

LECTURE 14

ERRORS IN HYPOTHESIS TESTING, EFFECT SIZE, & POWER

PSY2002

Hye Won Suk

ERRORS IN HYPOTHESIS TESTING

		The truth (Real difference)	
		The null hypothesis is true.	The alternative hypothesis is true.
Study findings (Decision)	Fail to reject the null hypothesis	Correct decision Probability = $1 - \alpha$	Type II error Probability = β
	Reject the null hypothesis	Type I error Probability = α	Correct decision Probability = $1 - \beta$ (power)

TYPE I ERROR

- Let's say that a researcher developed a new medication for the coronavirus.
- The medication is in fact not effective.
- However, the researcher performed a study to examine the effectiveness of the new medication and found a significant result (i.e., rejected the null hypothesis and concluded that the medication is effective).
- This is an example of a type I error. And the consequences of making a type I error can be very serious as you can see from the above example.

TYPE I ERROR

- A type I error (false-positive) occurs when a researcher rejects the null hypothesis that is actually true in the population.
- The probability of making a type I error (erroneously rejecting the null hypothesis) is equal to α (the level of significance).
- Type I error rate is set by the researcher. Oftentimes, $\alpha = .05$ is used. That is, the researcher has set 5% as the maximum chance of incorrectly rejecting the null hypothesis.

TYPE II ERROR

- Let's say that a researcher developed a new medication for the coronavirus.
- The medication is in fact effective.
- However, the researcher performed a study to examine the effectiveness of the new medication and failed to obtain a significant result (i.e., failed to reject the null hypothesis and concluded that the medication is not effective).
- This is an example of a type II error. And the consequences of making a type II error can be also very serious.

TYPE II ERROR

- A type II error (false-negative) occurs when a researcher fails to reject the null hypothesis that is actually false in the population.
- The probability of making a type II error (failing to detect the real difference or effects) is called β (beta).
- Type II error rate is not set by the researcher to a specific level in a hypothesis testing.

POWER

- The quantity, $1 - \beta$, is called power (or statistical power).
- The power indicates the probability of getting a significant result when there is a true difference in the population.
- A power of .90, for example, indicates there is a 90% chance of finding a significant result for the true difference. That is, roughly speaking, if a researcher performs an experiment to detect the true difference 100 times, he will get a significant results in 90 of them.

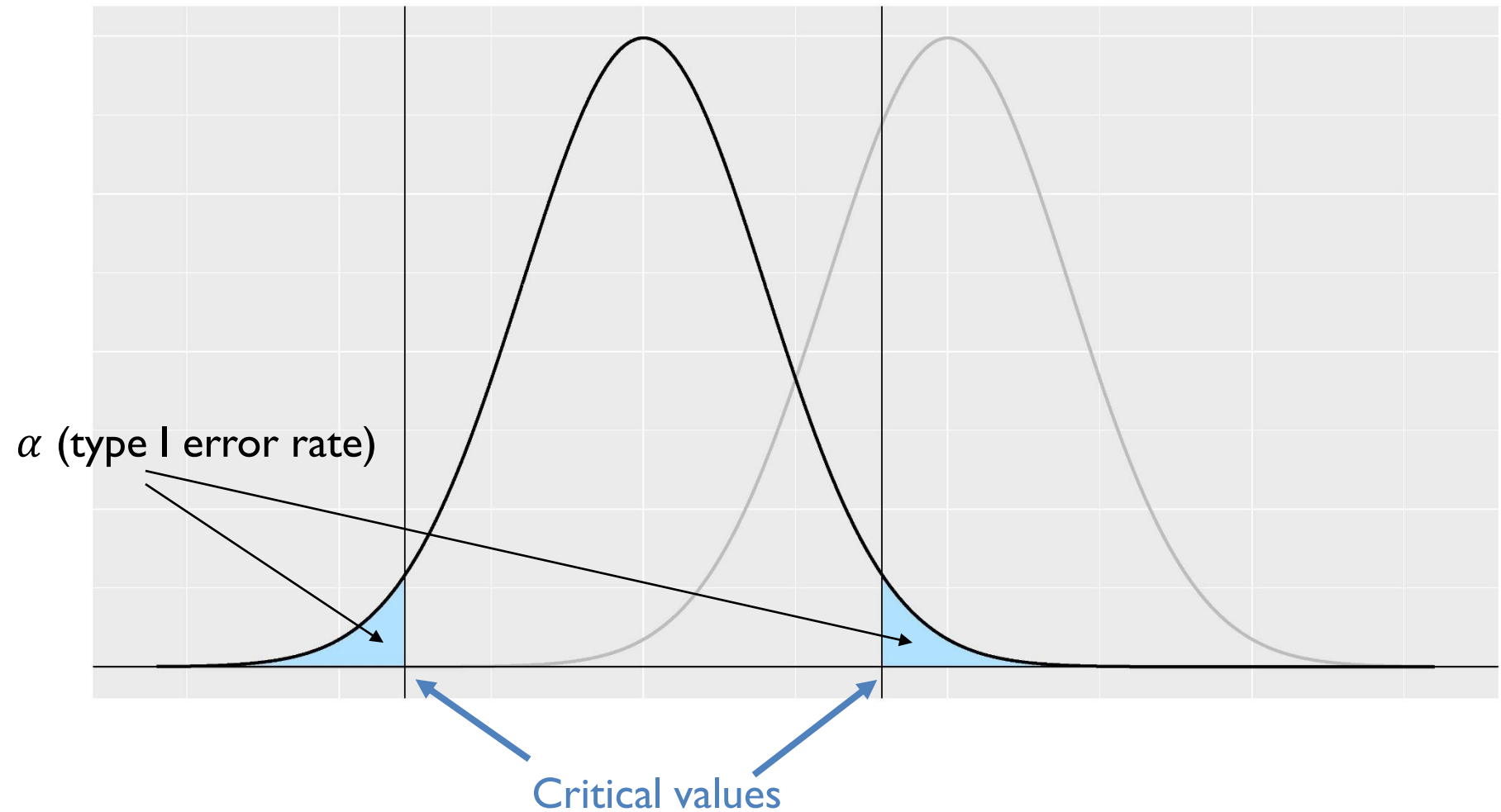
- A low power, say a power of .20, indicates that there is only a 20% chance of detecting the true difference. Roughly speaking, if a research repeats an experiment to detect the true difference 100 times, he will get a significant result only in 20 of them.
- Therefore, a low powered experiment is highly risky especially when the experiment requires a lot of resources such as money and time.

GRAPHICAL REPRESENTATION

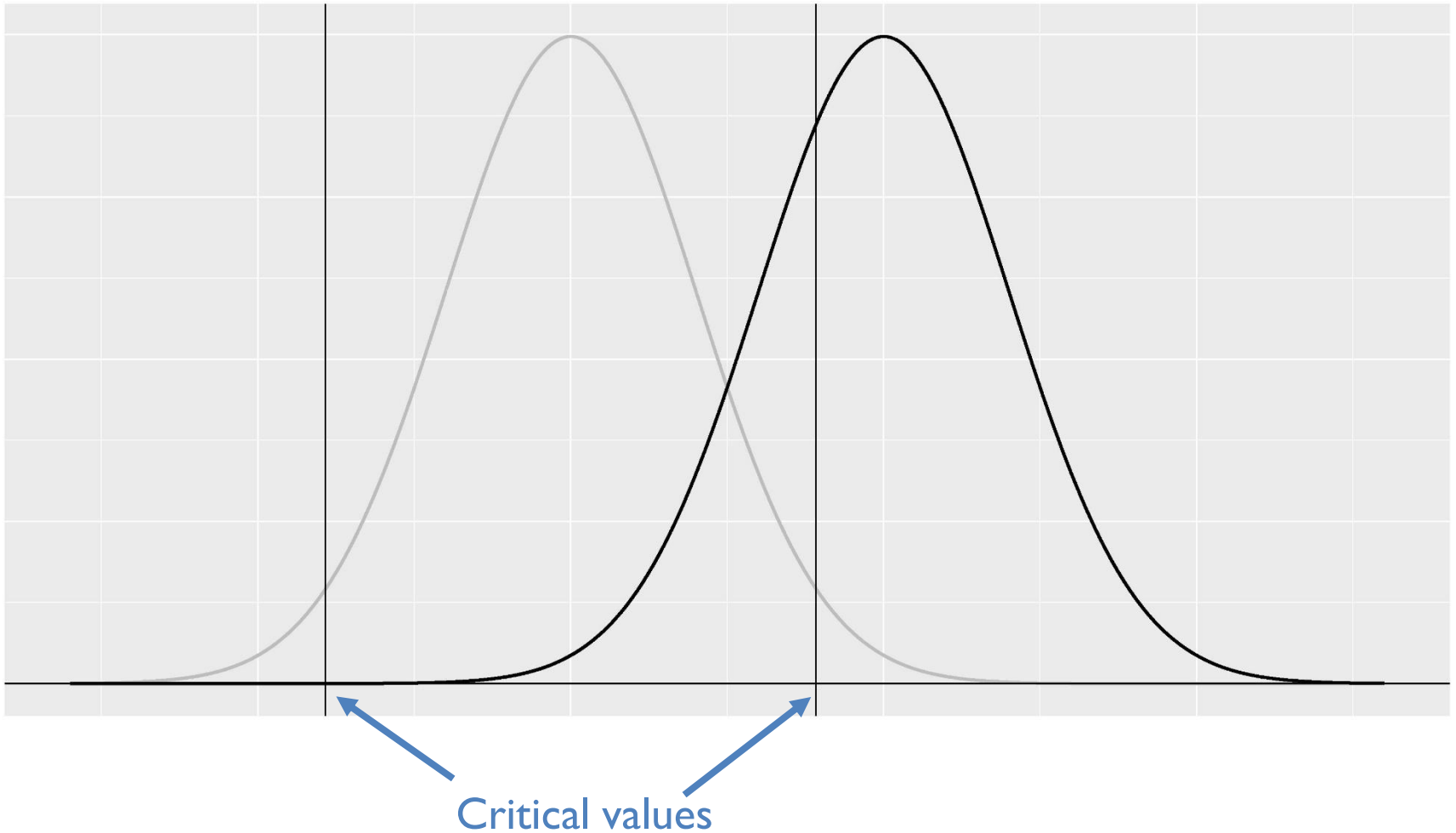
The distribution of the statistic under the null hypothesis



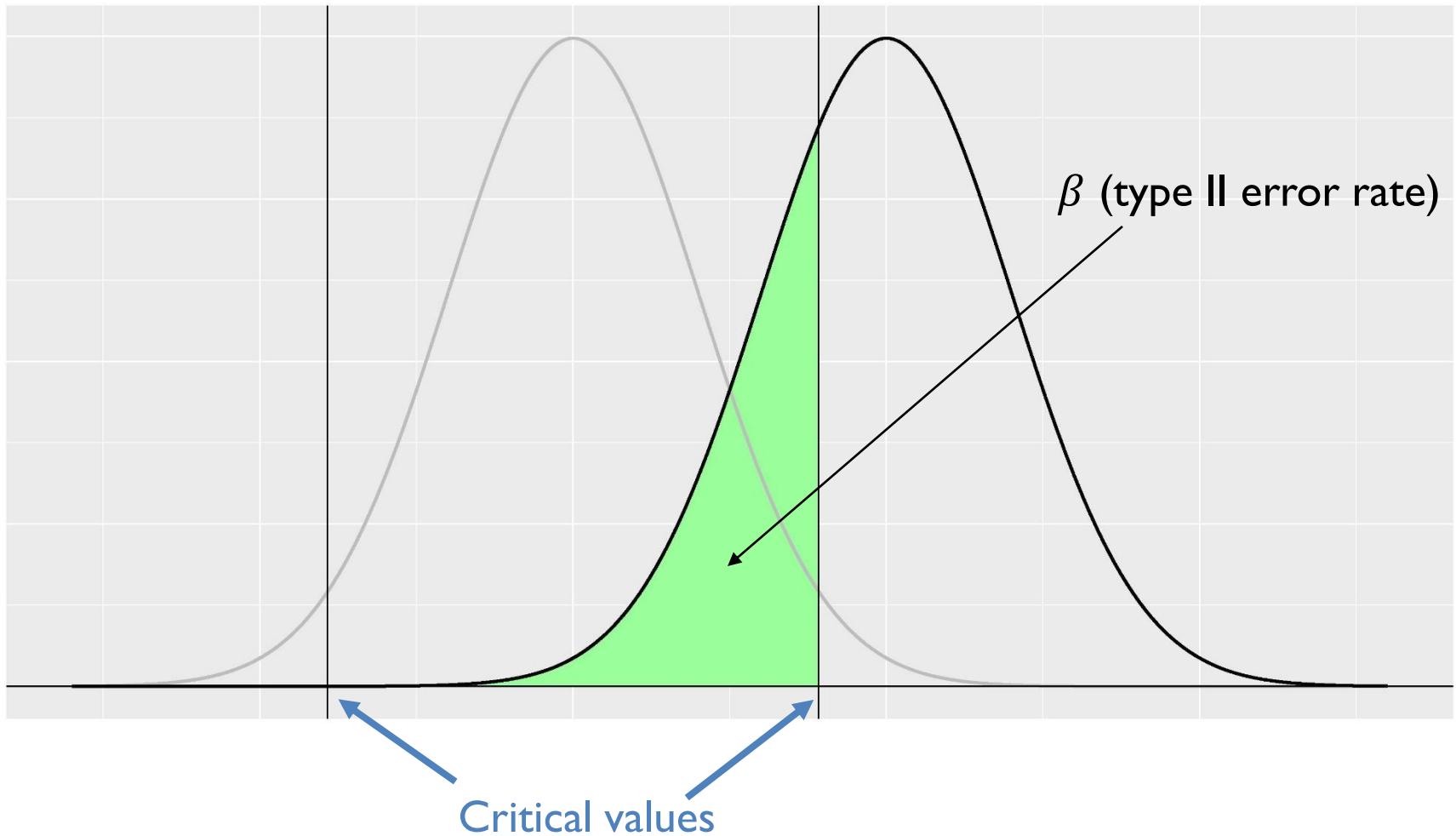
The distribution of the statistic
under the null hypothesis



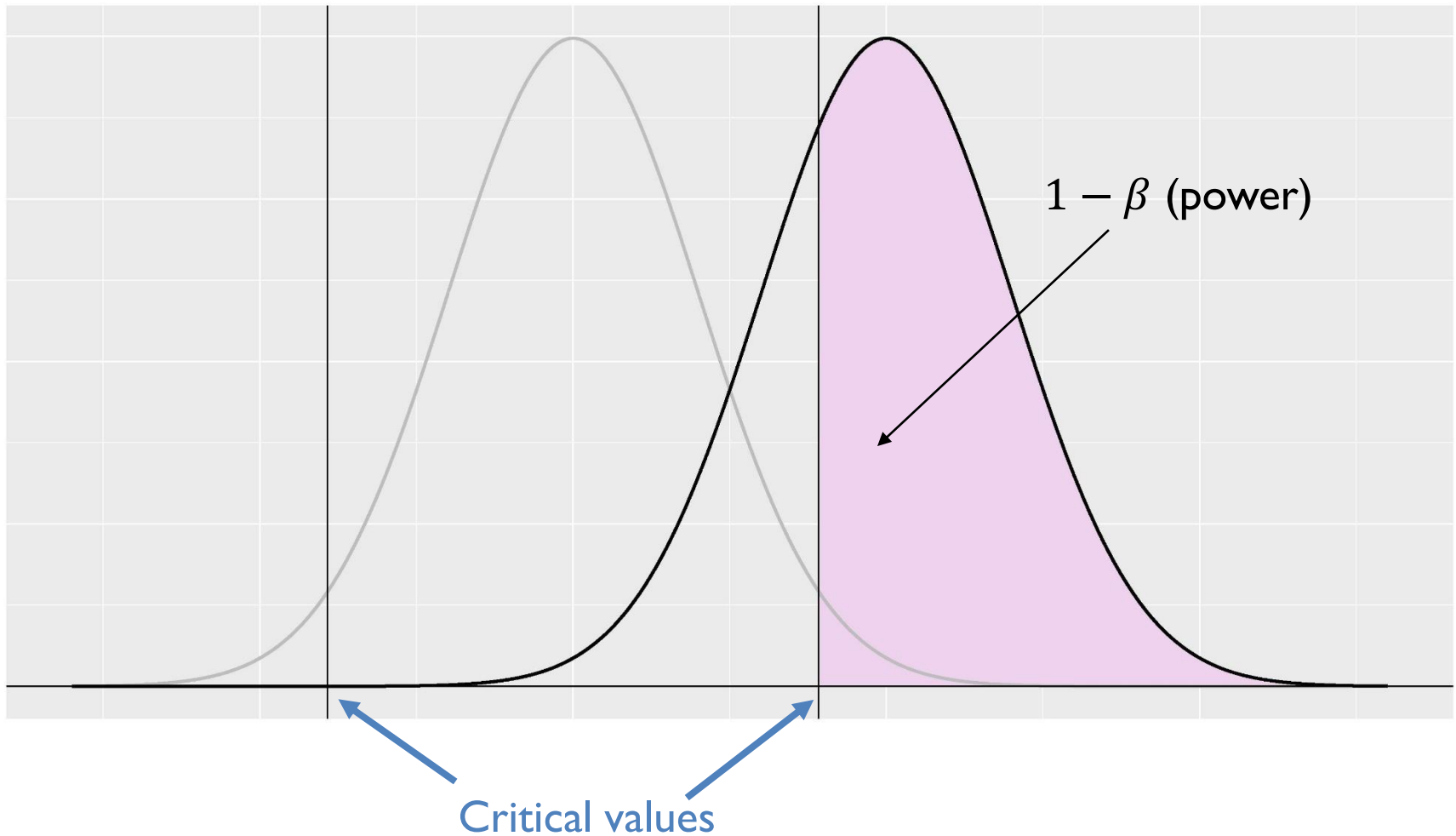
The distribution of the statistic
under the alternative hypothesis



The distribution of the statistic
under the alternative hypothesis



The distribution of the statistic
under the alternative hypothesis

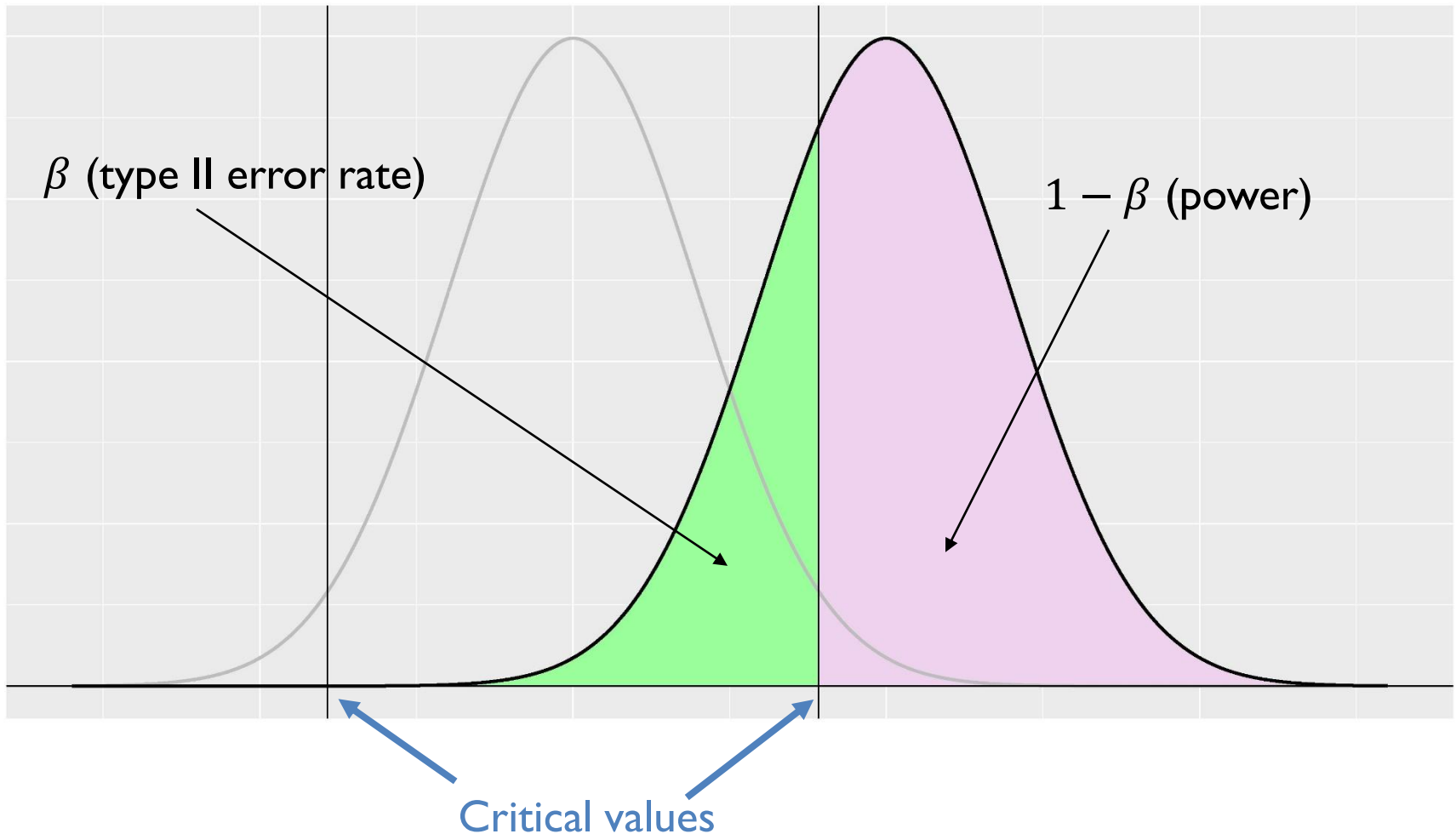


FACTORS AFFECTING POWER

- How to increase the power of a study?
- Let's consider the factors that affect the power.
 - Level of significance (α)
 - Effect size
 - Sample size

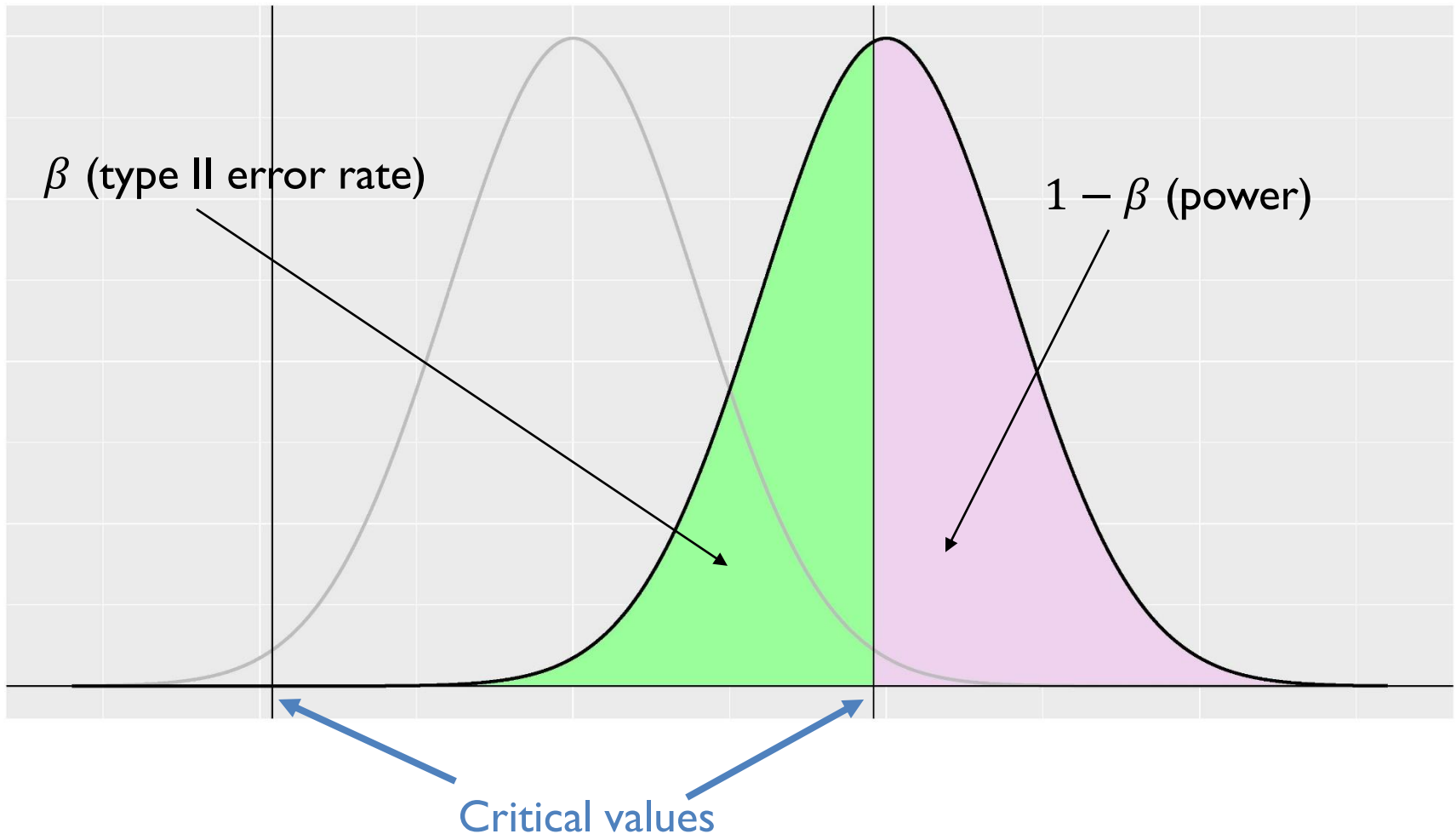
LEVEL OF SIGNIFICANCE (α)

$$\alpha = .05$$



LEVEL OF SIGNIFICANCE (α)

$$\alpha = .01$$



LEVEL OF SIGNIFICANCE (α)

- Decreasing the level of significance (0.05 to 0.01) leads to
 - More conservative test (Decreased type I error rate)
 - Increased type II error rate
 - Decreased power
- Increasing the level of significance (0.01 to 0.05) leads to
 - Less conservative test (Increased type I error rate)
 - Decreased type II error rate
 - Increased power
- Other things being equal, the higher the level of significance (α is larger) the higher the power is.

EFFECT SIZE

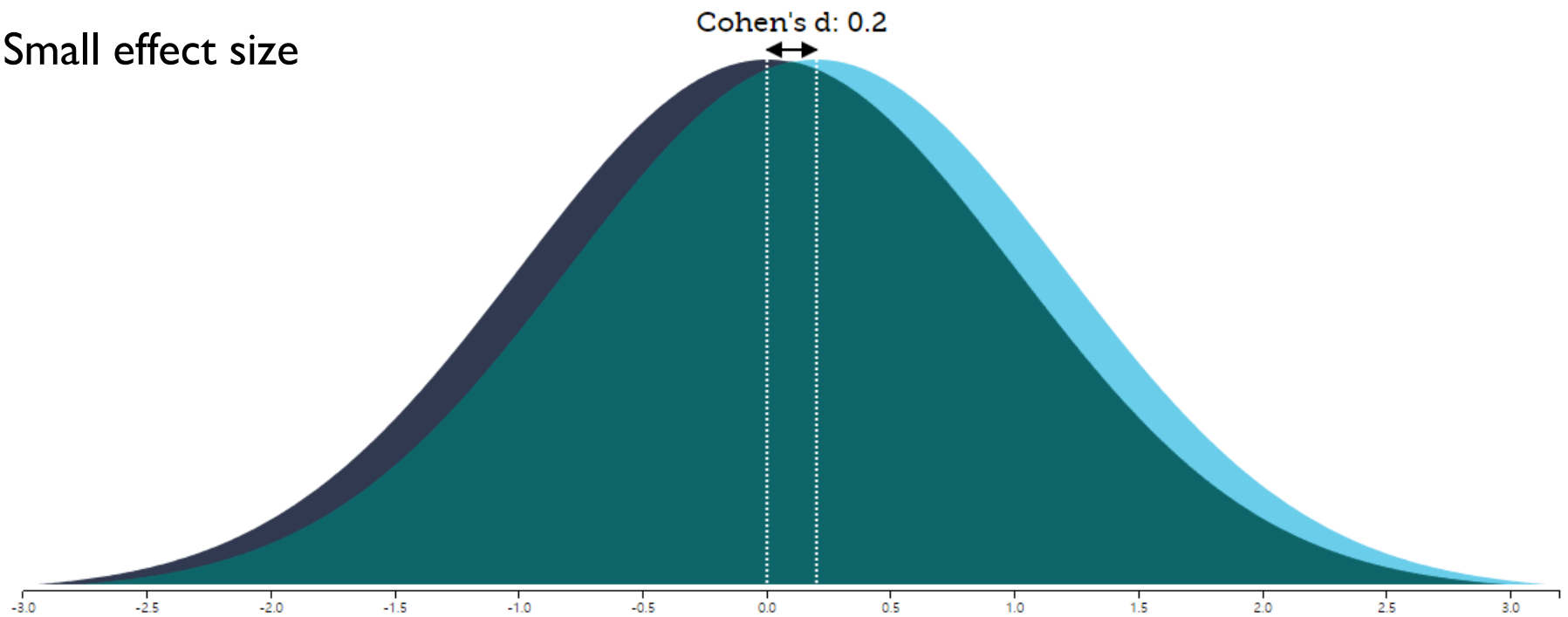
- The effect size means “the degree to which the null hypothesis is false” and serves as an index of degree of departure from the null hypothesis.
- Cohen’s d is one of the most commonly used effect size measure.

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

- μ_0 : population mean of the statistic under the null hypothesis
- μ_1 : population mean of the statistic under the alternative hypothesis
- σ : common standard deviation of the two distributions

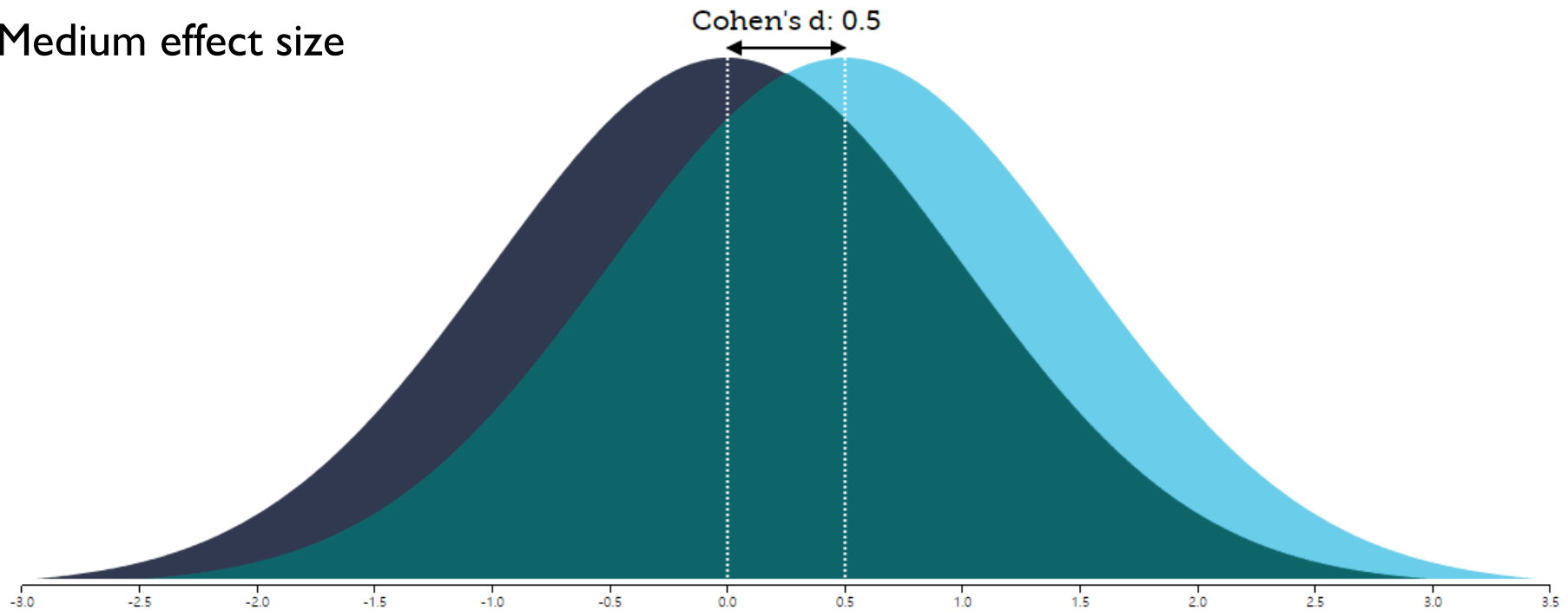
- A larger effect size means that the two distributions are less overlapping.
- <http://rpsychologist.com/d3/cohend/>
- Cohen (1988) Statistical power analysis for the behavioral sciences (second edition).
 - $d = .2$: small effect size
 - $d = .5$: medium effect size
 - $d = .8$: large effect size

Small effect size



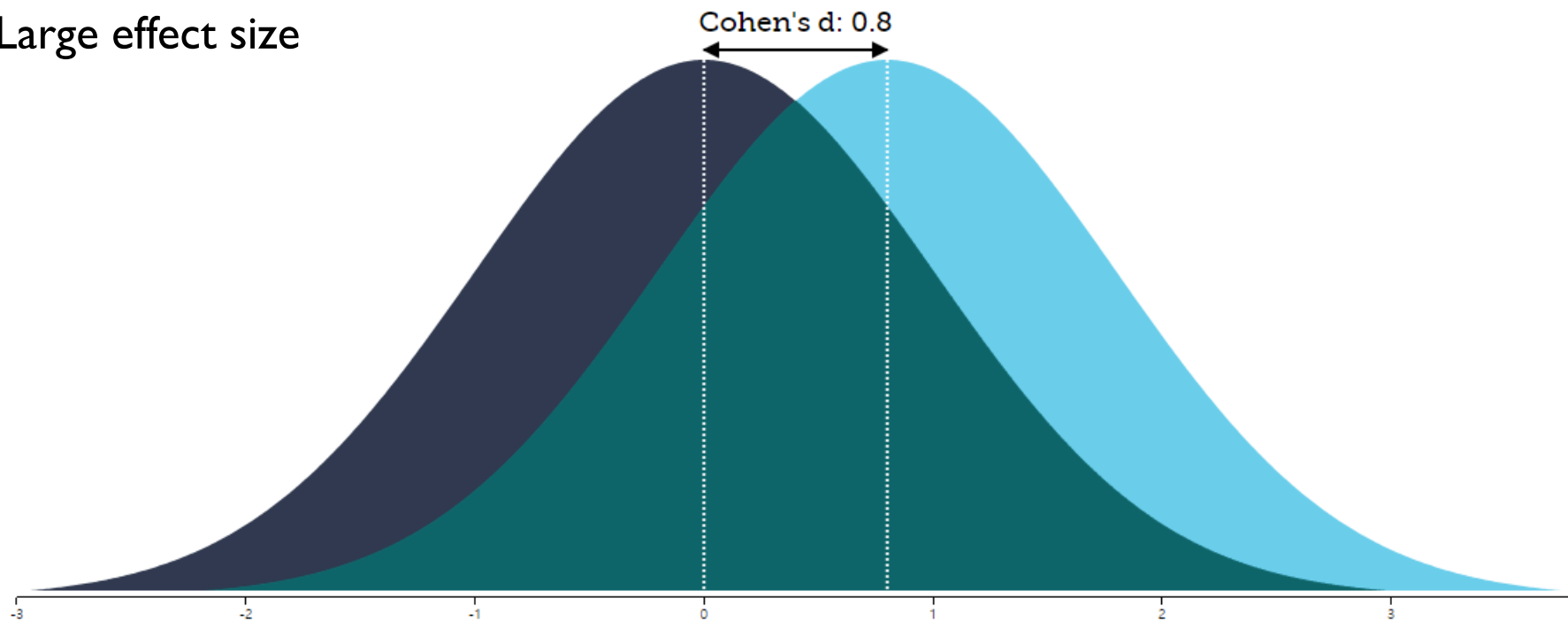
- Difference in heights between 15- and 16-year-old girls

Medium effect size



- Difference in heights between 14- and 18-year-old girls

Large effect size



- Difference in heights between 13- and 18-year-old girls

ONE-SAMPLE T-TEST

- Estimating Cohen's d

Population :

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

Sample :

$$d = \frac{\bar{X} - \mu_0}{s}$$

- Cohen's d should be estimated.
- μ_1 is estimated by \bar{X} .

ONE-SAMPLE T-TEST

- Estimating Cohen's d

Population :

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

Sample :

$$d = \frac{\bar{X} - \mu_0}{s}$$

- Cohen's d should be estimated.
- μ_1 is estimated by \bar{X} .
- σ is estimated by s .

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

ONE-SAMPLE T-TEST

- Estimating Cohen's d

Population :

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

Sample :

$$d = \frac{\bar{X} - \mu_0}{s}$$

- μ_0 doesn't need to be estimated because its value is assumed in the null hypothesis.

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}}$$

ONE-SAMPLE T-TEST

- Recall that $t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$

- This implies that

$$t \frac{1}{\sqrt{n}} = \frac{\bar{X} - \mu_0}{s} = d$$

- A larger t value does not necessarily indicate a larger effect size, and vice versa.

ONE-SAMPLE T-TEST

The TTEST Procedure

Variable: weight

N	Mean	Std Dev	Std Err	Minimum	Maximum
27	2945.8	177.6	34.1780	2619.0	3189.0

Mean	95% CL Mean	Std Dev	95% CL Std Dev
2945.8	2875.5 3016.0	177.6	139.9 243.4

DF	t Value	Pr > t
26	4.27	0.0002

- $$\begin{aligned}
 \text{Cohen's } d &= \frac{\bar{X} - \mu_0}{s} \\
 &= \frac{2945.8 - 2800}{177.6} \\
 &= 0.82
 \end{aligned}$$

- $$\begin{aligned}
 \text{Cohen's } d &= t \frac{1}{\sqrt{n}} \\
 &= 4.27 \frac{1}{\sqrt{27}} \\
 &= 0.82
 \end{aligned}$$

ONE-SAMPLE T-TEST

The TTEST Procedure

Variable: weight

N	Mean	Std Dev	Std Err	Minimum	Maximum
27	2945.8	177.6	34.1780	2619.0	3189.0

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
2945.8	2875.5	3016.0	177.6	139.9	243.4

DF	t Value	Pr > t
26	4.27	0.0002

- The prenatal care has a significant effect on the birthweights of babies born to poor women ($t(26) = 4.27$, $p = .0002$, $d = 0.82$).

PAIRED-SAMPLES T-TEST

- Cohen's d can be estimated in exactly the same way as for the one-sample t-tests.

Population :

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

Sample :

$$d = \frac{\bar{D} - 0}{s_D}$$

- Cohen's d should be estimated.
- μ_1 is estimated by \bar{D} .

PAIRED-SAMPLES T-TEST

- Cohen's d can be estimated in exactly the same way as for the one-sample t-tests.

Population :

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

- Cohen's d should be estimated.
- μ_1 is estimated by \bar{D} .
- σ is estimated by s_D .

Sample :

$$d = \frac{\bar{D} - 0}{s_D}$$

$$s_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n - 1}}$$

PAIRED-SAMPLES T-TEST

- Cohen's d can be estimated in exactly the same way as for the one-sample t-tests.

Population :

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

- μ_0 doesn't need to be estimated because its value is assumed to be zero in the null hypothesis.

Sample :

$$d = \frac{\bar{D} - 0}{s_D}$$

$$s_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n - 1}}$$

PAIRED-SAMPLES T-TEST

- Recall that $t = \frac{\bar{D} - 0}{\frac{s_D}{\sqrt{n}}}$

- This implies that

$$t \frac{1}{\sqrt{n}} = \frac{\bar{D}}{s_D} = d$$

- A larger t value does not necessarily indicate a larger effect size, and vice versa.

PAIRED-SAMPLES T-TEST

The TTEST Procedure

Difference: calc3 - psychstats

N	Mean	Std Dev	Std Err	Minimum	Maximum
6	3.0000	3.6332	1.4832	-2.0000	9.0000

Mean	95% CL Mean	Std Dev	95% CL Std Dev
3.0000	-0.8128 6.8128	3.6332	2.2679 8.9108

DF	t Value	Pr > t
5	2.02	0.0990

- $$\text{Cohen's } d = \frac{\bar{D}}{s_D} = \frac{3}{3.6332} = 0.83$$

- $$\text{Cohen's } d = t \frac{1}{\sqrt{n}} = 2.02 \frac{1}{\sqrt{6}} = 0.83$$

PAIRED-SAMPLES T-TEST

The TTEST Procedure

Difference: calc3 - psychstats

N	Mean	Std Dev	Std Err	Minimum	Maximum
6	3.0000	3.6332	1.4832	-2.0000	9.0000

Mean	95% CL Mean	Std Dev	95% CL Std Dev
3.0000	-0.8128 6.8128	3.6332	2.2679 8.9108

DF	t Value	Pr > t
5	2.02	0.0990

- There is no significant difference in sleeping time between the two classes ($t(5) = 2.02$, $p = .0990$, $d = .83$). This implies that the class materials for the two classes, Calculus III and Psych Stats, are equally interesting.

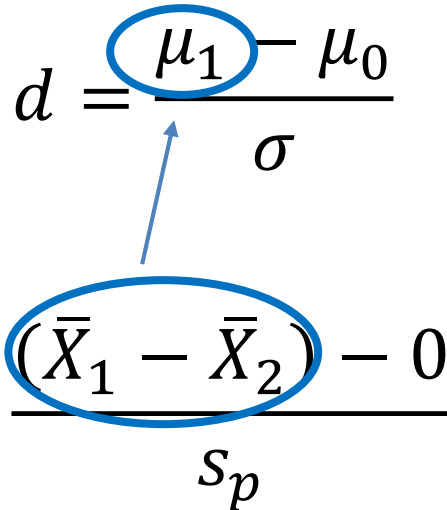
INDEPENDENT-SAMPLES T-TEST

- Estimating Cohen's d

Population :

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

Sample :

$$d = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{s_p}$$


- Cohen's d should be estimated.
- μ_1 is estimated by $\bar{X}_1 - \bar{X}_2$.

INDEPENDENT-SAMPLES T-TEST

- Estimating Cohen's d

Population :

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

Sample :

$$d = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{s_p}$$

- Cohen's d should be estimated.
- μ_1 is estimated by $\bar{X}_1 - \bar{X}_2$.
- σ is estimated by s_p .

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

INDEPENDENT-SAMPLES T-TEST

- Estimating Cohen's d

Population :

$$d = \frac{\mu_1 - \mu_0}{\sigma}$$

Sample :

$$d = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{s_p}$$

- μ_0 doesn't need to be estimated.
- In the null hypothesis, it is assumed to be zero.

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

INDEPENDENT-SAMPLES T-TEST

- Recall that $t = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

- This implies that

$$t \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \frac{(\bar{X}_1 - \bar{X}_2)}{s_p} = d$$

- A larger t value does not necessarily indicate a larger effect size, and vice versa.

INDEPENDENT-SAMPLES T-TEST

The TTEST Procedure

Variable: speed

group	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	15	43.0707	6.7814	1.7509	28.5700	57.6800
2	15	36.9580	5.1020	1.3173	23.5400	43.7500
Diff (1-2)		6.1127	6.0007	2.1911		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		43.0707	39.3153	46.8261	6.7814	4.9648	10.6949
2		36.9580	34.1326	39.7834	5.1020	3.7353	8.0463
Diff (1-2)	Pooled	6.1127	1.6243	10.6010	6.0007	4.7620	8.1157
Diff (1-2)	Satterthwaite	6.1127	1.6087	10.6166			

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	28	2.79	0.0094
Satterthwaite	Unequal	26.003	2.79	0.0097

- $$Cohen's d = \frac{(\bar{X}_1 - \bar{X}_2)}{s_p}$$

$$= \frac{6.1127}{6.0007} = 1.02$$

- $$Cohen's d = t \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= 2.79 \sqrt{\frac{1}{15} + \frac{1}{15}} = 1.02$$

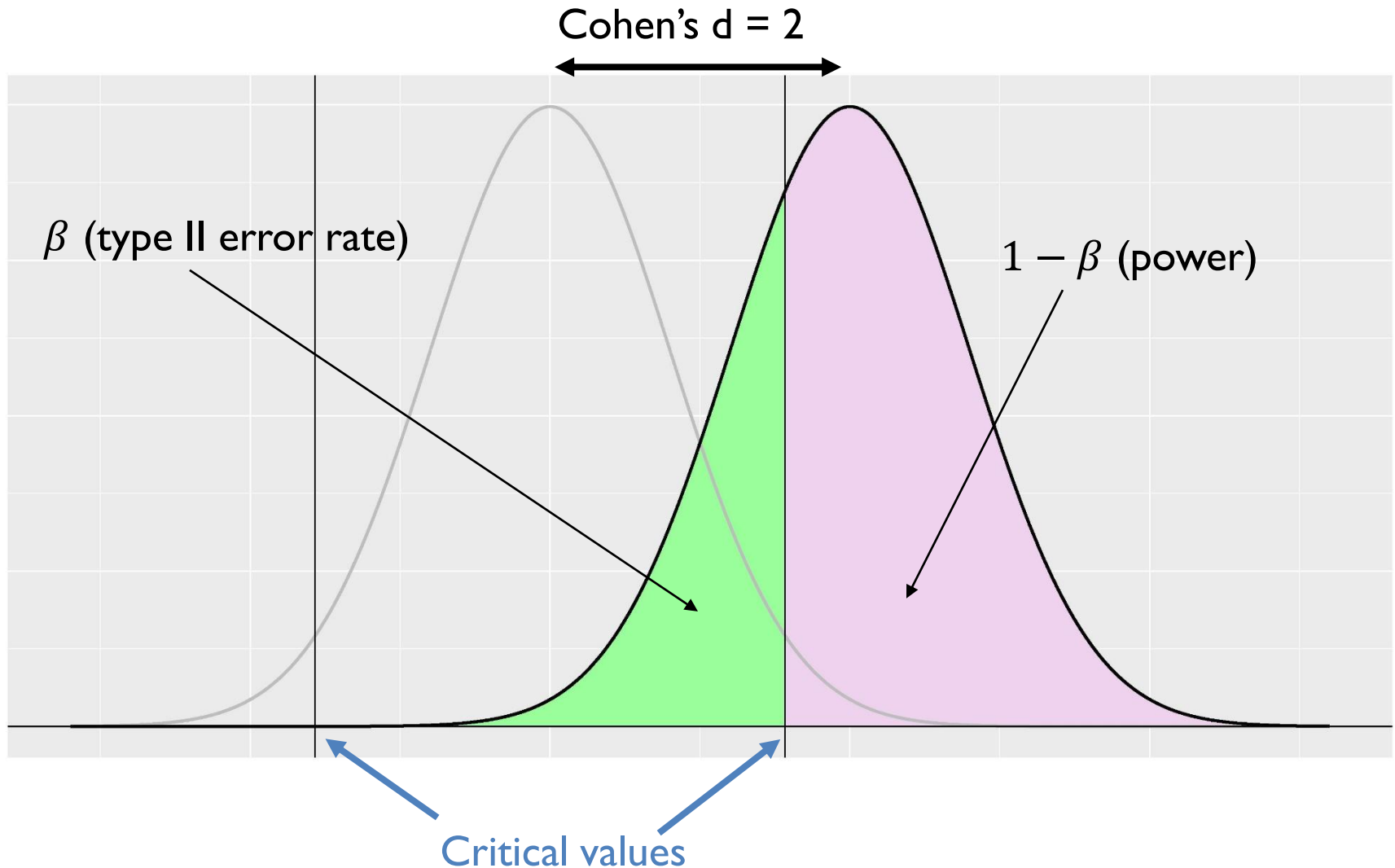
INDEPENDENT-SAMPLES T-TEST

- The mean of the estimated speed of the SMASH group and that of the HIT group are significantly different ($t(28) = 2.79$, $p = .0094$, $d = 1.02$). This result implies that the wording used to ask a question can influence eyewitness's memory.

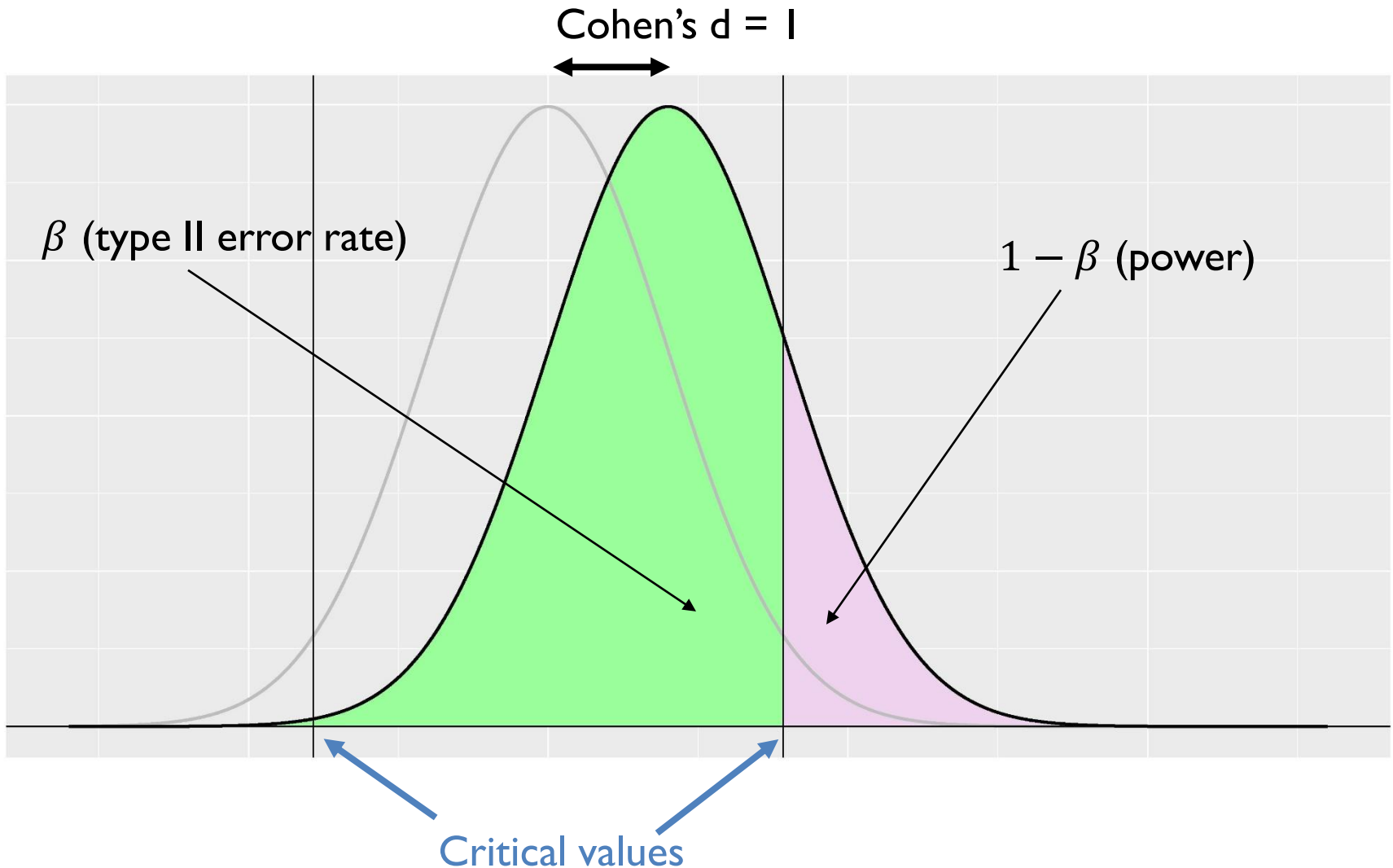
EFFECT SIZE

- In a hypothesis testing, it is **HIGHLY RECOMMENDED** to report an effect size as well as a p -value.
- As we have seen, a smaller p -value (i.e., a larger t value) does not necessarily indicates a larger effect size.

EFFECT SIZE AND POWER



EFFECT SIZE AND POWER



EFFECT SIZE AND POWER

- A smaller effect size is associated with
 - Increased type II error rate
 - Decreased power
 - (no effect on type I error rate)
- A larger effect size is associated with
 - Decreased type II error rate
 - Increased power
 - (no effect on type I error rate)
- Other things being equal, the larger the effect size the higher the power is.

SAMPLE SIZE AND POWER

- Sample size affects the standard error of a test statistic.
- Standard error decreases as sample size increases.

- Sampling distribution of mean

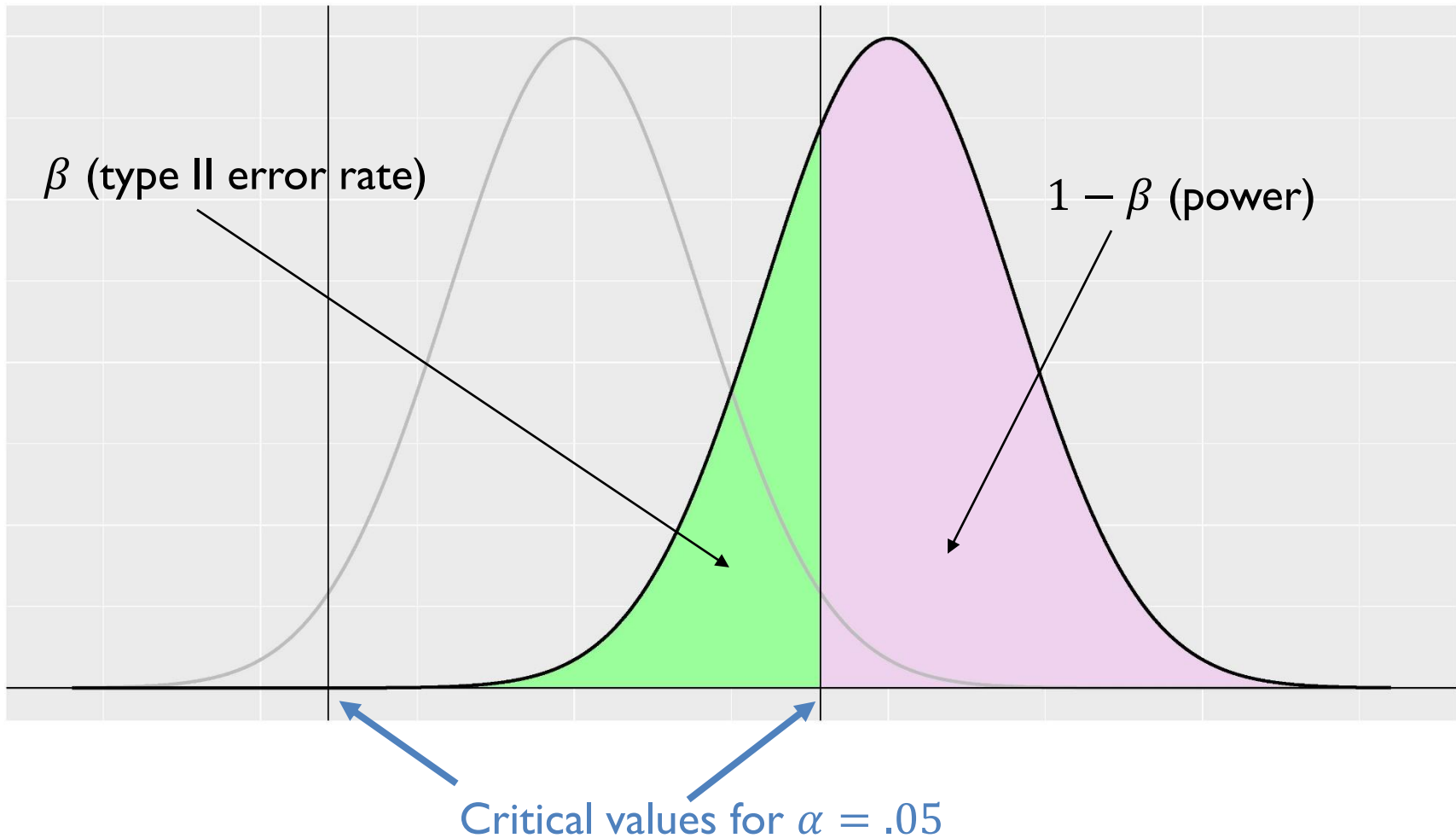
- Standard error of mean = $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

- Sampling distribution of mean difference

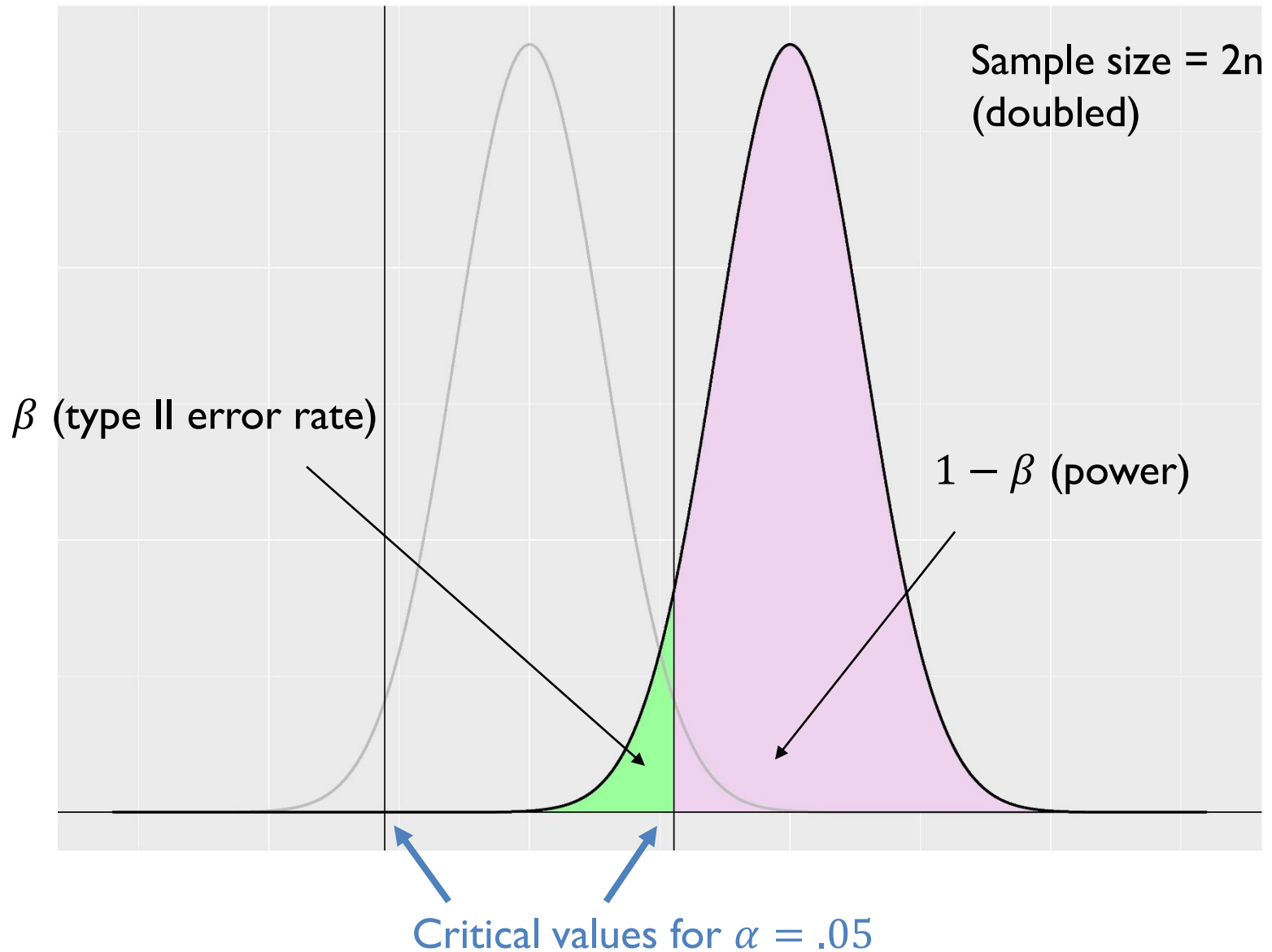
- Standard error of mean difference = $\sigma_{\bar{X}_1 - \bar{X}_2} = \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

SAMPLE SIZE AND POWER

Sample size = n



SAMPLE SIZE AND POWER



SAMPLE SIZE AND POWER

- A smaller sample size leads to
 - Increased type II error rate
 - Decreased power
 - (no effect on type I error rate)
- A larger sample size leads to
 - Decreased type II error rate
 - Increased power
 - (no effect on type I error rate)
- Other things being equal, the larger the sample size the higher the power is.

POWER ANALYSIS

- Type I error rate (or α), effect size, sample size, and power are related to each other.
- If three of them are known, the fourth one is determined.
- Two different types of power analysis
 - Post hoc power analysis
 - A prior power analysis

POST HOC POWER ANALYSIS

- A post hoc power analysis is done to find the power of a test after the experiment has been performed.
- α , effect size, and sample sizes are given.
- The effect size can be obtained from the estimate from the current study. It can also be obtained from:
 - Previous studies
 - Literature review

POST HOC POWER ANALYSIS

- An example
 - An educational researcher has developed two different programs (A and B) for teaching elementary algebra. He recruited 100 pupils, randomly assigned 50 of them to the A program and the rest to the B program. Then he tested all participants on a common algebra achievement test. He obtained $d = .40$ from this study. At $\alpha = .05$, what was the power of this study?

A PRIORI POWER ANALYSIS

- A priori power analysis is used to find the required sample size to achieve a desired level of power (usually .80) before an experiment is performed.
- α , effect size, and desired power level are given.
- The effect size can be obtained from
 - Experiences
 - Previous studies
 - Literature review

A PRIORI POWER ANALYSIS

- An example
 - An educational researcher has developed two different programs (A and B) for teaching elementary algebra. He anticipates a small-to-medium effect, $d = .40$, from literature review. At $\alpha = .05$, what should be the sample size (in each group) to achieve the power level of .80?

A PRIORI POWER ANALYSIS

- In general, a larger sample size enables us
 - to achieve a higher level of power for a given effect size and a given type I error rate;
 - to detect an effect of a smaller size at a given level of power and a given type I error rate;
 - to achieve the same level of power for a given effect size at a lower type I error rate.

G*POWER

- G*Power is free software for power analysis.
- You can download the current version of G*Power from <http://www.gpower.hhu.de/>.
- You can also download the manual and user guide from this website.

SUMMARY

- Errors in hypothesis testing
 - Type I error
 - Type II error
- Factors affecting the power
 - Other things being equal...
 - The higher the level of significance (α is larger) the higher the power is.
 - The larger the effect size the higher the power is.
 - The larger the sample size the higher the power is.
- Power analysis
 - Post hoc
 - A priori
- <http://webzine.kpsy.co.kr/2017autumn/sub.html?psyNow=12&UID=225> (in Korean)