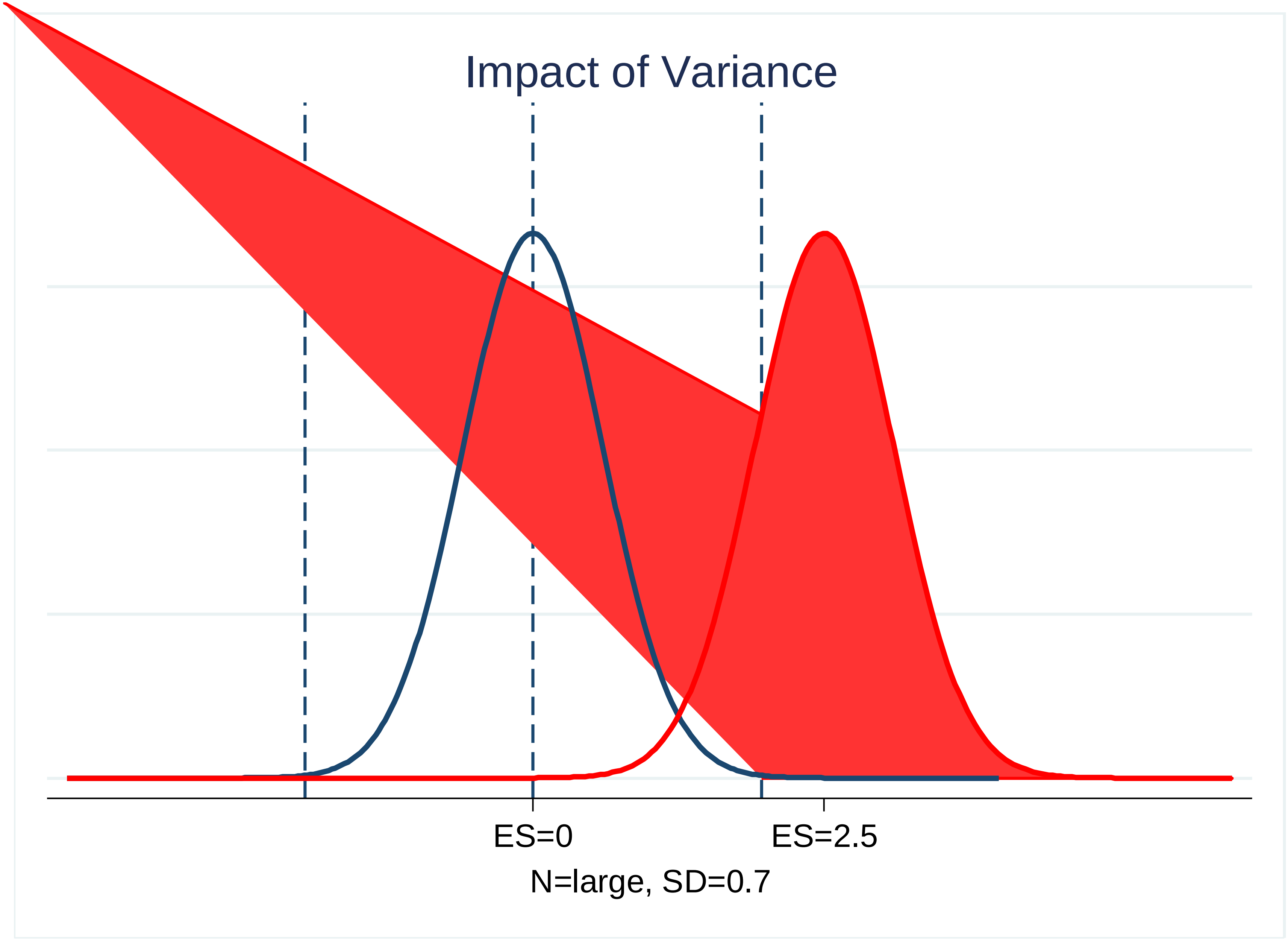
- There is sometimes very little we can do to reduce the noise
- The underlying variance is what it is
- We can try to “absorb” variance:
  - using a baseline
  - controlling for other variables
    - In practice, controlling for other variables (besides the baseline outcome) buys you very little


Impact of Variance
ES=0
ES=2.5
N=large, SD=1

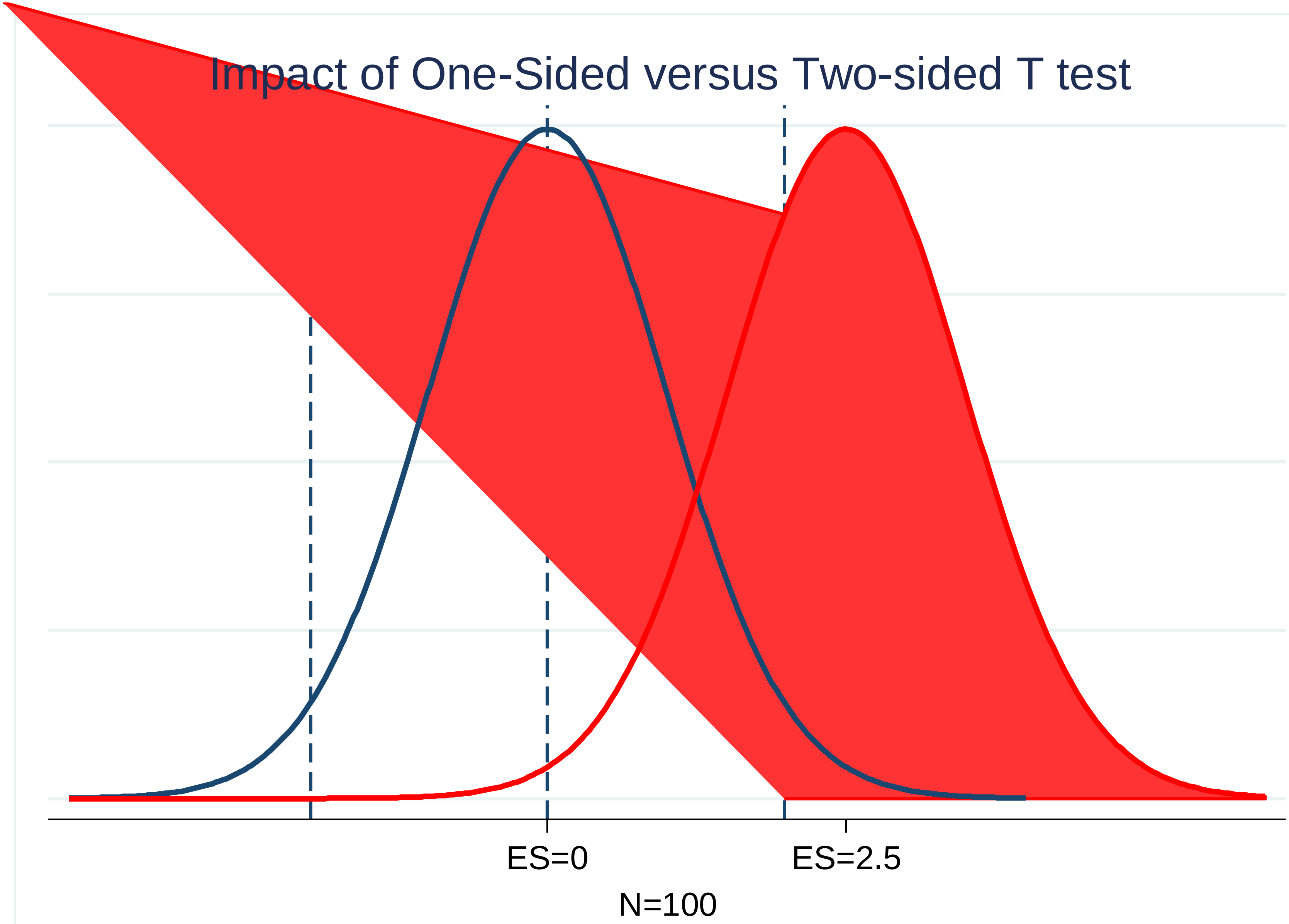
Impact of Variance



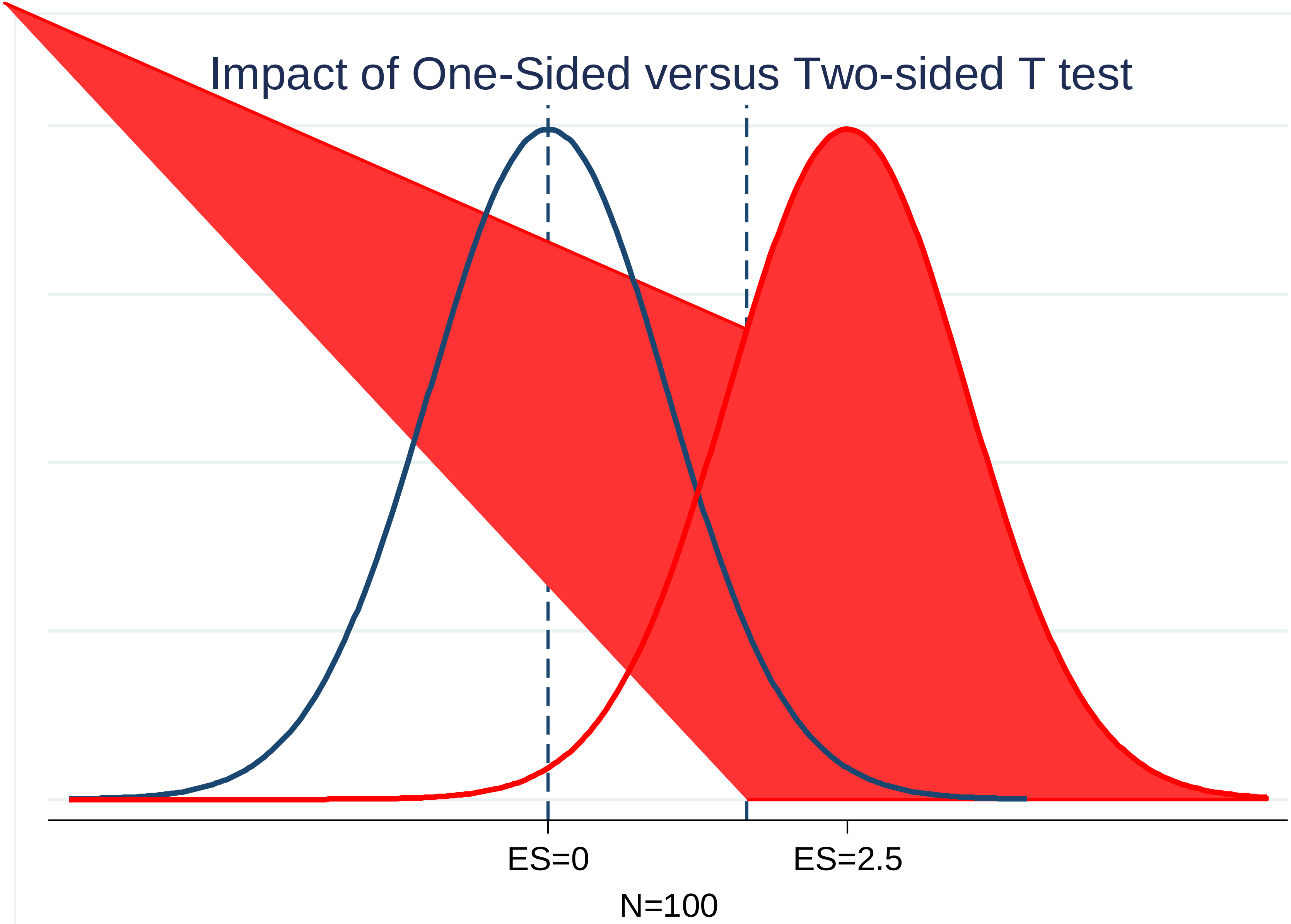
Impact of Variance



Impact of One-Sided versus Two-sided T test



Impact of One-Sided versus Two-sided T test



Allocation to Treatment versus Control

$$sd(X_1 - X_2) = \sqrt{\frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}}$$

$$sd(X_1 - X_2) = \sqrt{\frac{1}{2} + \frac{1}{2}} = \sqrt{\frac{2}{2}} = 1$$

$$sd(X_1 - X_2) = \sqrt{\frac{1}{3} + \frac{1}{1}} = \sqrt{\frac{4}{3}} = 1.15$$

2.3 Standardized Effect Size

- Recall the definition of **Minimum Detectable Effect** (MDE):

$$MDE(k,\alpha,N,P)=(t_{(1-k)}+t_{\alpha})*\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\sigma^2}{N}}$$

- An standardized version can be formulated if you normalized it in terms of standard deviations:

$$\beta=Y_T-Y_C\Rightarrow\delta=\frac{\beta}{\sigma}=\frac{Y_T-Y_C}{\sigma}$$

- Then:

$$SMDE(k,\alpha,N,P)=\frac{MDE(k,\alpha,N,P)}{\sigma}$$

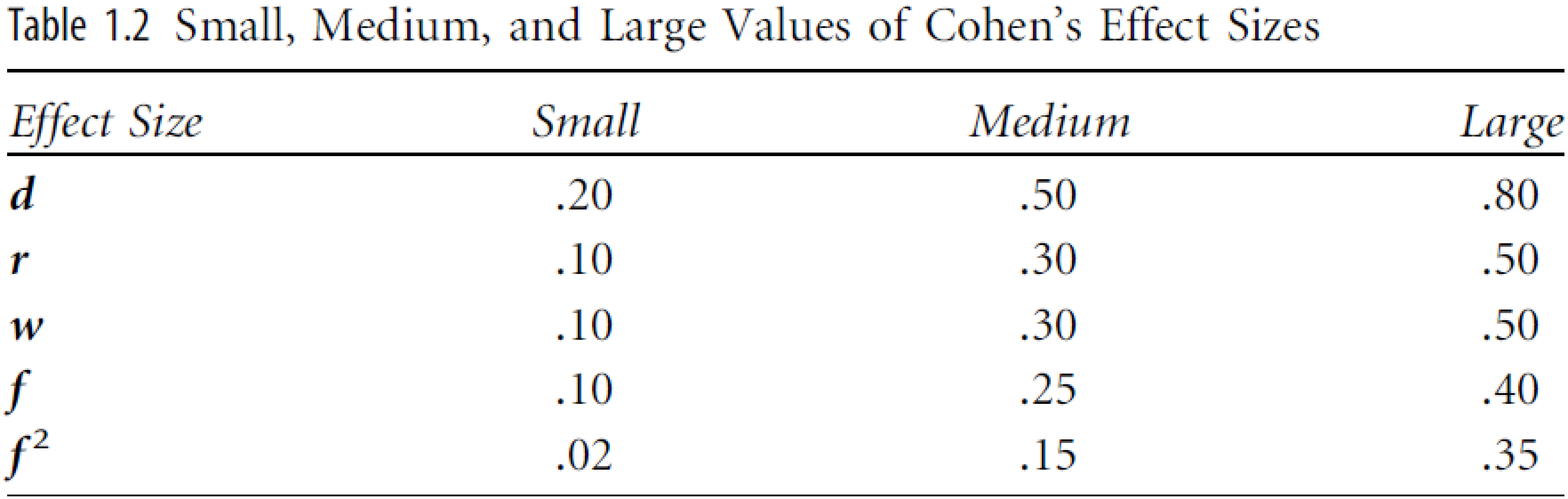


Table 1.2 Small, Medium, and Large Values of Cohen's Effect Sizes

Effect Size	*Small*	*Medium*	*Large*
d	.20	.50	.80
r	.10	.30	.50
w	.10	.30	.50
f	.10	.25	.40
f^2	.02	.15	.35

2.4 Using control variables

- Consider the following OLS regression:

$$Y_i = \alpha + \beta T + X_i \gamma + e_i$$

- Then the treatment effect can be written in the following way:

$$\beta = Y_T - Y_C - \gamma(\overline{X}_T - \overline{X}_C)$$

- Then,

$$Var(\beta) = \frac{1}{P(1-P)} \frac{\sigma^2(1-R_X^2)}{N}$$

- Recall the following result:

$$R_X^2 = 1 - \frac{\sigma_{|X}^2}{\sigma^2}$$

- Adding covariates reduces the residual variance and thereby tends to reduce the variance of parameter estimates (Duflo et al 2008)

- The **MDE with covariates** can be written as follows:

$$MDE(k,\alpha,N,P) = (t_{(1-k)} + t_\alpha) * \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2(1-R_X^2)}{N}}$$

2.5 Accounting for Imperfect Compliance

- Imperfect compliance in an issue in impact evaluation with implications for power calculation

- Let's define the following groups:

  $c:$ the fraction of those initially assigned to the treatment who were actually treated

  $s:$ the share of subjects assigned to control group who receive the treatment

- Then, the **MDE accounting for imperfect compliance**:

$$MDE(k,\alpha,N,P) = (t_{(1-k)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2}{N}} \frac{1}{c-s}$$

2.6 Power Calculations in Practice

- Power calculations implies a substantial guess work in practice:
  1. Data on mean and variance of outcome is required. Previously collected data is helpful
  2. Need to chose the level of the test
  3. Effect size must be specified (using SD can simplify this)
  4. Level of power (80% to 90% as standard)

Power Analysis for Complex Randomized Designs

3. Power Analysis for Complex Randomized Designs

- Individual-level randomized designs are less common in practice and may not be the most efficient/realistic way to allocate the treatment:
  - Allocation to treatment may follow a hierarchical structure (schools rather than individuals)
  - Researchers want to reduce sampling variability
- Two most common approaches:
  - Cluster Randomized Designs
  - Block Randomized Designs

3.1 Cluster Randomized Designs

- Programs randomized at group level are common (OPORTUNIDADES, Familias en Accion, etc)
- Error term may not be independent across individuals in a given group (common shocks to all individuals in a treated area)
- Formally:

$$Y_{ij} = \alpha + \beta T + \upsilon_j + \omega_{ij}$$

- Standard error:

$$SE(\beta) = \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{n\tau^2 + \sigma^2}{nJ}}$$

- If randomization had been conducted at the level of the individual:

$$SE(\beta) = \sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\tau^2+\sigma^2}{nJ}}$$

- This implies that the ratio between SE (known as **Design Effect**) is:

$$D = \frac{SE(\beta)_{cluster}}{SE(\beta)_{individual}} = \frac{\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{n\tau^2+\sigma^2}{nJ}}}{\sqrt{\frac{1}{P(1-P)}}\sqrt{\frac{\tau^2+\sigma^2}{nJ}}} = \sqrt{1+(n-1)\rho}$$

- Notice that D is increasing in ICC and n, leading to an increase in variance respect to individual randomization
- Key result: sample size for clustered randomized design can be obtained by multiplying the design effect with the sample size computed under individual randomized design:

$$MDE_{cluster} = MDE_{individual} * D$$

$$= (t_{(1-k)} + t_{\alpha}) * \sqrt{\frac{1}{P(1-P)}} \sqrt{\frac{\sigma^2}{nJ}} * \sqrt{1 + (n-1)\rho}$$

- Where: $\rho = \frac{\tau^2}{\tau^2 + \sigma^2}$
- Bloom (2005) shows that the MDE with J groups of size n each is given by:

$$MDE = \frac{M_{J-2}}{\sqrt{P(1-P)J}} \sqrt{\rho + \frac{1-\rho}{n}\sigma}$$

$$M_{J-2} = t_{\alpha/2} + t_{1-k}$$

- Three implications:
  - MDE varies proportionally to J
  - n affects precision much less, especially for large ICC
  - J and n depend critically on ICC

ICC in Educational Programs

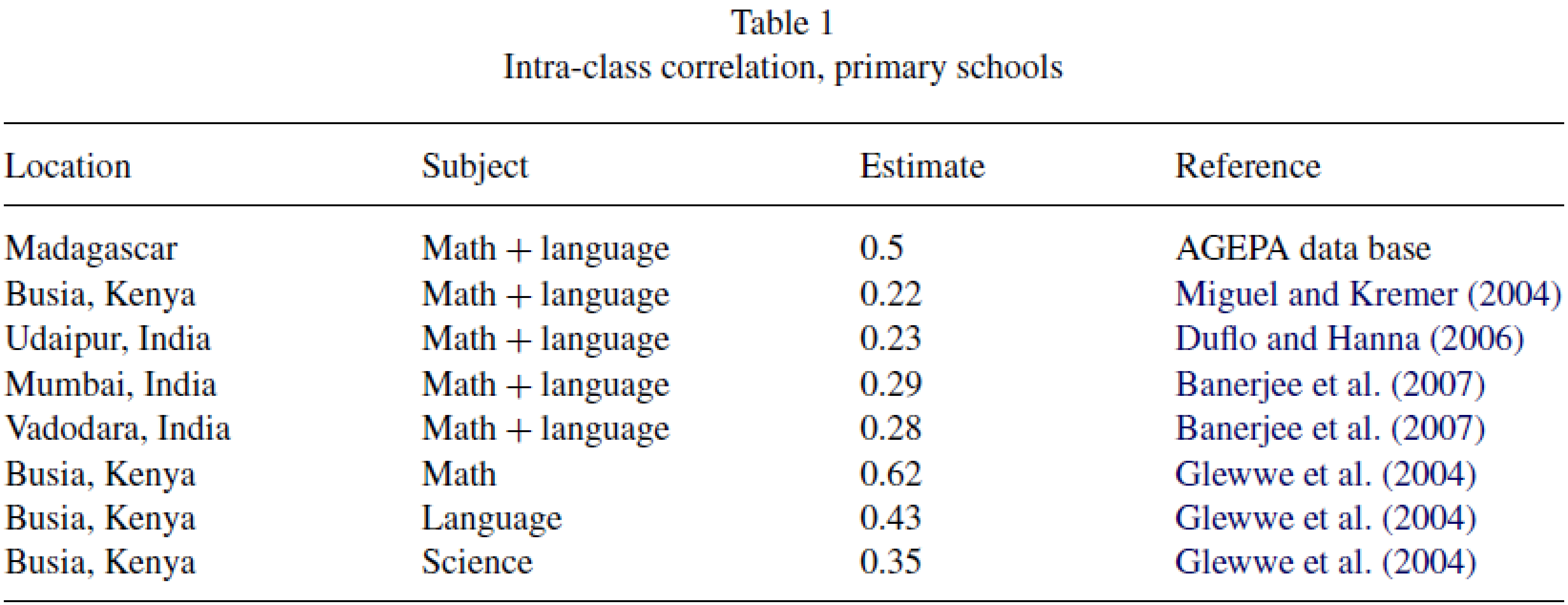


Table 1
Intra-class correlation, primary schools

Location	Subject	Estimate	Reference
Madagascar	Math + language	0.5	AGEPA data base
Busia, Kenya	Math + language	0.22	Miguel and Kremer (2004)
Udaipur, India	Math + language	0.23	Duflo and Hanna (2006)
Mumbai, India	Math + language	0.29	Banerjee et al. (2007)
Vadodara, India	Math + language	0.28	Banerjee et al. (2007)
Busia, Kenya	Math	0.62	Glewwe et al. (2004)
Busia, Kenya	Language	0.43	Glewwe et al. (2004)
Busia, Kenya	Science	0.35	Glewwe et al. (2004)

Duflo et al (2008)

Power surveys show that most research lacks of adequate sample sizes to detect causal effects

- Low power is common in many sciences:
  - Psychology (Cohen 1962, Sedlmeier et al 1989)
  - Management (Cashen et al 2004)
  - Behavioral ecology (Jennions et al 2003)
  - Psychiatry (Brown and Hale 1992)
  - Biology (Thomas and Juanes 1996)
  - Education (West 1985)
- What about economics?

Low power in business training programs (McKenzie et al 2012)

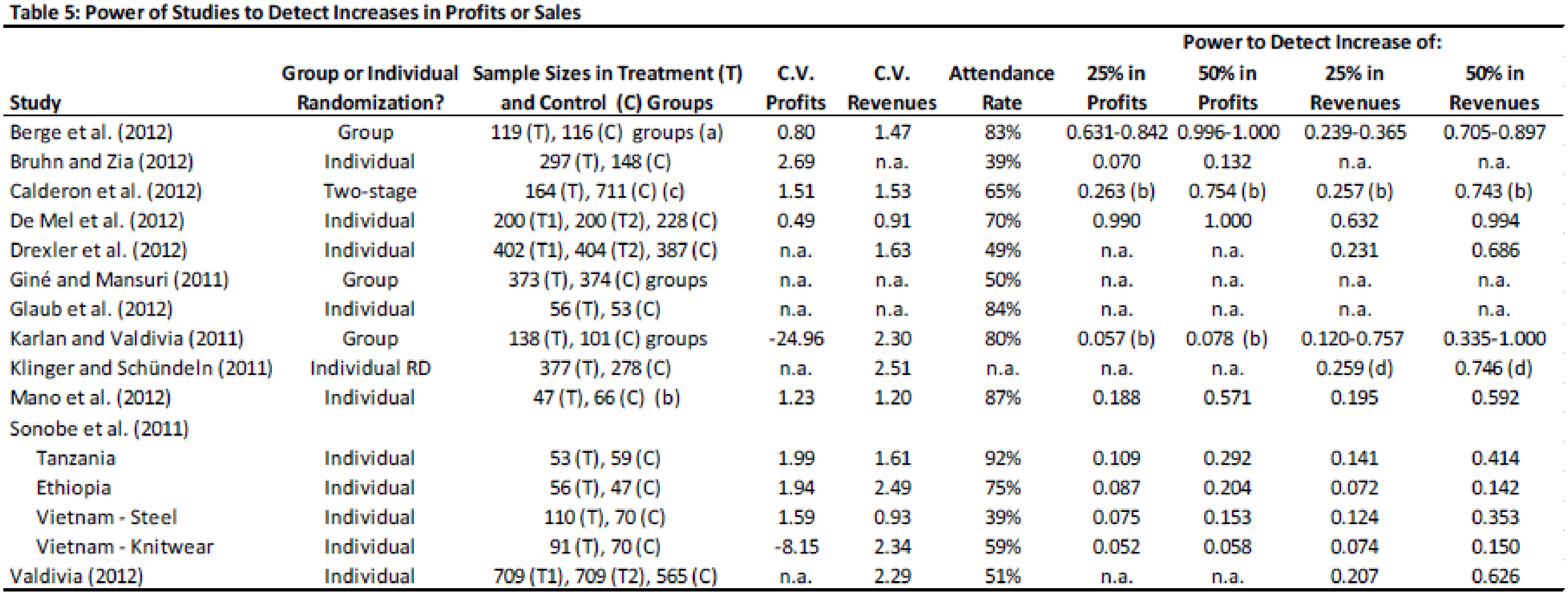
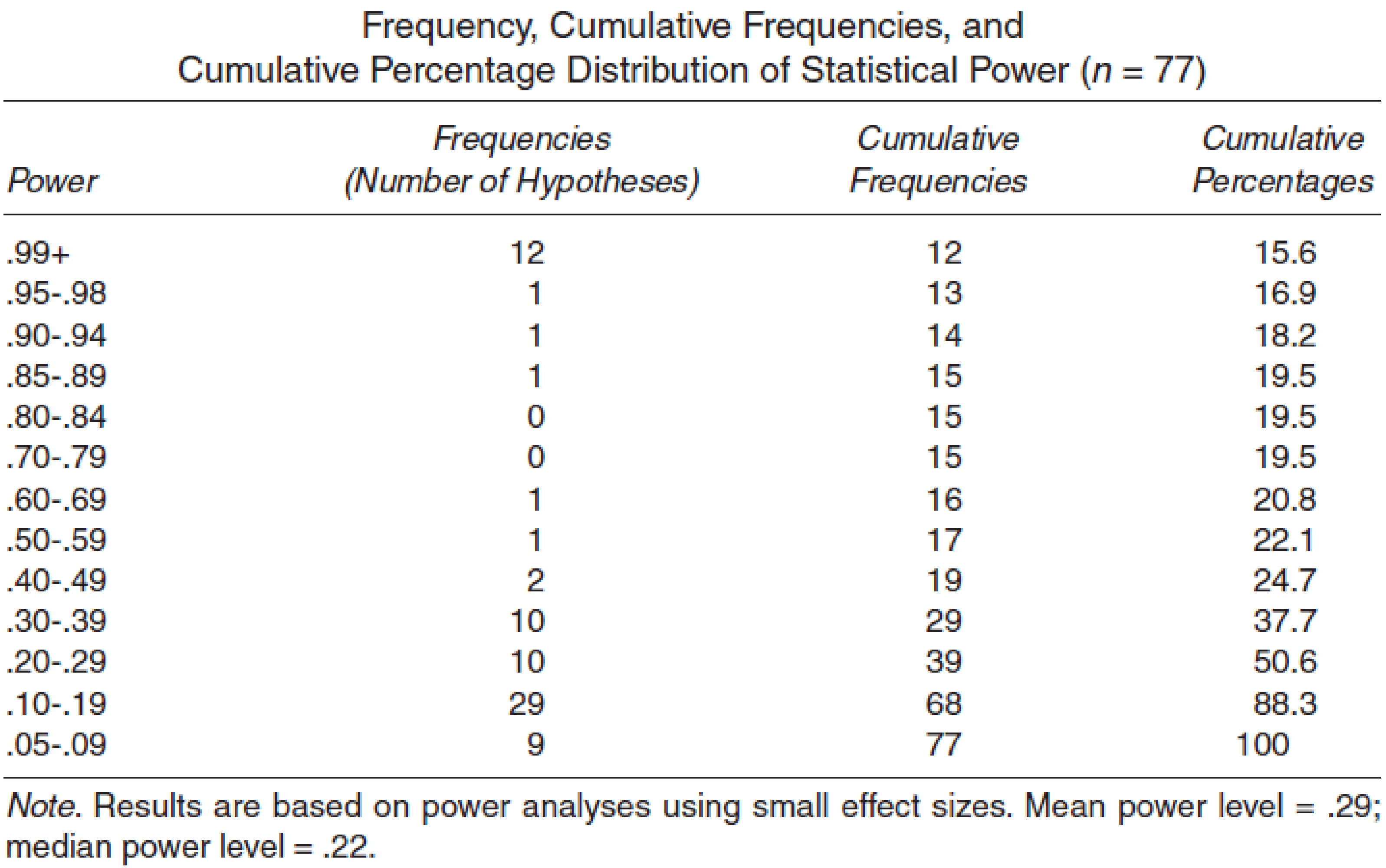


Table 5: Power of Studies to Detect Increases in Profits or Sales

						Power to Detect Increase of:			
Study	Group or Individual Randomization?	Sample Sizes in Treatment (T) and Control (C) Groups	C.V. Profits	C.V. Revenues	Attendance Rate	25% in Profits	50% in Profits	25% in Revenues	50% in Revenues
Berge et al. (2012)	Group	119 (T), 116 (C) groups (a)	0.80	1.47	83%	0.631-0.842	0.996-1.000	0.239-0.365	0.705-0.897
Bruhn and Zia (2012)	Individual	297 (T), 148 (C)	2.69	n.a.	39%	0.070	0.132	n.a.	n.a.
Calderon et al. (2012)	Two-stage	164 (T), 711 (C) (c)	1.51	1.53	65%	0.263 (b)	0.754 (b)	0.257 (b)	0.743 (b)
De Mel et al. (2012)	Individual	200 (T1), 200 (T2), 228 (C)	0.49	0.91	70%	0.990	1.000	0.632	0.994
Drexler et al. (2012)	Individual	402 (T1), 404 (T2), 387 (C)	n.a.	1.63	49%	n.a.	n.a.	0.231	0.686
Giné and Mansuri (2011)	Group	373 (T), 374 (C) groups	n.a.	n.a.	50%	n.a.	n.a.	n.a.	n.a.
Glaub et al. (2012)	Individual	56 (T), 53 (C)	n.a.	n.a.	84%	n.a.	n.a.	n.a.	n.a.
Karlan and Valdivia (2011)	Group	138 (T), 101 (C) groups	-24.96	2.30	80%	0.057 (b)	0.078 (b)	0.120-0.757	0.335-1.000
Klinger and Schündeln (2011)	Individual RD	377 (T), 278 (C)	n.a.	2.51	n.a.	n.a.	n.a.	0.259 (d)	0.746 (d)
Mano et al. (2012)	Individual	47 (T), 66 (C) (b)	1.23	1.20	87%	0.188	0.571	0.195	0.592
Sonobe et al. (2011)									
Tanzania	Individual	53 (T), 59 (C)	1.99	1.61	92%	0.109	0.292	0.141	0.414
Ethiopia	Individual	56 (T), 47 (C)	1.94	2.49	75%	0.087	0.204	0.072	0.142
Vietnam - Steel	Individual	110 (T), 70 (C)	1.59	0.93	39%	0.075	0.153	0.124	0.353
Vietnam - Knitwear	Individual	91 (T), 70 (C)	-8.15	2.34	59%	0.052	0.058	0.074	0.150
Valdivia (2012)	Individual	709 (T1), 709 (T2), 565 (C)	n.a.	2.29	51%	n.a.	n.a.	0.207	0.626

Low power in management (Cashen et al 2004)



Frequency, Cumulative Frequencies, and
Cumulative Percentage Distribution of Statistical Power (n = 77)

Power	*Frequencies (Number of Hypotheses)*	*Cumulative Frequencies*	*Cumulative Percentages*
.99+	12	12	15.6
.95-.98	1	13	16.9
.90-.94	1	14	18.2
.85-.89	1	15	19.5
.80-.84	0	15	19.5
.70-.79	0	15	19.5
.60-.69	1	16	20.8
.50-.59	1	17	22.1
.40-.49	2	19	24.7
.30-.39	10	29	37.7
.20-.29	10	39	50.6
.10-.19	29	68	88.3
.05-.09	9	77	100

Note. Results are based on power analyses using small effect sizes. Mean power level = .29; median power level = .22.