Experimetrics and Power Analysis
實驗計量與統計檢定力分析

Joseph Tao-yi Wang (王道一)
EE-BGT, Lecture 1b (Experimetrics Module 1)

The Replication Size Trinity

1. Sample Size n : # of observations/subjects
2. Effect Size d : How big is the true result
3. Power $(1\text{-}\beta)$: How likely will your test show significance if there is truly an effect

Why Do We Care About This?

- Editor's Preface (JEEA 2015):
  - "A necessary (but not sufficient) condition for publishing a replication study or null result will be
  - the presentation of power calculations."
- Test Resolution: $\Pr(\text{confirm} \mid \text{infected patient})$
  - Discharge of COVID requires 3 consecutive negatives (三採陰)
  - Because even PCR has insufficient power (around 70%)...
- But what about structural estimation?

Key Concepts and Definitions

- Treatment Test:
  - Null ($H_0 : \theta = \theta_0$) Hypothesis - No Effect!
  - Alternative ($H_1 : \theta = \theta_1$) Hypothesis - Effective!
- Effect Size ($\theta_1 - \theta_0$): True size of effect
- Alternative Hypothesis can be Directional:
  1. One-sided Alternative - One-tailed test
     - Usually comes from prior beliefs based on theory
  2. Two-sided Alternative - Two-tailed test

Key Concepts and Definitions

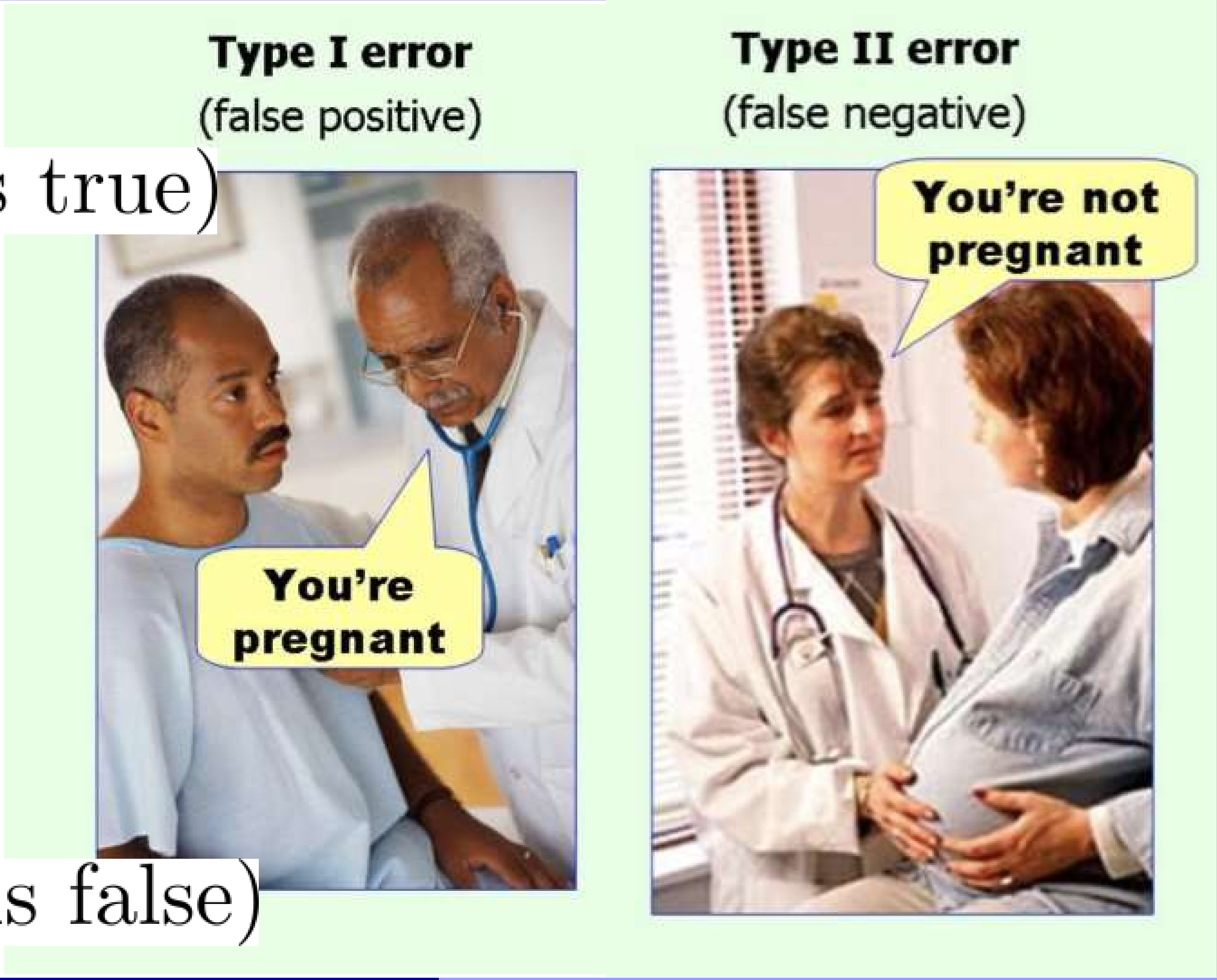
- Two Stages of the Treatment Test:
  1. Compute Test Statistic of sample size n
  2. Compare Test Statistic with null distribution
- Rejection Region = Tail of null distribution
  - of a Size $\alpha = \Pr(\text{reject null} \mid \text{null is true})$
  - Critical Value: Rejection region starting point
- p-value $= \Pr\left(|T| \geq T_{CV} \middle| \text{null is true}\right)$
  - $p < 0.05$ (Evidence) vs. $p < 0.01/0.001$ (Strong/Overwhelming Evidence)

Key Concepts and Definitions

- Type 1 Error:
  $\alpha = \Pr(\text{reject null} \mid \text{null is true})$

- But what is Power?

- Type 2 Error:
  $\beta = \Pr(\text{accept null} \mid \text{null is false})$



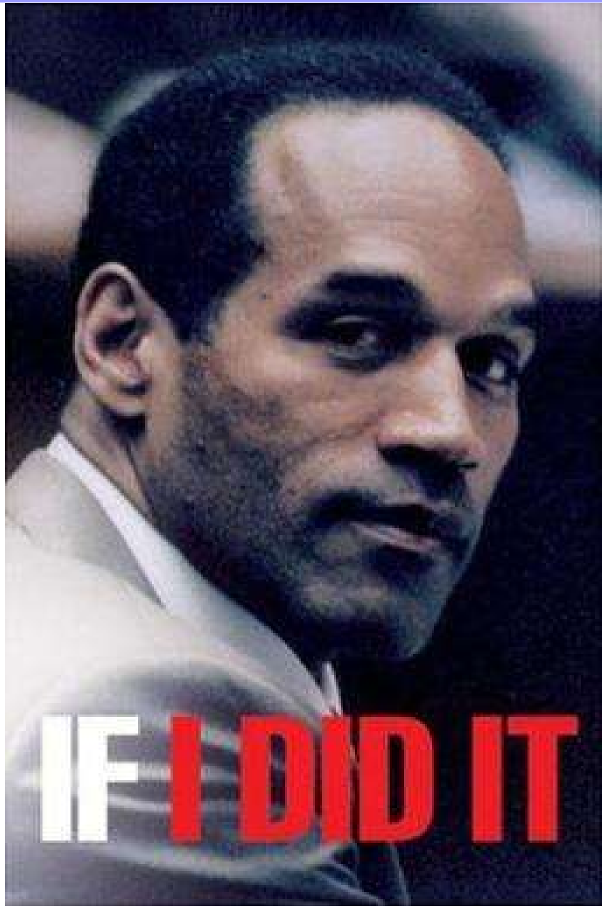
Key Concepts and Definitions

- Type 1 Error: $\alpha = \Pr(\text{reject null} \mid \text{null is true})$
- Type 2 Error: $\beta = \Pr(\text{accept null} \mid \text{null is false})$
- Power(π):$1 - \beta = \Pr(\text{reject null} \mid \text{null is false})$
  1. True effect size $\theta_1 - \theta_0$ (and one/two-tailed)
  2. Sample size n
  3. Size of the test α
- Trade-off: The higher α/n, the higher is π
  1. Power Analysis: Compute power $\pi = 1 - \beta$, or
  2. Find n to meet power requirement $\pi(n) \geq \overline{\pi}$

Choosing the Value of α

- How big can we allow Type 1 Error to be?
- To convict a crime suspect,
  - Null Hypothesis: Not Guilty
  - Alternative Hypothesis: Guilty
  - Type 1: $\alpha = \Pr(\text{convict} \mid \text{innocent suspect})$
  - Type 2: $\beta = \Pr(\text{acquit} \mid \text{guilty suspect})$
- Type 1 Error more serious than Type 2 Error
  - Choose a low α at the expense of power:

$$1 - \beta = \Pr(\text{convict} \mid \text{guilty suspect})$$



Choosing the Value of α

- How big can we allow Type 1 Error to be?
- To test for COVID-19,
  - Null Hypothesis: Healthy
  - Alternative Hypothesis: Infected by COVID-19
  - Type 1: $\alpha = \Pr(\text{confirm} \mid \text{healthy patient})$
  - Type 2: $\beta = \Pr(\text{discharge} \mid \text{infected patient})$
- Type 2 Error more serious than Type 1 Error
  - Choose a higher α to get higher of power:

$$1 - \beta = \Pr(\text{confirm} \mid \text{infected patient})$$

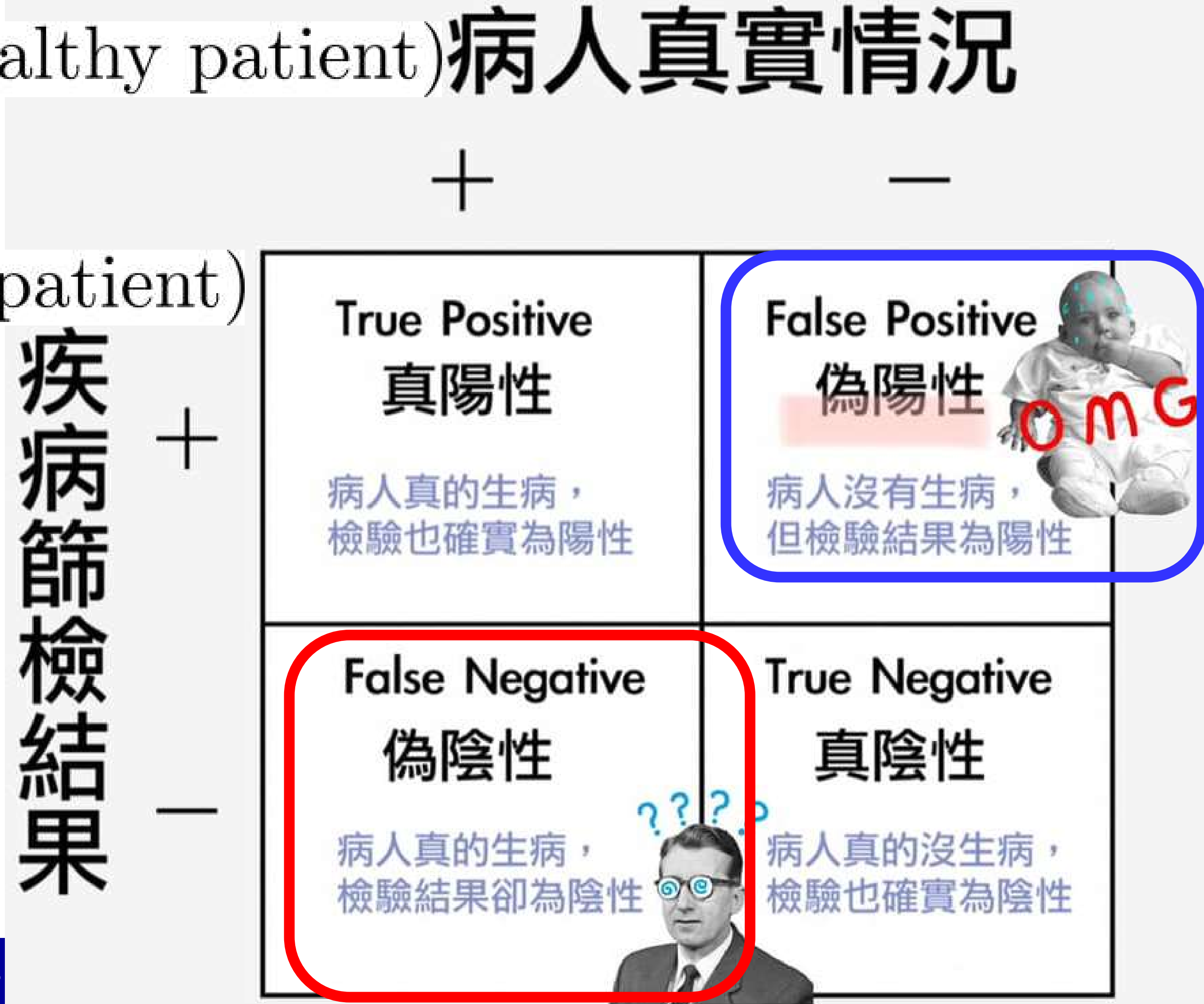
Choosing the Value of α

- Type 1: $\alpha = \Pr(\text{confirm} \mid \text{healthy patient})$
- Type 2:

  $\beta = \Pr(\text{discharge} \mid \text{infected patient})$
- Both errors not fatal in Experimental Economics,
- Convention is: $\alpha = 0.05$

  $\pi = 1 - \beta = 0.80$

  $\beta = 0.20$



Course Material for the Experimetrics Module

- Joseph's Experimetrics Module Website:
  - Peter G. Moffatt (2019): Experimetrics Lecture Notes with Joseph's Notes:
  - Data and Code Package:

    https://homepage.ntu.edu.tw/~josephw/MiniCourseExperimetrics.zip

Treatment Testing Toolkit

- One-sample t-Test
  - Does WTP $=$ £3 ($=$ retail value of coffee mug)?
- Two-sample t-Test (with equal variance)
  - If passes variance ratio test
  - Can be done using OLS!
- Two-sample t-Test (with unequal variance)
  - If fails variance ratio test
- Skewness-kurtosis test

Need CLT (large n)!

But is $n \geq 30$ sufficient?

Treatment Testing Toolkit

- What if we do not have CLT/large n?
  - Use non-parametric tests instead!
- Mann-Whitney Test (aka Ranksum Test)
  - Between-subject non-parametric treatment test
- Kolmogorov-Smirnov (KS) Test
- Epps-Singleton Test (discrete version KS Test)
  - Tests comparing entire distributions

Treatment Testing Toolkit

- What if we have within-subject data?
  - Can use within-subject tests, but watch out for order effect!
- Paired t-Test (assume CLT)
- Wilcoxon Signed Rank Test
  - Within-subject non-parametric treatment test
  - Assume symmetric distribution around median
  - (regarding paired difference). Without it, use:
- Paired-sample Sign Test

Treatment Testing Example: WTP - WTA Gap

- Isoni et al. (AER 2011)
  - Replicate Plott and Zeiler (AER 2007), which in turn
  - Replicate Kahneman et al. (JPE 1990) (KKT)
- Measure WTP and/or WTA
  - Becker–DeGroot–Marschak (BDM) mechanism
  - 2nd price auction against (randomizing) computer
- Treatment Test:
  - Does WTP/WTA $= £3$ ($=$ retail value of the coffee mug)?

Power Analysis: Theory

1. Power Analysis: Find test power $\pi = 1 - \beta$, or
2. Find n to meet power requirement $\pi(n) \geq \overline{\pi}$

- One-sample t-Test (Rarely used in experimental economics)
  - But, Isoni et al. (2011) test WTP of coffee mug $= £3$
- Y: Continuous outcome measure with mean μ
  - Null Hypothesis: $H_0 : \mu = \mu_0$
  - Alternative Hypothesis: $H_1 : \mu = \mu_1 > \mu_0$
- Collect data of sample size n

Power Analysis: Theory

1. What is the power of this test?
2. How big should sample size n be?

$\overline{y} = \text{sample mean}$

$s^2 = \text{sample variance}$

- Test Size $\alpha = 0.05 = \Pr(\text{reject null} \mid \text{null is true})$
- Type 2 Error $\beta = 0.20 = \Pr(\text{accept null} \mid \text{null is false})$
- Power $\pi = 1 - \beta = 0.80$
- One-sample t-test Test Statistic: $t = \dfrac{\overline{y} - \mu_0}{s/\sqrt{n}} \sim t(n-1)$
  - Reject if $t > t_{n-1,\alpha}$ $(t > z_\alpha \text{ for large } n)$

Power Analysis: Power of the Test

$$\pi = \Pr(t > z_\alpha | \mu = \mu_1) = \Pr\left(\frac{\bar{y} - \mu_0}{s/\sqrt{n}} > z_\alpha \middle| \mu = \mu_1\right)$$

$n = 30, \quad \alpha = 0.05$

$\mu_0 = 10$

$\mu_1 = 12$

$z_\alpha = 1.645, \quad s = 5$

$$= \Pr\left(\bar{y} > \mu_0 + z_\alpha(s/\sqrt{n}) \middle| \mu = \mu_1\right)$$

$$= \Pr\left(\frac{\bar{y} - \mu_1}{s/\sqrt{n}} > \frac{\mu_0 + z_\alpha(s/\sqrt{n}) - \mu_1}{s/\sqrt{n}} \middle| \mu = \mu_1\right)$$

$$= \Phi\left(\frac{\mu_1 - \mu_0 - z_\alpha(s/\sqrt{n})}{s/\sqrt{n}}\right) = \Phi\left(\frac{12 - 10 - 1.645(5/\sqrt{30})}{5/\sqrt{30}}\right)$$

$$= \underline{0.71}$$

- What n is required to get $\pi = 0.80$?

Power Analysis: How Big Should n Be?

- Power $\pi = 1 - \beta = \Phi\left(\frac{\mu_1 - \mu_0 - z_\alpha(s/\sqrt{n})}{s/\sqrt{n}}\right)$

$$\Rightarrow z_\beta = \frac{\mu_1 - \mu_0 - z_\alpha(s/\sqrt{n})}{s/\sqrt{n}}$$

$$\Rightarrow z_\beta + z_\alpha = \frac{\mu_1 - \mu_0}{s/\sqrt{n}}$$

$\alpha = 0.05, \quad \beta = 0.20$

$z_\alpha = 1.645, \quad z_\beta = 0.842$

$\mu_0 = 10$

$\mu_1 = 12$

$s = 5$

$$\Rightarrow n = \frac{s^2(z_\alpha + z_\beta)^2}{(\mu_1 - \mu_0)^2} = \frac{5^2(1.645 + 0.842)^2}{(12 - 10)^2} = \underline{38.66}$$

- So we need $n \geq 39$

Power Analysis: Power in STATA

- What is the power for sample size $n = 30$?
  - STATA command for power calculation

μ_0 / μ_1

```
power onemean 10 12 , sd(5) n(30) oneside
```

  - sample std; sample size
  - 1-sample t-test

one-tailed test

Power Analysis: Power Results in STATA

- What is the pow
  - STATA Results:

power onemear

Slightly different since STATA did not use normal approximation...

```
Estimated power for a one-sample mean test
t test
Ho: m = m0  versus  Ha: m > m0

Study parameters:

        alpha =    0.0500
            N =        30
        delta =    0.4000
           m0 =   10.0000
           ma =   12.0000
           sd =    5.0000

Estimated power:

        power =    0.6895
```

Power Analysis: Sample Size in STATA

- What is the sample size to get power $\pi = 0.80$?
  - STATA command for power calculation

μ_0 / μ_1

```
power onemean 10 12 , sd(5) oneside p(0.8)
```

  - sample std; required power
  - 1-sample t-test one-tailed test

Power Analysis: Sample Size Result/Stata

- What is the sam
  - STATA Results:

    power onemea .8)

Slightly larger n since STATA did not use normal approximation...

```
Performing iteration ...

Estimated sample size for a one-sample mean test
t test
Ho: m = m0  versus  Ha: m > m0

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    0.4000
           m0 =   10.0000
           ma =   12.0000
           sd =    5.0000

Estimated sample size:

            N =        41
```

Power Analysis: Graph Power in STATA

- Plot power against sample size with `graph`
  - STATA command for power calculation

$\mu_0/\mu_1 \in [10.5, 12.5]$

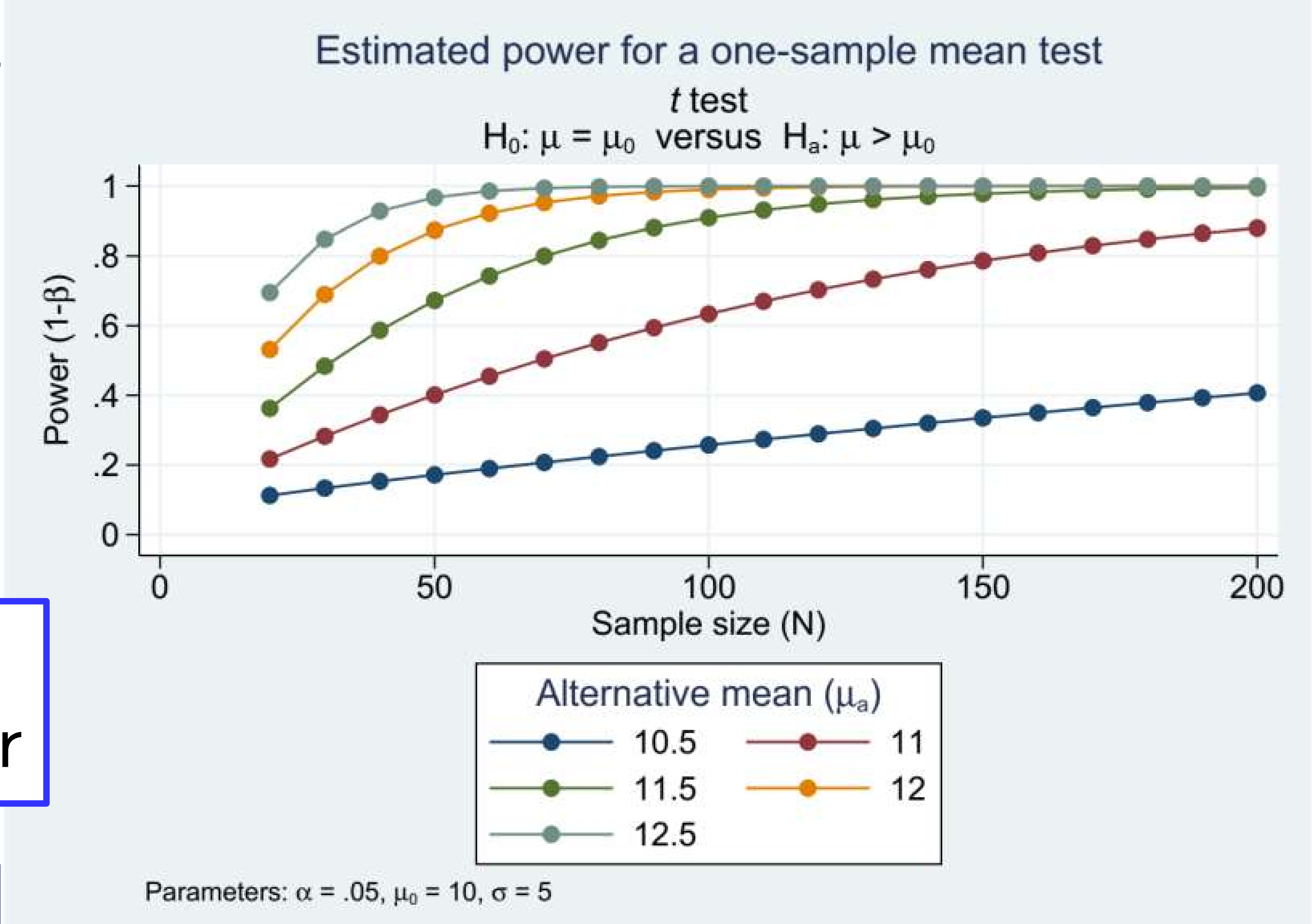
```
power onemean 10 (10.5(0.5)12.5), sd(5) n(20(10)200) oneside graph
```

  - sample std; n=20-200
  - 1-sample t-test
    one-tailed test

Power Analysis: Graph Power in STATA

- Plot power agair
  - STATA Results:

```
power onemean 10 (10.5
```



Larger effect size yields higher power

Power Analysis: Graph Sample Size/Stata

- Plot sample size against effect size
  - STATA command for power calculation

$\mu_0 / \mu_1 \in [10.5, 12.5]$

```
power onemean 10 (10.5(0.25)12.5), sd(5) p(0.6(0.1)0.9) oneside graph
```

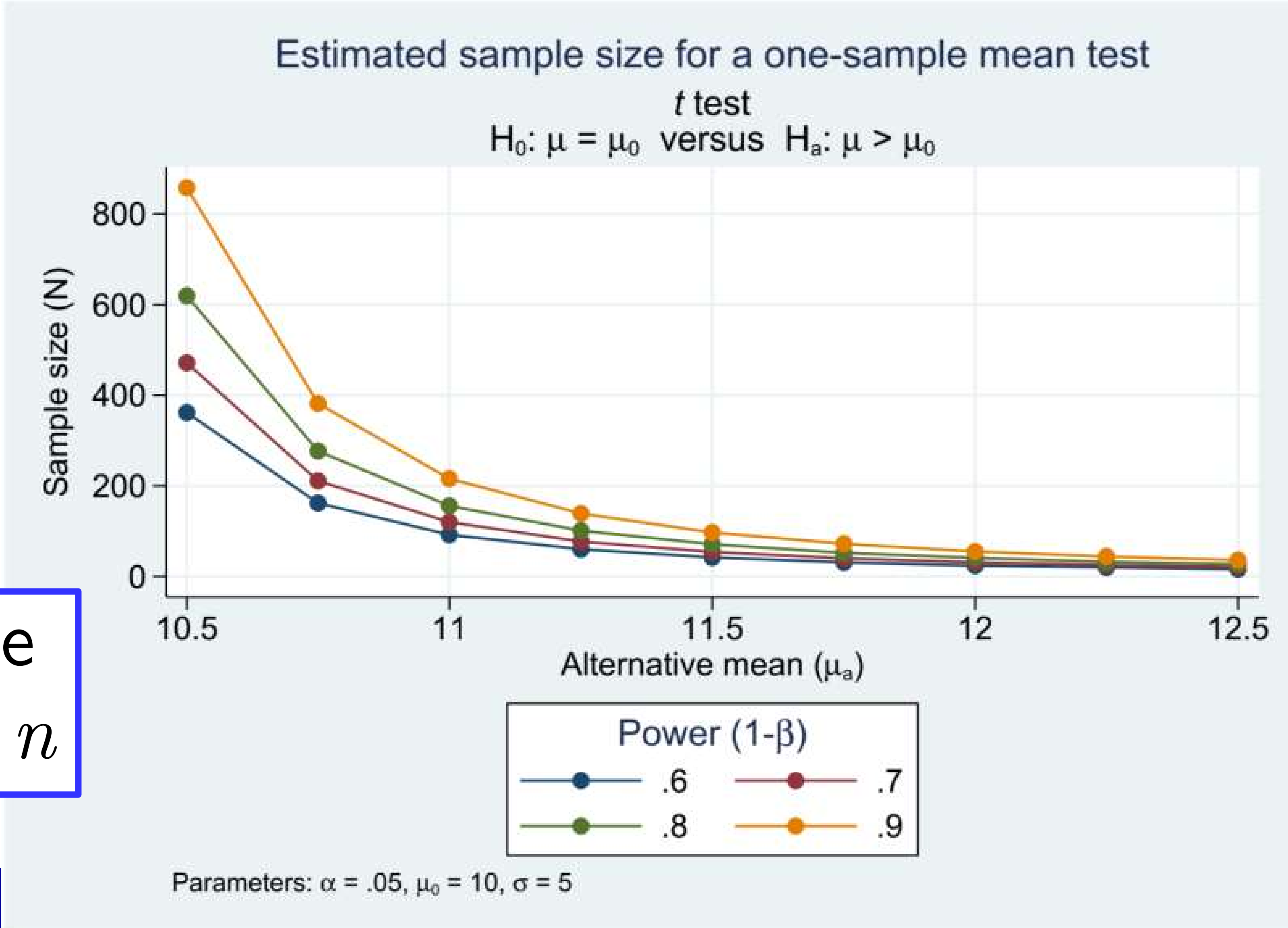
  - sample std; power=0.6-0.9
  - 1-sample t-test one-tailed test

Power Analysis: Graph Sample Size/Stata

- Plot sample size
  - STATA Results:

```
power onemean 10 (10.5(
```

```
ph
```



Larger effect size requires smaller n

Power Analysis: Two-sample t-test

1. Power Analysis: Find test power $\pi = 1 - \beta$, or
2. Find n to meet power requirement $\pi(n) \geq \overline{\pi}$

- Two-sample t-test (Common in experimental economics...)
- μ_1: Population mean of control group
- μ_2: Population mean of treatment group
  - Null Hypothesis: $H_0 : \mu_2 - \mu_1 = 0$
  - Alternative Hypothesis: $H_1 : \mu_2 - \mu_1 = d$
- Collect data of sample size n_1 and n_2

Effect Size from prior

Power Analysis: Two-Sample t-Test

- Test Size: $\alpha = 0.05 = \Pr(\text{reject null} \mid \text{null is true})$
- Type 2: $\beta = 0.20 = \Pr(\text{accept null} \mid \text{null is false})$
- Power: $\pi = 1 - \beta = 0.80$

$s_1^2, s_2^2 = \text{sample variances}$

- Pooled Sample STD: (with $\sigma_1^2 = \sigma_2^2$)

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$\overline{y}_1, \overline{y}_2 = \text{sample means}$

- Test Statistic: $t = \dfrac{\overline{y}_2 - \overline{y}_1}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$
- Reject if $t > t_{n_1+n_2-2,\alpha}$ $(t > z_\alpha \text{ for large } n)$

Power Analysis: Two-Sample t-Test

- If Equal Sample Size: $n_1 = n_2 = n$
- Pooled Sample STD (with $\sigma_1^2 = \sigma_2^2$):

$$s_p = \sqrt{\frac{s_1^2 + s_2^2}{2}}$$

$\bar{y}_1, \bar{y}_2 = \text{sample means}$

$s_1^2, s_2^2 = \text{sample variances}$

- Test Statistic:

$$t = \frac{\bar{y}_2 - \bar{y}_1}{s_p\sqrt{\frac{2}{n}}} \sim t(2n-2)$$

- Reject if $t > t_{2n-2,\alpha}$ $(t > z_\alpha \text{ for large } n)$

Power Analysis: Power of the Test

$$\pi = \Pr(t > z_\alpha | \mu_2 - \mu_1 = d) = \Pr\left[\frac{\bar{y}_2 - \bar{y}_1}{s_p\sqrt{2/n}} > z_\alpha \middle| \mu_2 - \mu_1 = d\right]$$

$$= \Pr\left(\bar{y}_2 - \bar{y}_1 > z_\alpha s_p \sqrt{2/n} \middle| \mu_2 - \mu_1 = d\right)$$

$$= \Pr\left(\frac{\bar{y}_2 - \bar{y}_1 - d}{s_p\sqrt{2/n}} > \frac{z_\alpha s_p \sqrt{2/n} - d}{s_p\sqrt{2/n}} \middle| \mu_2 - \mu_1 = d\right)$$

$$= \Phi\left(\frac{d - z_\alpha s_p \sqrt{2/n}}{s_p\sqrt{2/n}}\right)$$

$$\Rightarrow z_\beta = \frac{d - z_\alpha s_p \sqrt{2/n}}{s_p\sqrt{2/n}}$$

Power Analysis: How Big Should n Be?

Power $\pi = 1 - \beta = \Phi\left(\frac{d - z_\alpha s_p\sqrt{2/n}}{s_p\sqrt{2/n}}\right)$

$\alpha = 0.05, \quad \beta = 0.20$

$z_\alpha = 1.645, \quad z_\beta = 0.842$

$$\Rightarrow z_\beta = \frac{d - z_\alpha s_p\sqrt{2/n}}{s_p\sqrt{2/n}} \Rightarrow z_\beta + z_\alpha = \frac{d}{s_p\sqrt{2/n}}$$

$$\Rightarrow n = \frac{2s_p^2(z_\alpha + z_\beta)^2}{d^2}$$

$s_1 = 4.0,$

$s_2 = 5.84$

$\Rightarrow s_p^2 = \frac{s_1^2 + s_2^2}{2} = 5.0^2$

$d = 2$

$$= \frac{2(5^2)(1.645 + 0.842)^2}{2^2} = \underline{77.32}$$

- So we need $n \geq 78$

Power Analysis: Sample Size in Stata

- What is the sample size to get power $\pi = 0.80$?
  - STATA command for power calculation

μ_0 / μ_1

```
power twomeans 10 12 , sd1(4.0) sd2(5.84) oneside p(0.8)
```

- 2 sample std's   required power
- 2-sample t-test   one-tailed test

Power Analysis: Sample Size Result in Stata

- What is the sam
  - STATA Results:

```
power twomeans 10 1                                          p(0.8)
```

```
Estimated sample sizes for a two-sample means test
Satterthwaite's t test assuming unequal variances
Ho: m2 = m1  versus  Ha: m2 > m1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    2.0000
           m1 =   10.0000
           m2 =   12.0000
          sd1 =    4.0000
          sd2 =    5.8400

Estimated sample sizes:

            N =       158
  N per group =        79
```

Slightly larger n since STATA did not use normal approximation...

Power Analysis: Graph Power in STATA

- Plot power against sample size with `graph`
  - STATA command for power calculation

μ_0 / μ_1

```
power twomeans 10 12, sd1(4.0) sd2(5.84) n(20(10)200) oneside graph
```

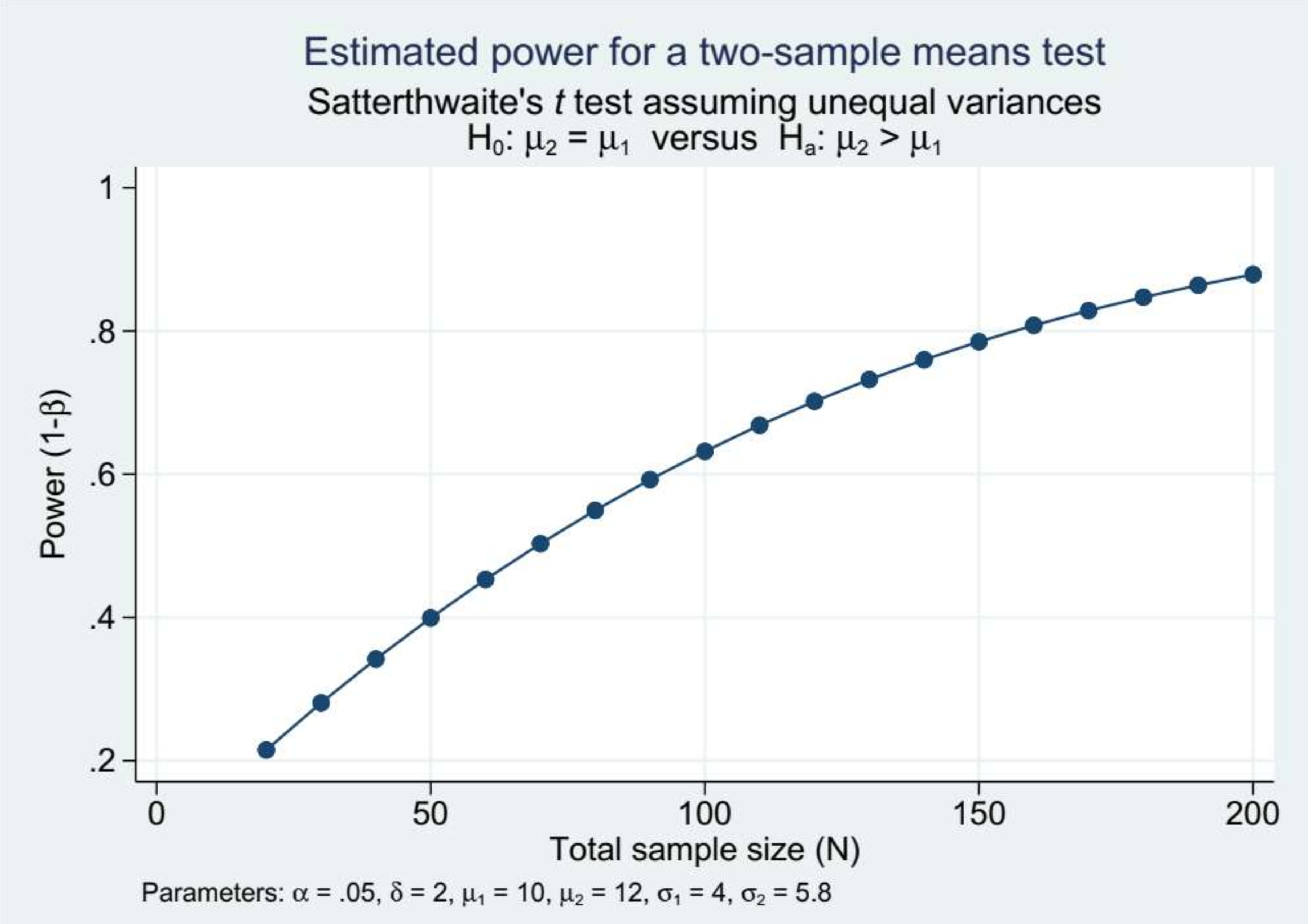
  - sample std; n=20-200
  - 2-sample t-test; one-tailed test

Power Analysis: Graph Power in STATA

- Plot power again
  - STATA Results:

```
power twomeans 10 12, s                                          h
```

Larger total same size yields higher power



Conclusion: The Replication Size Trinity

1. Sample Size n : # of observations/subjects
2. Effect Size d : How big is the true result
3. Power (1-β): How likely will your test show significance if there is truly an effect

- 1-sample t-Test vs. 2-sample t-Test

$$\Rightarrow n = \frac{s^2 (z_\alpha + z_\beta)^2}{(\mu_1 - \mu_0)^2}$$

$$\Rightarrow n = \frac{2 s_p^2 (z_\alpha + z_\beta)^2}{d^2}$$

Acknowledgement

- This presentation is based on
  - Section 1.1-1.4 of the lecture notes of Experimetrics,
- prepared for a mini-course taught by Peter G. Moffatt (UEA) at National Taiwan University in Spring 2019