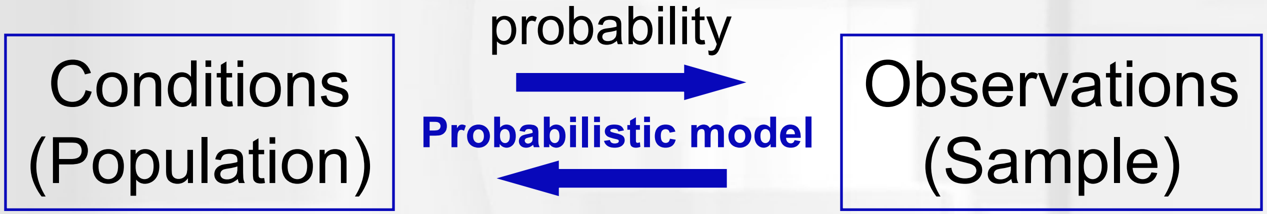
Lecture 06: Introduction to Hypothesis Testing

Flashback: induction and deduction

- Probability versus statistics:



Statistics
(estimation and hypothesis testing)
估值和假设检验

Inferential Statistics

Two major forms:

Estimation is concerned with predicting values or intervals of specific population parameters, based on a set of observed data.

Hypothesis Testing is concerned with testing whether the value of a population parameter is equal to some specific value (or another parameter), based on a set of observed data.

Both are based on a sample from the population.

- The setting of hypothesis testing
  - Null hypothesis
  - Critical region
  - Test statistics
  - Significance and p-value
- Errors in Hypothesis Tests
- Effect size and power

Research starts with questions

1. Is the mean self-esteem level of kids in big cities equal to that in general population?
2. Is the new psychological therapy more effective than old ones?
3. Do mothers in north China deliver babies with higher birthweight than mothers in south China?

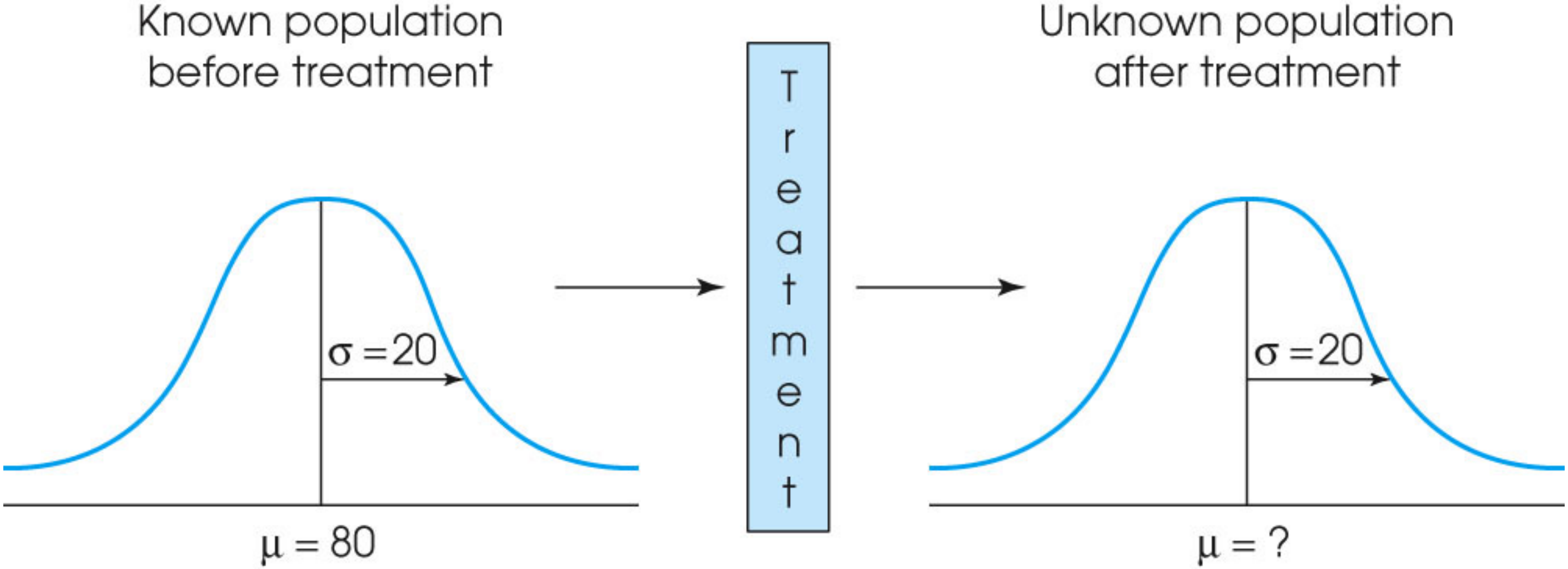
Hypothesis Testing

- The general goal of a hypothesis test is to rule out chance (sampling error) as a plausible explanation for the results from a research study.
  - Hypothesis testing is a technique to help determine whether a specific treatment has an effect on the individuals in a population
  - or whether the value of a population parameter is equal to some specific value (or another parameter), based on a set of observations.

Hypothesis Testing

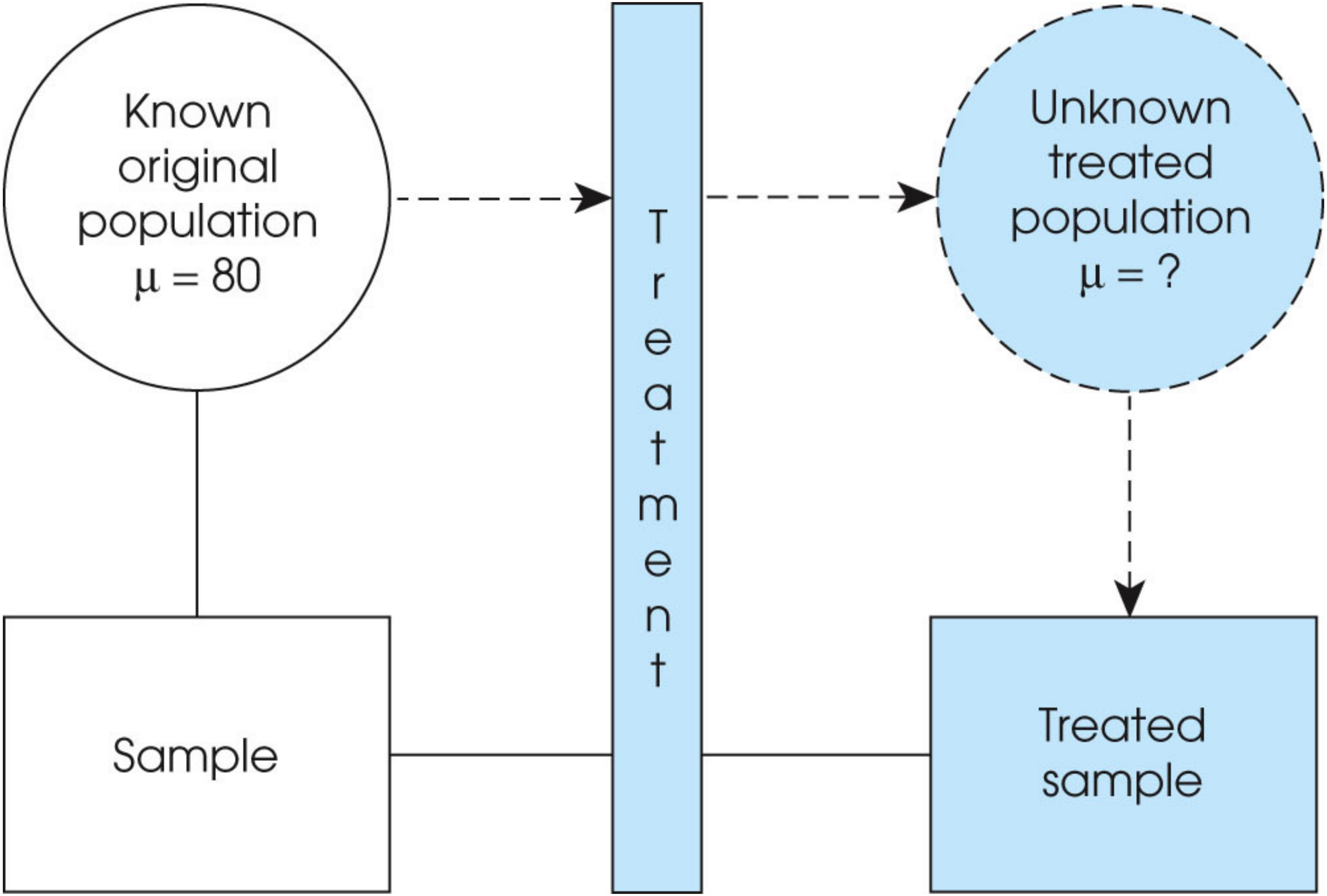
The hypothesis test is used to evaluate the results from a research study in which

1. A sample is selected from the population.
2. The treatment is administered to the sample.
3. After treatment, the individuals in the sample are measured.


Known population
before treatment
σ = 20
μ = 80
Treatment
Unknown population
after treatment
σ = 20
μ = ?

Hypothesis Testing (cont.)

- If the individuals in the sample are noticeably different from the individuals in the original population, we have evidence that the treatment has an effect.
- However, it is also possible that the difference between the sample and the population is simply a sampling error.


Known original population $\mu = 80$

Treatment

Unknown treated population $\mu = ?$

Sample

Treated sample

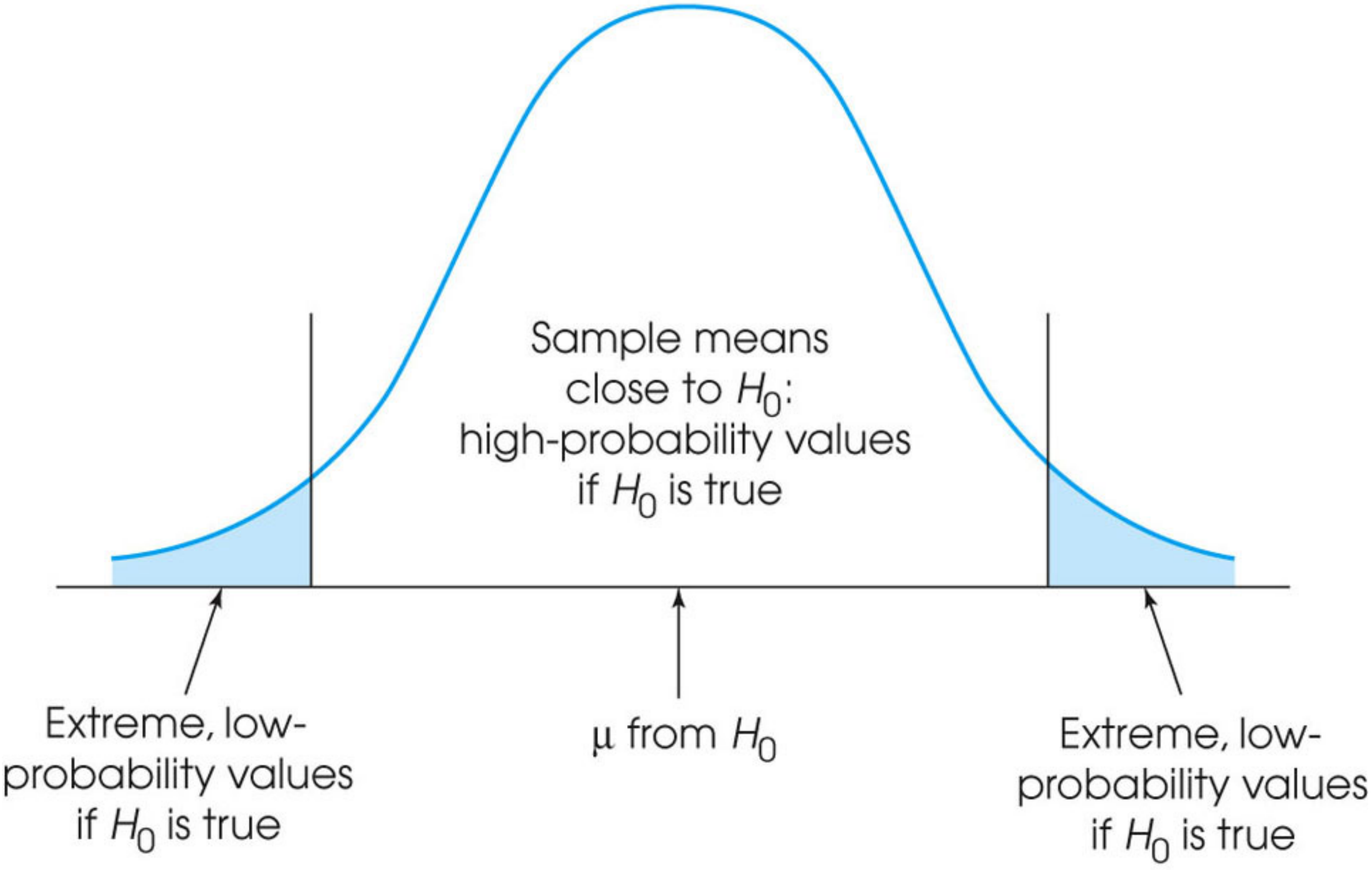
Hypothesis Testing (cont.)

The purpose of the hypothesis test is to decide between two explanations:

1. The difference between the sample and the population can be explained by sampling error (there does not appear to be a treatment effect).

2. The difference between the sample and the population is too large to be explained by sampling error (there does ***appear*** to be a treatment effect).

The distribution of sample means
if the null hypothesis is true
(all the possible outcomes)



Sample means
close to H_0:
high-probability values
if H_0 is true

Extreme, low-
probability values
if H_0 is true

μ from H_0

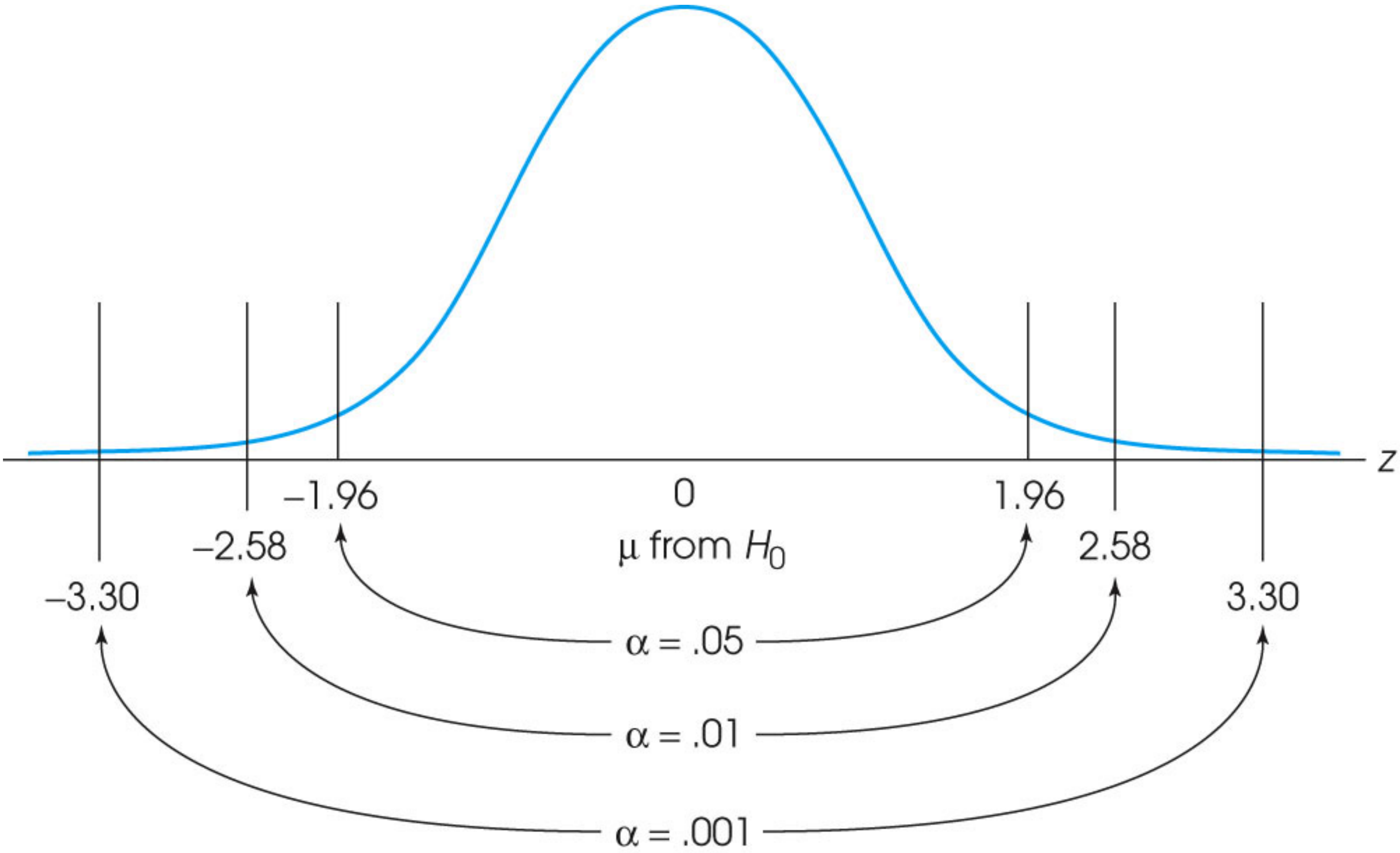
Extreme, low-
probability values
if H_0 is true

The Null Hypothesis, the Alpha Level, the Critical Region, and the Test Statistic

- The following four steps outline the process of hypothesis testing and introduce some of the new terminology:

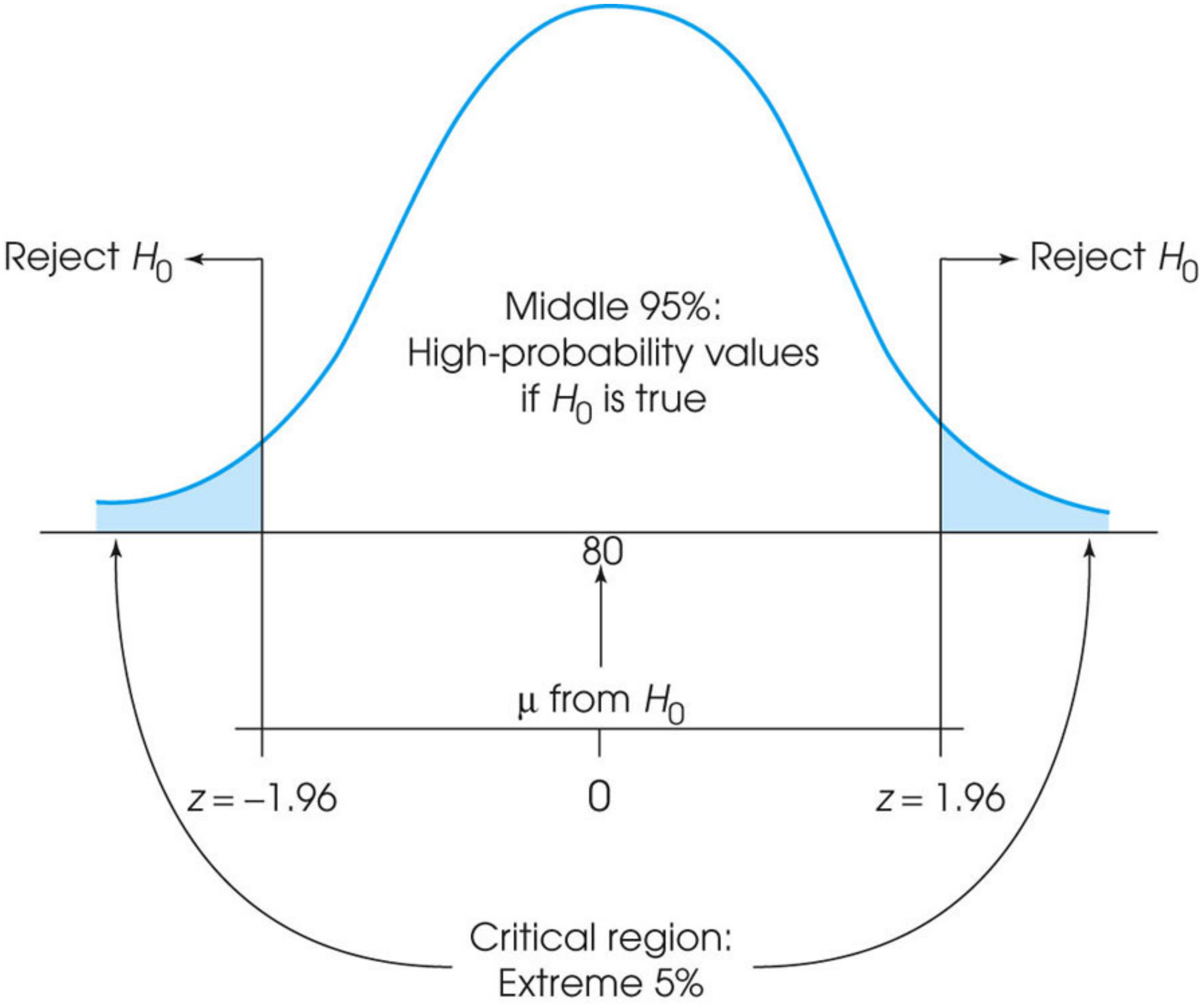
Step 1

State the hypotheses and select an α level. The **null hypothesis**, H_0, always states that the treatment has no effect (no change, no difference). According to the null hypothesis, the population mean after treatment is the same as it was before treatment. The **α level** establishes a criterion, or "cut-off", for making a decision about the null hypothesis.


z
0
μ from H_0
−1.96
1.96
−2.58
2.58
−3.30
3.30
α = .05
α = .01
α = .001

Step 2

Locate the critical region. The **critical region** (aka rejection region) consists of outcomes that are very unlikely to occur if the null hypothesis is true. That is, the critical region is defined by sample means that are almost impossible to obtain if the treatment has no effect. The phrase “almost impossible” means that these samples have a probability (p) that is ***less than*** the alpha level.


Reject H_0
Reject H_0
Middle 95%:
High-probability values
if H_0 is true
80
μ from H_0
$z = -1.96$
0
$z = 1.96$
Critical region:
Extreme 5%

Step 3

Compute the test statistic. The **test statistic** (in this chapter a z-score) forms a ratio comparing the obtained *difference* between the sample mean and the hypothesized population mean versus the amount of *difference* we would expect without any treatment effect (the standard error).

Step 4

A large value for the test statistic shows that the obtained difference in the mean is more than would be expected if there is no treatment effect. If it is large enough to be in the critical region, we conclude that the difference is **significant** or that the treatment has a significant effect. In this case, we reject the null hypothesis. If the mean difference is relatively small, the test statistic will have a low value. In this case, we conclude that the evidence from the sample is not sufficient, and we fail to reject the null hypothesis.

Typical Question:

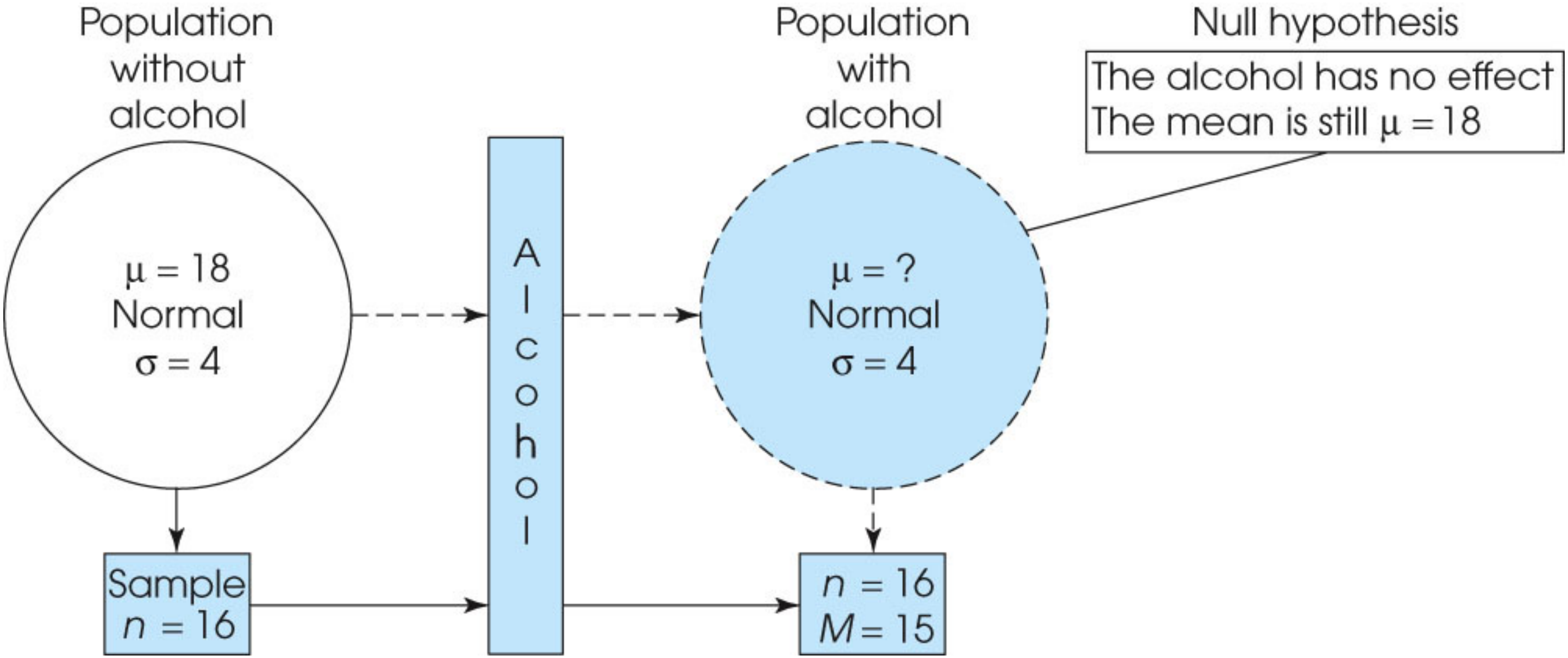
Do the data in the sample provide sufficient evidence to indicate that *one group is different from the other*?

Reasoning used:

Court trial analogy:

1. The court assumes that the accused is innocent until proven guilty **(H_0).**

2. The prosecution collects and presents available evidence in an attempt to **contradict the “not guilty” hypothesis (H_0)** and hence obtain a conviction.



The effect of alcohol on birth weight

Errors in Hypothesis Tests

- Just because the sample mean (following treatment) is different from the original population mean does not necessarily indicate that the treatment has caused a change.
- You should recall that there usually is some discrepancy between a sample mean and the population mean simply as a result of ***sampling error***.

Errors in Hypothesis Tests (cont.)

- Because the hypothesis test relies on sample data, and because sample data are not completely reliable, there is always the risk that misleading data will cause the hypothesis test to reach a wrong conclusion.
- Two types of error are possible.

Errors in Hypothesis Testing

e.g., H_0: μ = 0; H_1: μ = -1.7
i.e., H_1 states that μ is from a different distribution.

Type I error: α
Type II error: β
power （效力/统计检验力） = 1 – β

Reality:

Decision:	H_0 True	H_0 False
Fail to Reject H_0	No error $\Pr\left(\text{Accept } H_0 \mid H_0 \text{ True}\right)$ $= 1-\alpha = 0.95$	Type II error $\Pr\left(\text{Accept } H_0 \mid H_0 \text{ False}\right)$ $= \beta = ?$
Reject H_0	Type I error $\Pr\left(\text{Reject } H_0 \mid H_0 \text{ True}\right)$ $= \alpha = 0.05$ **You to avoid!**	No error $\Pr\left(\text{Reject } H_0 \mid H_0 \text{ False}\right)$ $= 1-\beta = ?$ **Your power!**

Type I Errors

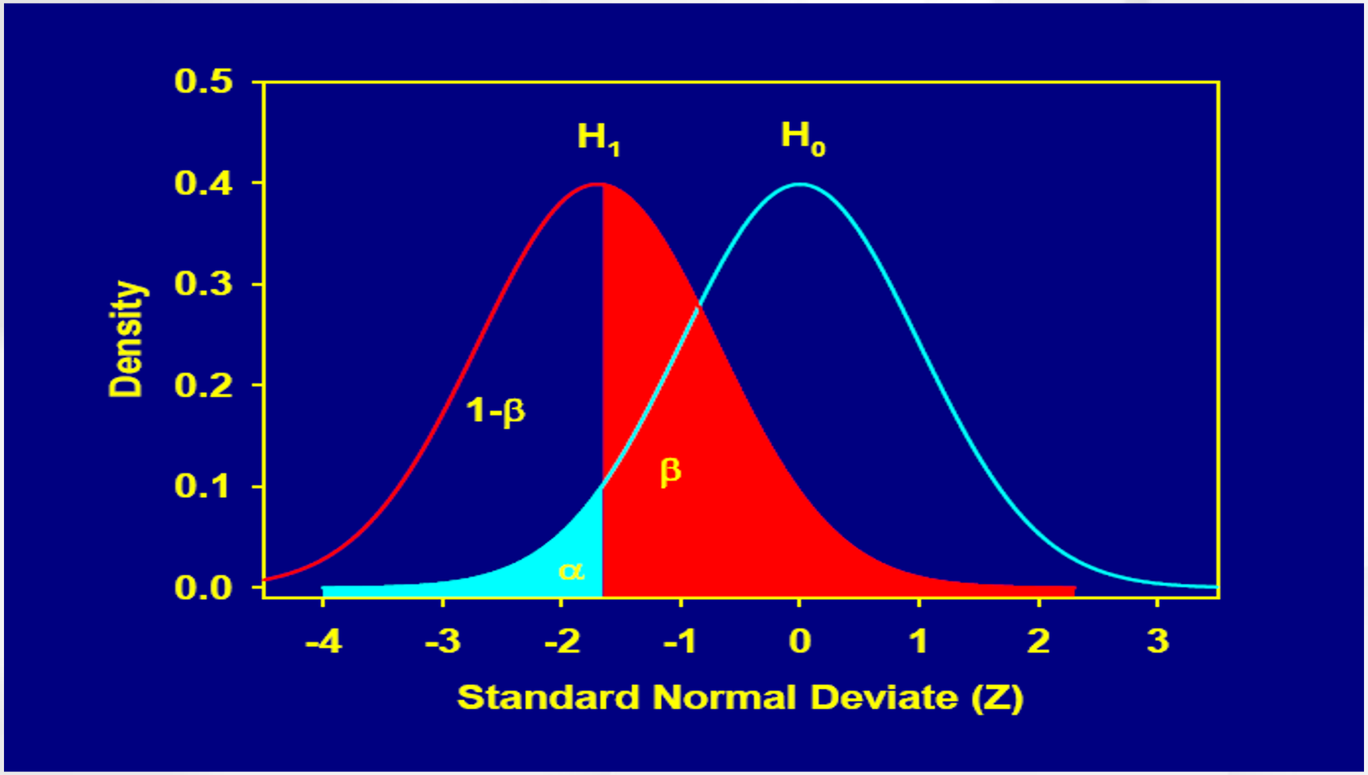
- A **Type I error** occurs when the sample data appear to show a treatment effect when, in fact, there is none.
- In this case the researcher will reject the null hypothesis and falsely conclude that the treatment has an effect.
- Type I errors are caused by unusual, unrepresentative samples. Just by chance the researcher selects an extreme sample with the result that the sample falls in the critical region even though the treatment has no effect.
- The hypothesis test is structured so that Type I errors are very unlikely; specifically, the probability of a Type I error is equal to the alpha level.

Type II Errors

- A **Type II error** occurs when the sample does not appear to have been affected by the treatment when, in fact, the treatment does have an effect.
- In this case, the researcher will fail to reject the null hypothesis and falsely conclude that the treatment does not have an effect.
- Type II errors are commonly the result of a very small treatment effect. Although the treatment does have an effect, it is not large enough to show up in the research study.

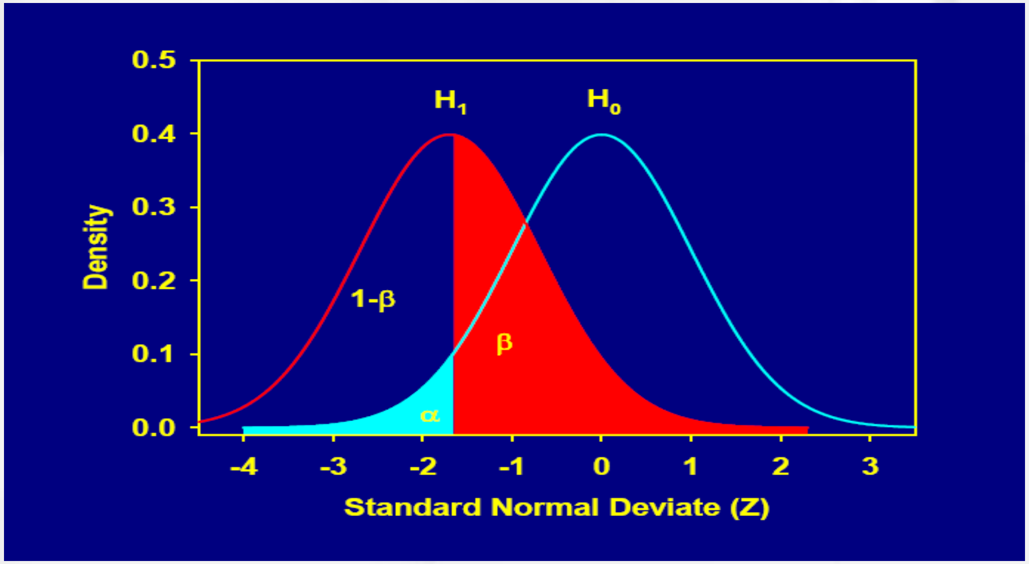
Performance Measures

- Type I error α: commonly referred to as the significance level of a test.
- power = 1 – β: the probability of rejecting H_0 hypothesis, given H_0 is indeed false.

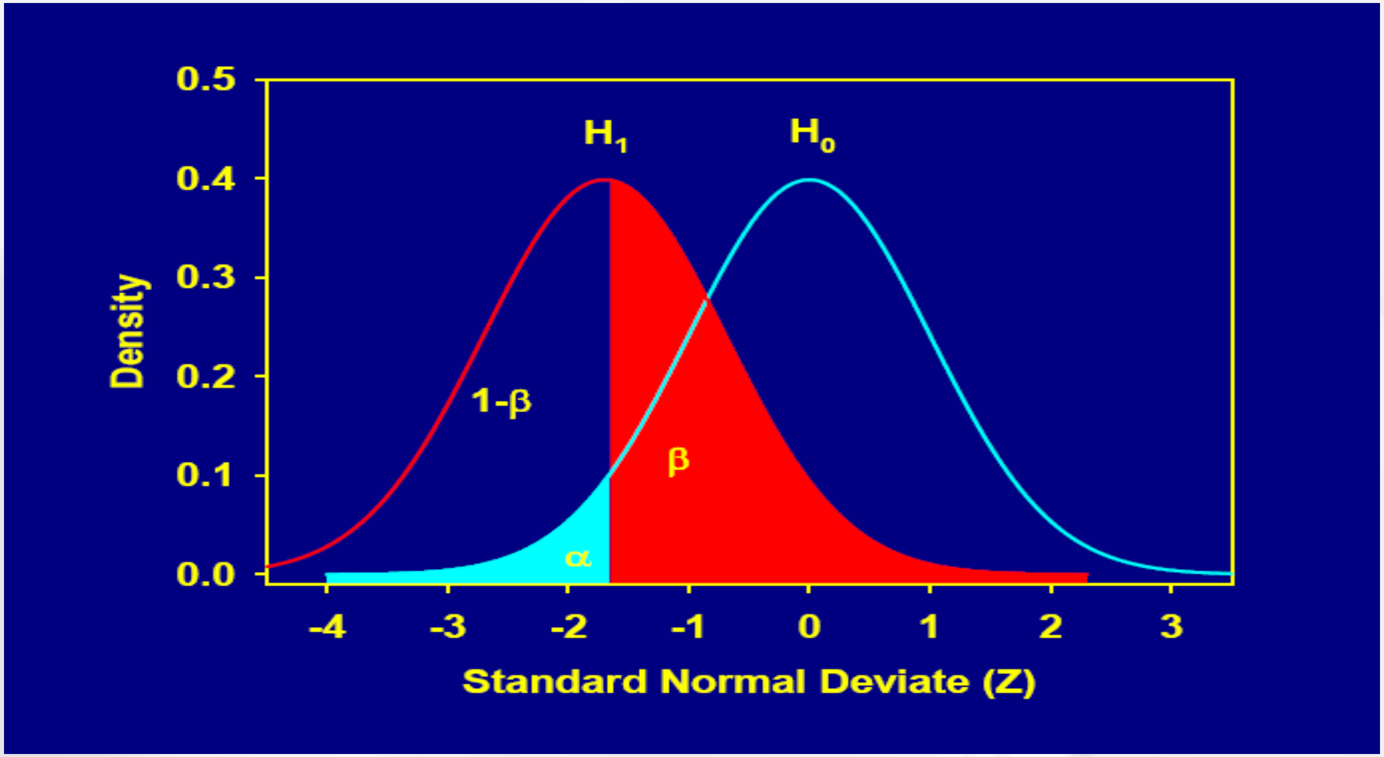


α and β are inversely related

- The probability of a Type I error, α, will increase as the size of the rejection region increases.
- Why not reduce the size of the rejection region and make α as small as possible?
- Unfortunately, decreasing α increases the probability of a type II error, β.



What to do?



Sample size:

- For a fixed sample size, α and β are inversely related.
- Increasing the sample size, provides more information directed at the decision, allowing β to be reduced as α is a pre-determined value ~~allowing α and β to both be reduced~~ (both H_0 and H_1 distributions become narrow!)

Convention:

Fix Level α =0.05, 0.01, etc; then choose a test that minimizes β; two-sided vs. one-sided; have enough samples.

Directional Tests

- When a research study predicts a specific direction for the treatment effect (increase or decrease), it is possible to incorporate the directional prediction into the hypothesis test.
- The result is called a **directional test** or a **one-tailed test**. A directional test includes the directional prediction in the statement of the hypotheses and in the location of the critical region.

Directional Tests (cont.)

- For example, if the original population has a mean of $\mu = 80$ and the treatment is predicted to ***increase*** the scores, then the null hypothesis would state that after treatment:

  H_0: $\mu = 80$ (there is no increase)

  H_1: $\mu > 80$ (there is an increase)

- In this case, the entire critical region would be located in ***the right-hand tail*** of the distribution because large values for M would demonstrate that there is an increase and would tend to reject the null hypothesis.
- If the test has no direction, the rejection region can be on either side. H_1: $\mu \neq 80$.

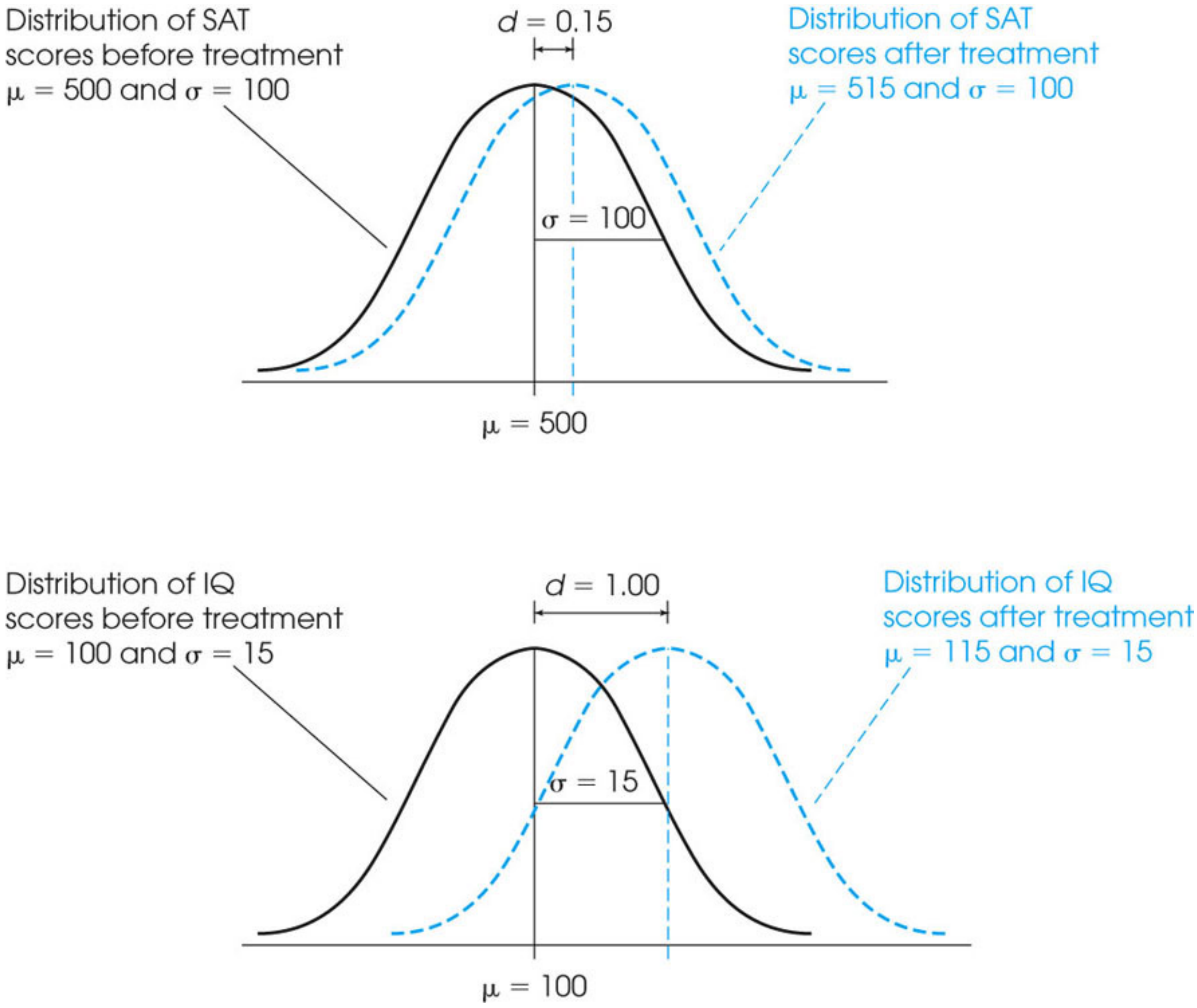
Measuring Effect Size

- A hypothesis test evaluates the ***statistical significance*** of the results from a research study.
- That is, the test determines whether or not it is likely that the obtained sample mean occurred without any contribution from a treatment effect.
- The hypothesis test is influenced not only by the size of the treatment effect but also by the size of the sample.
- Thus, even a very small effect can be significant if it is observed in a very large sample.

Measuring Effect Size

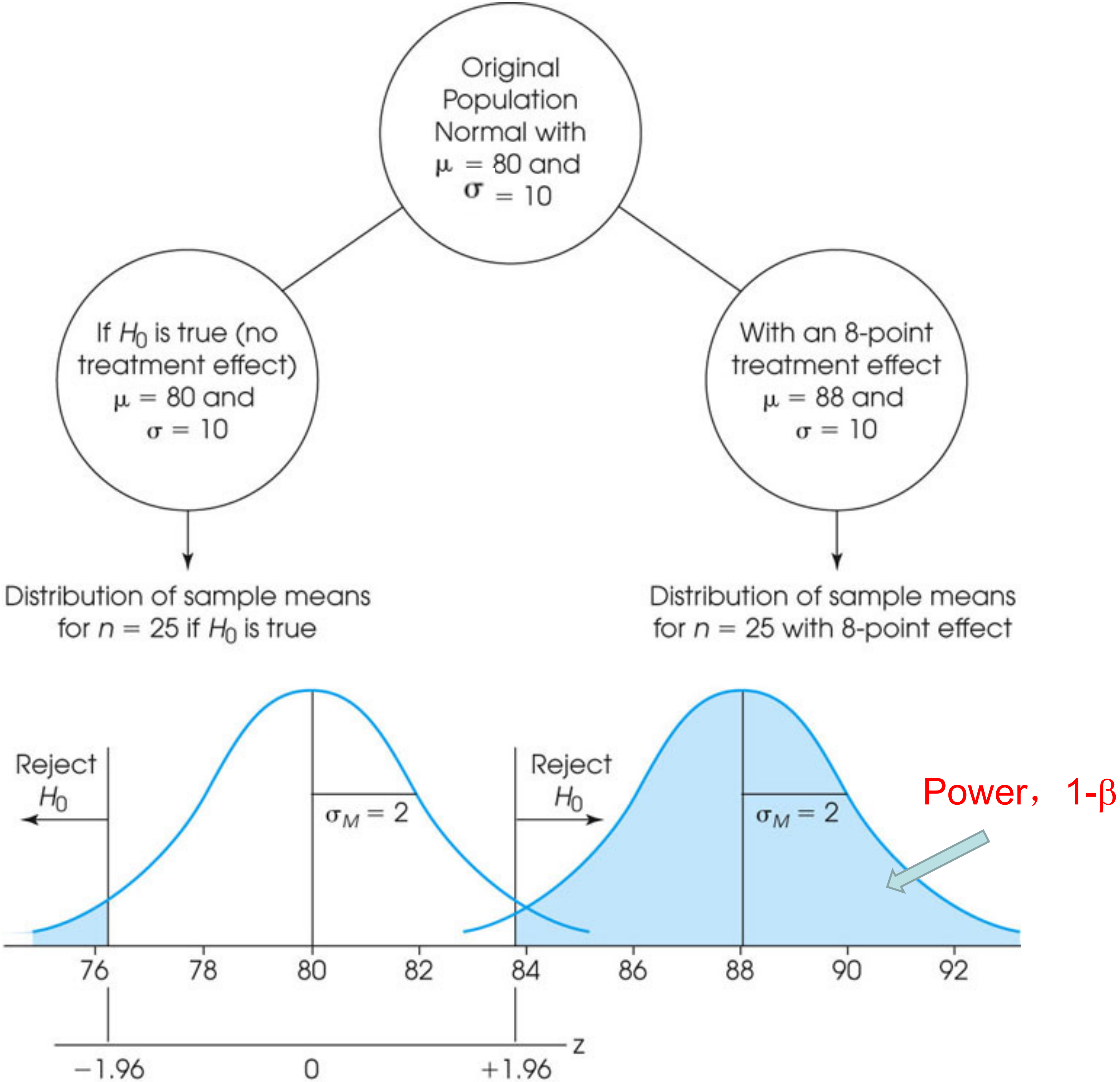
- Because a significant effect does not necessarily mean a large effect (but a large effect is more likely to be significant), it is recommended that the hypothesis test be accompanied by a measure of the **effect size**.
- We use **Cohen's d** as a standardized measure of effect size.
- Much like a z-score, **Cohen's d** measures the size of the mean difference in terms of the standard deviation.

The same difference of means, 15


Distribution of SAT scores before treatment
μ = 500 and σ = 100
d = 0.15
Distribution of SAT scores after treatment
μ = 515 and σ = 100
σ = 100
μ = 500
Distribution of IQ scores before treatment
μ = 100 and σ = 15
d = 1.00
Distribution of IQ scores after treatment
μ = 115 and σ = 15
σ = 15
μ = 100

Power of a Hypothesis Test

- The **power** of a hypothesis test is defined is the probability that the test will reject the null hypothesis when the treatment does have an effect.
- The power of a test depends on a variety of factors including the size of the treatment effect and the size of the sample.



Original Population Normal with $\mu = 80$ and $\sigma = 10$

If H_0 is true (no treatment effect) $\mu = 80$ and $\sigma = 10$

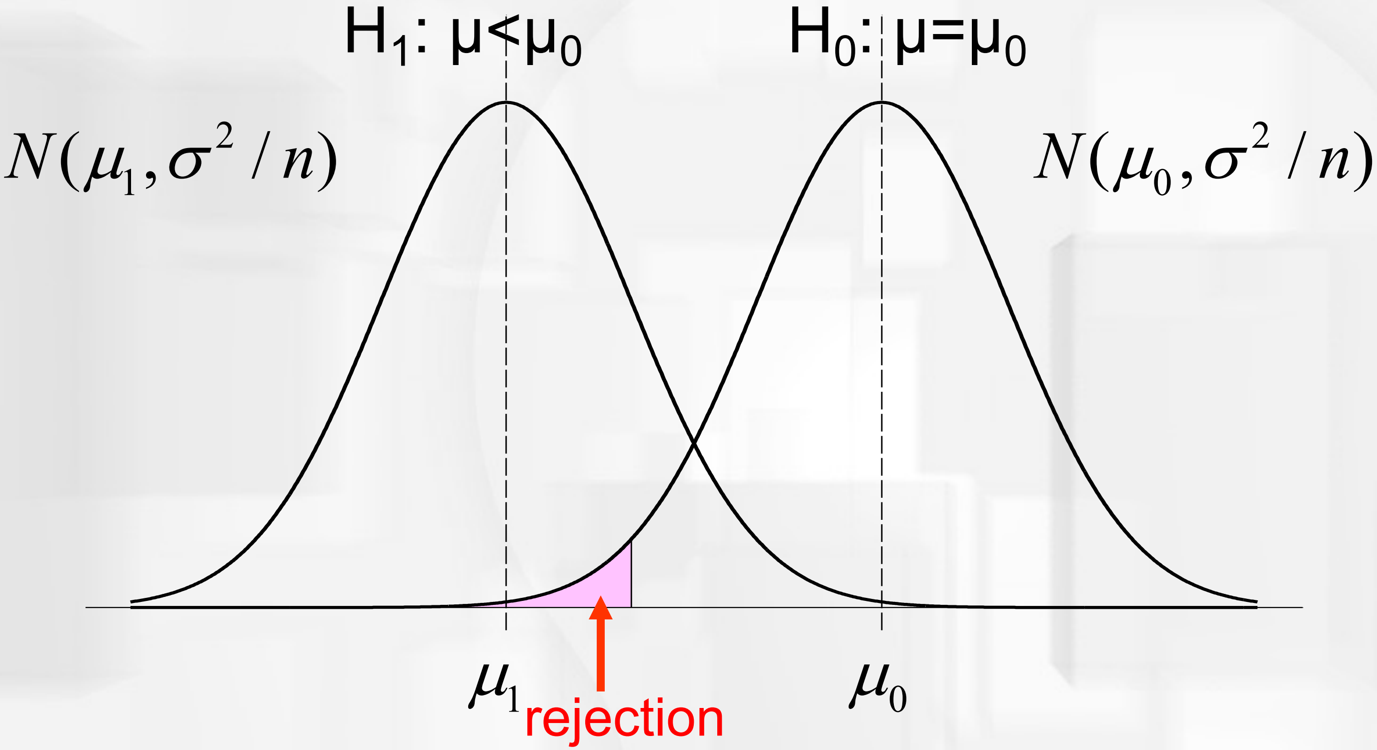
With an 8-point treatment effect $\mu = 88$ and $\sigma = 10$

Distribution of sample means for $n = 25$ if H_0 is true

Distribution of sample means for $n = 25$ with 8-point effect

Power analysis

Power Calculation
Z test as an example



$$\bar{X} < \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \quad or \quad \bar{X} < \mu_0 - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$$

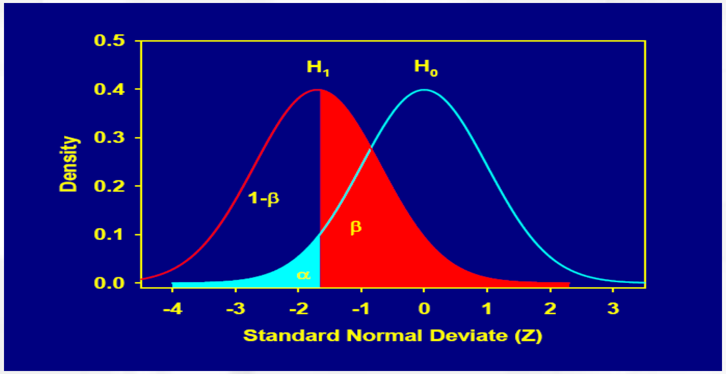
- The hypothesis testing does not depend on the alternative mean chosen, μ_1;
- However, the power of the test depends on μ_1.

Reject H_0 if: $\bar{X} < \mu_0 + z_\alpha \cdot \sigma \big/ \sqrt{n}$

$$Power = \Pr(reject\ \ H_0 \mid H_1\ \ true) = 1 - \beta$$

$$= \Pr(\bar{X} < \mu_0 + z_\alpha \cdot \sigma / \sqrt{n} \mid \underline{\mu = \mu_1})$$

$Note: under\ \ H_1:\qquad \bar{X} \sim N(\mu_1, \sigma^2 / n)$



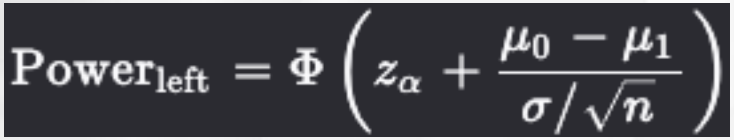
Now Standardize:

$$1 - \beta = Power = \Pr\left(\frac{\bar{X} - \mu_1}{\sigma \big/ \sqrt{n}} < \frac{\mu_0 - \mu_1}{\sigma \big/ \sqrt{n}} + z_\alpha \right)$$

$$= \Pr\left(Z < \frac{\mu_0 - \mu_1}{\sigma \big/ \sqrt{n}} + z_\alpha \right) = \phi\left(z_\alpha + \frac{\mu_0 - \mu_1}{\sigma \big/ \sqrt{n}} \right)$$

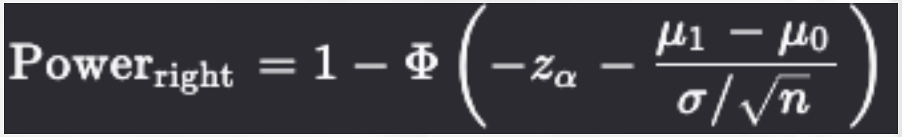
Note: z_α is the critical value under a standard normal distribution, a negative value.

Left-tailed Z test, $\mu_1 < \mu_0$



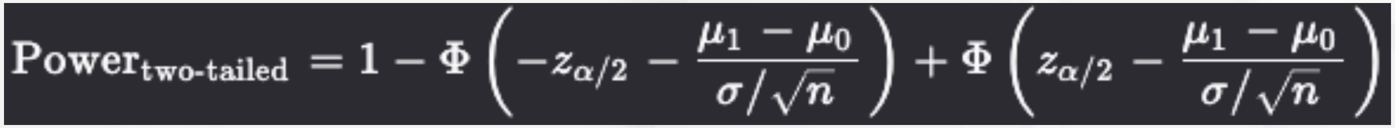
$$\text{Power}_{\text{left}} = \Phi\left(z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right)$$

Right-tailed Z test, $\mu_1 > \mu_0$



$$\text{Power}_{\text{right}} = 1 - \Phi\left(-z_\alpha - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$$

Two-tailed Z test, $\mu_1 \neq \mu_0$



$$\text{Power}_{\text{two-tailed}} = 1 - \Phi\left(-z_{\alpha/2} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right) + \Phi\left(z_{\alpha/2} - \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}\right)$$

Example for Power Calculation

Power to detect a 5 Oz difference

$$H_0 : \mu = \mu_0, \sigma = \sigma_0 \ vs.\ H_1 : \mu = \mu_1 < \mu_0, \sigma = \sigma_0$$

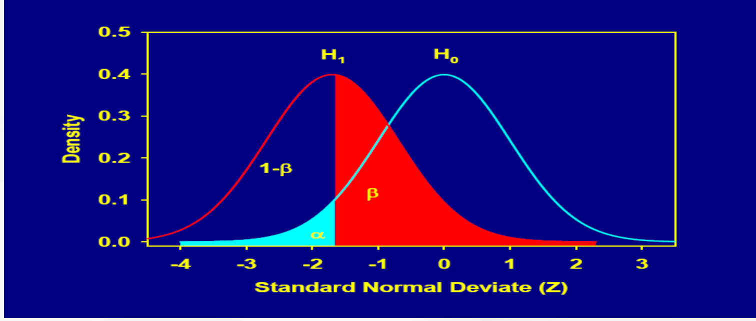
$$\mu_0 = 120\ oz.,\ \mu_1 = 115\ oz.,\ \alpha = 0.05,\ n = 100,\ \sigma = 25\ oz.$$

$$Power = \phi\left(z_{0.05} + \frac{120 - 115}{25 \big/ \sqrt{100}} \right) = \phi\left(-1.645 + \frac{5(10)}{25} \right)$$

$$= \phi(0.355) = 0.639$$

There is a 64% chance of detecting a significant difference
that is really there using a 5% significance level
with this sample size

Factors Affecting the Power



1) Significance level, α.

As $\alpha \downarrow$, $z_\alpha \downarrow$, power $\downarrow$.

Significance level

2) As alternative means become farther from null mean, power $\uparrow$.

That is, $|\mu_0 - \mu_1| \uparrow$, then power $\uparrow$

Effect size

3) If SD $\uparrow$, then power $\downarrow$

Measurement error

4) If sample size $\uparrow$, then power $\uparrow$

Sample Size

$$Power = \phi\left(z_\alpha + \frac{|\mu_1 - \mu_0|}{\sigma / \sqrt{n}} \right)$$

Note: This formula is for the two types of one-tailed Z test

Sample size determination

For planning purposes we sometimes need an appropriate sample size prior to the beginning of a study.

Problem:

Given that a significance test will be conducted at the level of α, the null mean is μ_0 and that the alternative mean is $\mu_1(\mu_1 < \mu_0)$. What sample size is needed to be able to detect a significant difference between μ_0 and μ_1 with probability 1–β?

H_0: $\mu=\mu_0$ vs. H_1: $\mu=\mu_1<\mu_0$

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

Solve for *n* in terms of α, β, $\mu_1 – \mu_0$ and σ.

$$z_\alpha + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}} = z_{1-\beta}$$

$$\Rightarrow z_{1-\beta} - z_\alpha = \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}}$$

$$\Rightarrow n = \frac{(z_{1-\beta}+z_{1-\alpha})^2\sigma^2}{(\mu_0-\mu_1)^2}$$

Factors affecting sample sizes

$$n = \frac{\sigma^2 \left(z_{1-\beta} + z_{1-\alpha}\right)^2}{\left(\mu_0 - \mu_1\right)^2}$$

1. As σ^2 ↑, n ↑
2. As α ↓, n ↑
3. As 1-β ↑, n ↑
4. As $\delta = |\mu_1 - \mu_0|$ ↑, n ↓
5. As $\delta = |\mu_1 - \mu_0|$ ↓, n ↑

Sample Size Determination Example

Suppose that μ_0=120 oz, μ_1=115 oz, σ^2=625, α=0.05, 1-β =0.8. Compute the appropriate sample size needed to conduct this test.

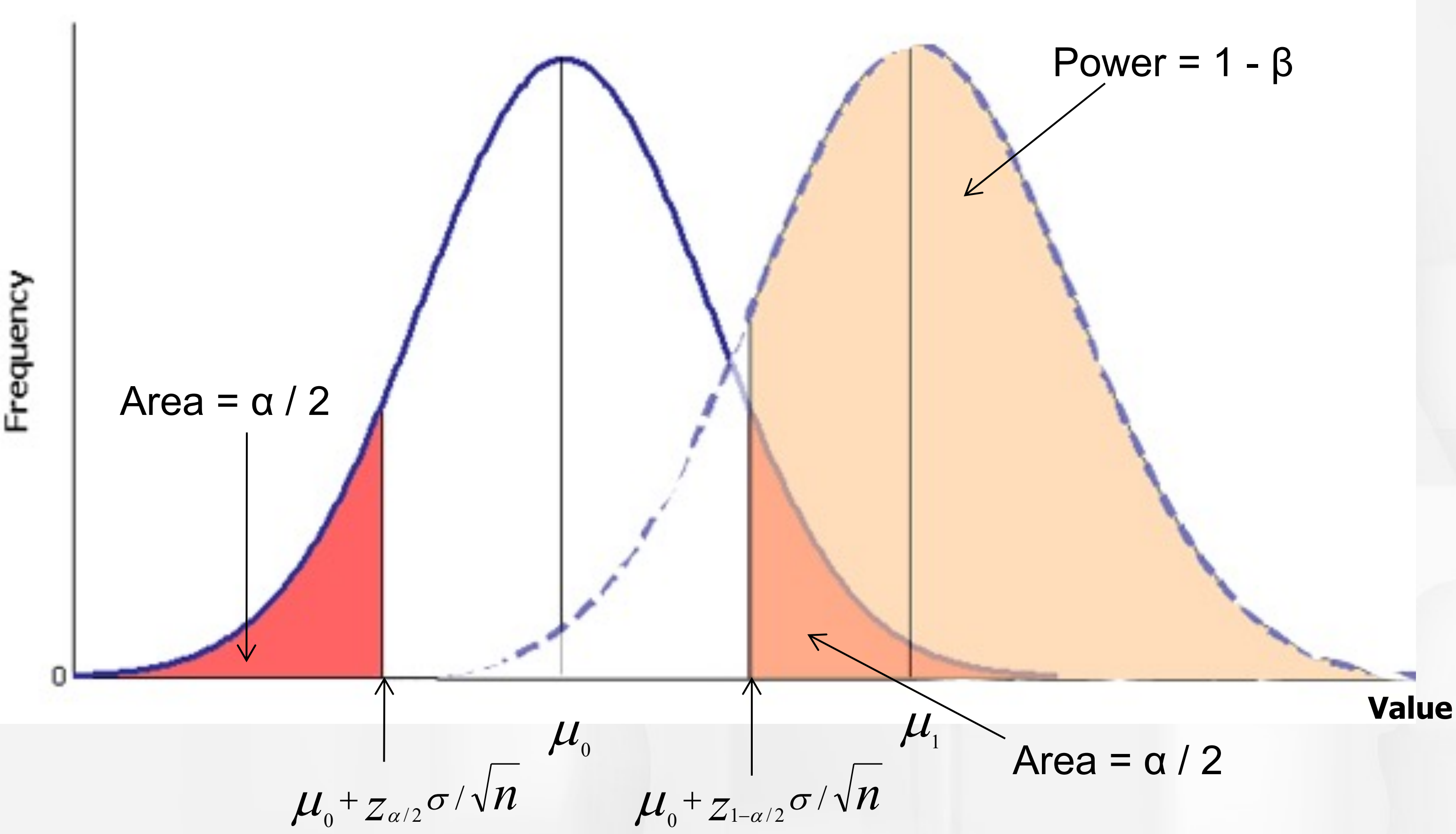
$$z_{1-\alpha} = z_{0.95} = 1.645,\ z_{1-\beta} = z_{0.8} = 0.84$$

$$n = \frac{(z_{1-\beta} + z_{1-\alpha})^2 \sigma^2}{(\mu_0 - \mu_1)^2} = \frac{(1.645 + 0.84)^2 625}{5^2} = 154.4$$

A sample size of 155 (round up) is required to have an 80% chance of detecting a significant difference at the 5% level, if the alternative mean is that 115 oz is the true mean.

One sample two-sided Z-test

H_0: $\mu = \mu_0$ vs. H_1: $\mu \neq \mu_0$



Reject H_0 if $\bar{X} > \mu_0 + z_{1-\alpha/2} \cdot \sigma / \sqrt{n}$ or $\bar{X} < \mu_0 + z_{\alpha/2} \cdot \sigma / \sqrt{n}$

Reject H_0 if $Z_{obs} > z_{1-\alpha/2}$ or $Z_{obs} < z_{\alpha/2}$, where $Z_{obs} = \dfrac{\bar{X} - \mu_0}{\sigma_0 / \sqrt{n}}$

H_0: μ = μ_0 vs. H_1: μ≠μ_0

What is the p-value of the two-sided Z-test?

$$p - value = 2 * \Pr\left(Z > z_{obs}\right) \; if \;\; z_{obs} > 0$$

$$p - value = 2 * \Pr\left(Z < z_{obs}\right) \; if \;\; z_{obs} < 0$$

What is the power of the two-sided Z-test?

$$Power = \Pr\left(\bar{X} > \mu_0 + z_{1-\alpha/2} \cdot {}^{\sigma}\!/_{\sqrt{n}} \middle| H_1 \right) + \Pr\left(\bar{X} < \mu_0 + z_{\alpha/2} \cdot {}^{\sigma}\!/_{\sqrt{n}} \middle| H_1 \right)$$

$$= \Phi\left(z_{\alpha/2} + \frac{\mu_1 - \mu_0}{{}^{\sigma}\!/_{\sqrt{n}}} \right) + \Phi\left(z_{\alpha/2} - \frac{\mu_1 - \mu_0}{{}^{\sigma}\!/_{\sqrt{n}}} \right)$$

$$\approx \Phi\left(z_{\alpha/2} + \frac{\left|\mu_1 - \mu_0\right|}{{}^{\sigma}\!/_{\sqrt{n}}} \right)$$

Sample-Size Estimation When Testing for the Mean of a Normal Distribution (Two-Sided Alternative)

Suppose we wish to test H_0: μ = μ_0 versus H_1: μ ≠ μ_0 , where the data are normally distributed with mean μ and known variance σ^2. The **sample size** needed to conduct a two-sided test with significance level α and power 1 – β is

$$n = \frac{\sigma^2 \left(z_{1-\beta} + z_{1-\alpha/2} \right)^2}{\left(\mu_0 - \mu_1 \right)^2}$$

For one-tailed test:

$$n = \frac{\sigma^2 \left(z_{1-\beta} + z_{1-\alpha} \right)^2}{\left(\mu_0 - \mu_1 \right)^2}$$

Summary for one-sample Z-test

- $H_1 : \mu > \mu_o$ reject H_0 if $\bar{x} > \mu_0 + z_{1-\alpha} \cdot \sigma / \sqrt{n}$

=> $Power = \Phi\left(z_\alpha + \frac{|\mu_1 - \mu_0|}{\sigma/\sqrt{n}} \right)$ $=> n = \frac{(z_{1-\beta} + z_{1-\alpha})^2 \sigma^2}{(\mu_0 - \mu_1)^2}$

- $H_1 : \mu < \mu_o$ reject H_0 if $\bar{x} < \mu_0 + z_\alpha \cdot \sigma / \sqrt{n}$

=> $Power = \Phi\left(z_\alpha + \frac{|\mu_1 - \mu_0|}{\sigma/\sqrt{n}} \right)$ $=> n = \frac{(z_{1-\beta} + z_{1-\alpha})^2 \sigma^2}{(\mu_0 - \mu_1)^2}$

- $H_1 : \mu \neq \mu_o$ reject H_0 if

$\bar{x} > \mu_0 + z_{1-\alpha/2} \cdot \sigma / \sqrt{n}$ *or* $\bar{x} < \mu_0 + z_{\alpha/2} \cdot \sigma / \sqrt{n}$

=> $Power \approx \Phi\left(z_{\alpha/2} + \frac{|\mu_1 - \mu_0|}{\sigma/\sqrt{n}} \right)$ $=> n = \frac{(z_{1-\beta} + z_{1-\alpha/2})^2 \sigma^2}{(\mu_0 - \mu_1)^2}$