Lecture 10

Power

The ability of a test to detect that H_0 is false is measured by the probability that the test will reject H_0 when an alternative is true. The higher this probability is, the more sensitive the test is.

Definition: The probability that a fixed level α test will reject H_0 when H_0 is false is called the **power** of the test.

- A powerful test has a large probability of rejecting H_0 when it is false.
- We want a powerful test!

Three steps to find the power of the test:

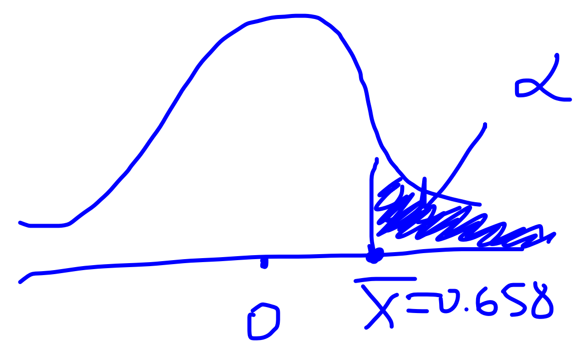
- State H_0, H_a, the particular alternative we want to detect, and the significance level α.
- Find the values of $\bar{x}$ (or other estimates) that will lead to reject H_0.
- Calculate the probability of observing these values of $\bar{x}$ when the alternative is true.

Example: Can a 6-month exercise program increase the total body bone mineral content (TBBMC) of young women? A team of researchers is planning a study to examine this question. Based on the results of a previous study, they are willing to assume that $\boxed{\sigma = 2}$ for the percent change in TBBMC over the 6-month period. A change in TBBMC of 1% would be considered important, and the researchers would like to have a reasonable chance of detecting a change this large or larger. Is 25 subjects a large enough sample for this project?

μ_a

Step 1: $H_0: \mu = 0$ $\qquad \mu_a = 1\%$

$H_a: \mu > 0$ $\qquad \alpha = 0.05$

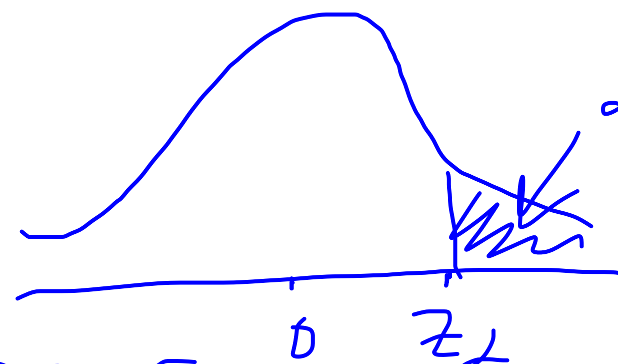


Step 2:

We reject H_0 when

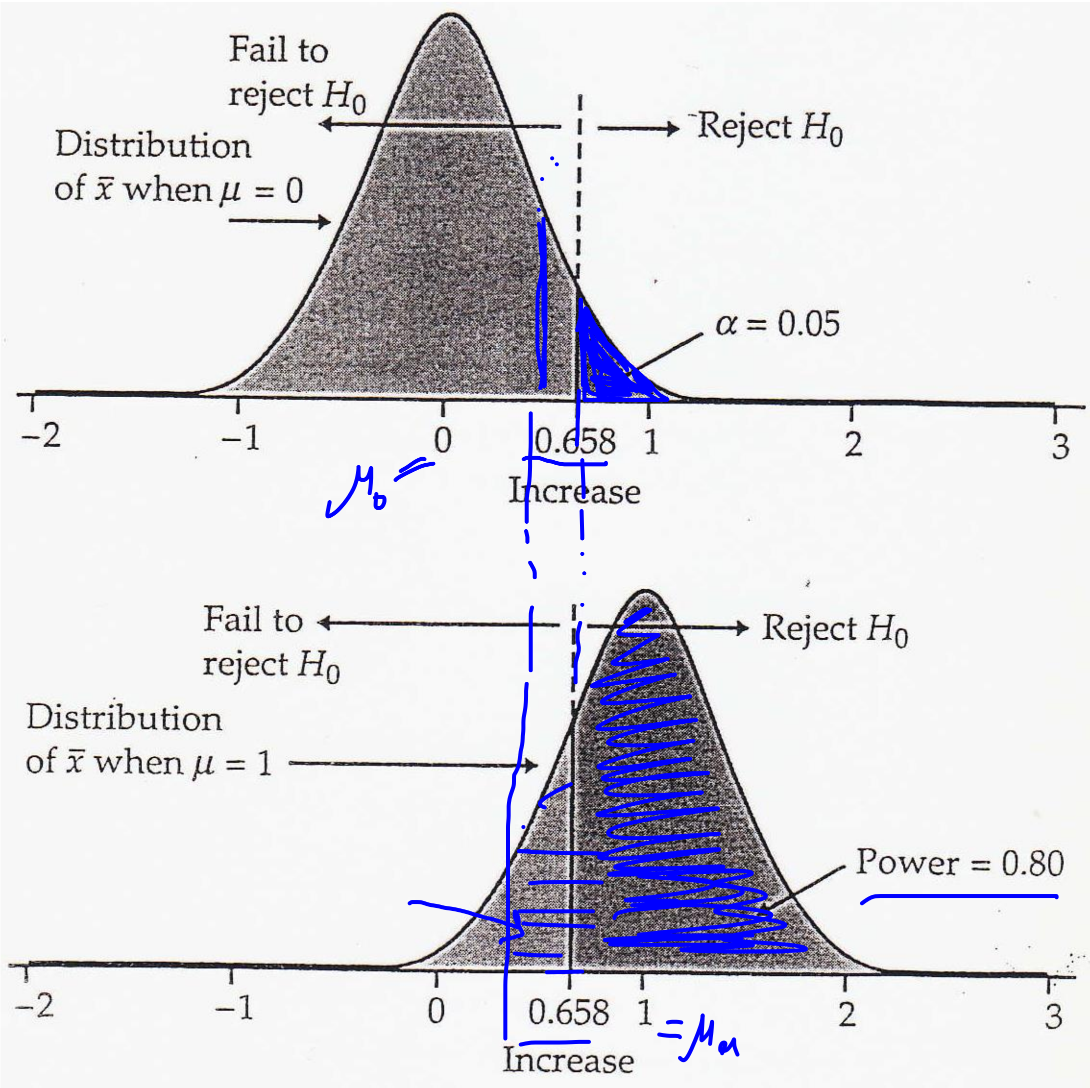
$$\frac{\bar{x} - 0}{2/\sqrt{25}} \geq z_\alpha = 1.645$$

$$\bar{x} \geq 1.645 \cdot \frac{2}{\sqrt{25}} = 0.658$$



$P(\bar{x} \geq 0.658 \text{ when } \mu = 0) = 0.05$

Step 3: $P(\bar{X} \geq 0.658 \text{ when } \mu = 1)$



$$= P\left(\frac{\bar{X}-1}{2/\sqrt{25}} \geq \frac{0.658-1}{2/\sqrt{25}}\right)$$

$$= P(Z \geq -0.855)$$

$$Z \sim N(0,1)$$

$$= 0.80$$

Power = 80%

The test rejects H_0 80% of the time if the true value $\mu = 1$.

Example: Power of the pharmaceutical product test (from the last lecture):

Step 1 :

$$H_0: \mu = 0.86$$

$$H_a: \mu \neq 0.86$$

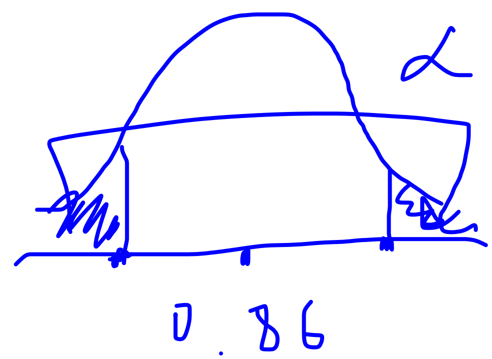
$$\alpha = 0.01, \qquad \sigma = 0.0068, \qquad n = 3$$

What is the power of the test against the specific alternative $\mu = 0.845$?

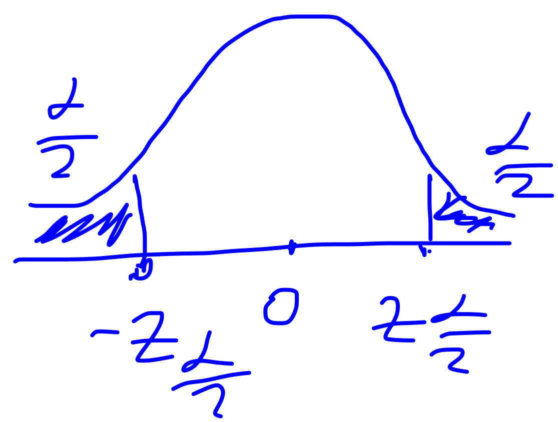
Step 2 :

We will reject H_0 when

$$\frac{\bar{X} - 0.86}{0.0068/\sqrt{3}} \geq z_{\frac{\alpha}{2}} = 2.575$$



$$\frac{\bar{X} - 0.86}{0.0068/\sqrt{3}} \leq -z_{\frac{\alpha}{2}} = -2.575$$



$$\Rightarrow \quad \bar{X} \geq 0.87$$

$$\bar{X} \leq 0.85$$

Step3: $P(\bar{X} \geq 0.87 \text{ when } \mu = 0.845)$

$$= P\left(\frac{\bar{X} - 0.845}{0.0068/\sqrt{3}} \geq \frac{0.87 - 0.845}{0.0068/\sqrt{3}}\right)$$

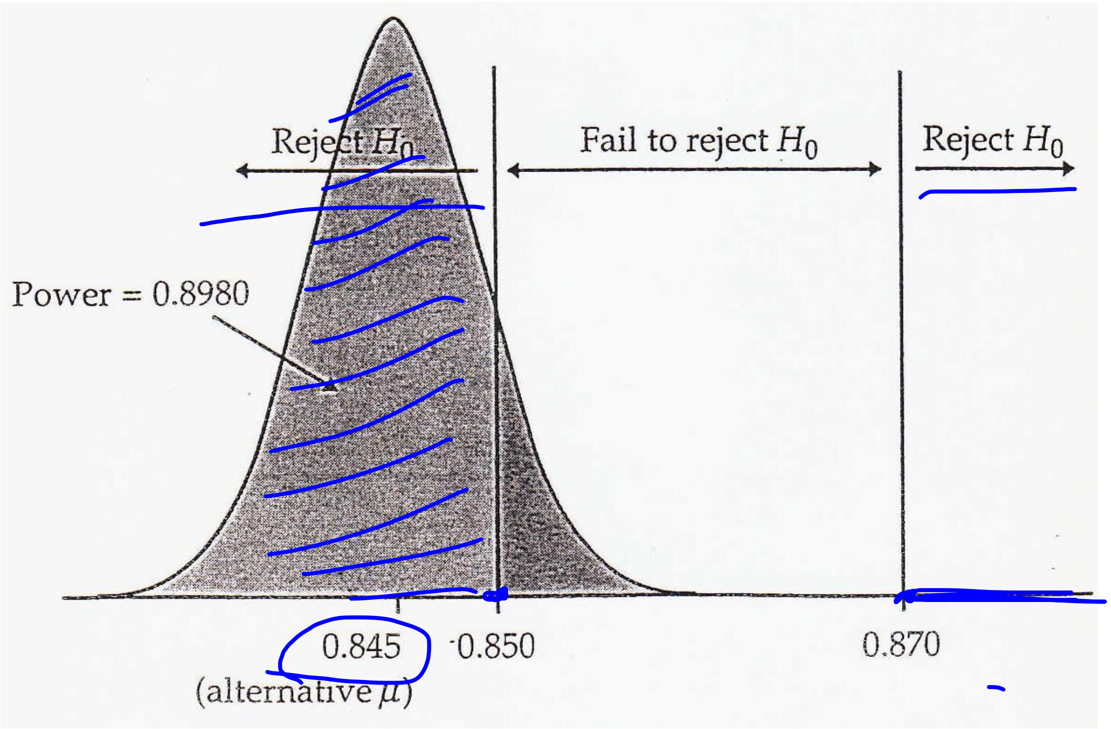
$$= P(Z \geq 6.37) \approx 0$$

$P(\bar{X} \leq 0.85 \text{ when } \mu = 0.845)$

$$= P\left(Z \leq \frac{0.85 - 0.845}{0.0068/\sqrt{3}}\right)$$

$$= P(Z \leq 1.27) \approx$$

$$\approx 0.8980$$



Power $= 0 + 0.8980$

$\approx 90\%$

How to increase the power?

- Increase α
- Consider an alternative that is farther away from μ_0
- Increase the sample size
- Decrease σ

Decision Errors

When we perform a statistical test we hope that our decision will be correct, but sometimes it will be wrong. There are two possible errors that can be made in hypothesis test.

<u>Definition</u>: The error made by rejecting the null hypothesis H_0 when in fact H_0 is true is called a **type I error**.
The error made by failing to reject the null hypothesis H_0 when in fact H_0 is false is called a **type II error**.

		Truth about the population	
		H_0 true	H_a true
Decision based on sample	Reject H_0	Type I error	Correct decision
	Accept H_0	Correct decision	Type II error

Example: When a parachute is inspected, the inspector is looking for anything that might indicate that the parachute might not open.

- H_0: The parachute will open
- H_a: The parachute will not open

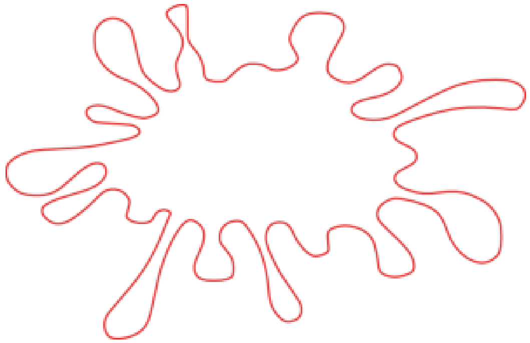
Which error is worse?

Type I Error: We conclude the parachute will not open when in fact, it will.

Consequences: The parachute will be rejected, and a new one put in its place. Money will be spend needlessly, and a perfectly good parachute will be wasted. But the parachutist is safe.

Type II Error: We conclude parachute will open when in fact it will not.

Consequences: Splat!



Example: Suppose that you have been put on trial for murder.

- H_0: You are innocent
- H_a: You are guilty

Which of the two errors is more serious?

Type I Error: You are found guilty of a murder that you did not commit.

Consequences: An innocent person will be sent to a jail.

Type II Error: You are guilty but are found not guilty.

Consequences: A murderer is on the loose!

Error Probabilities

	meets standards	does not
reject lot	Type I error	correct
accept lot	correct	Type II error

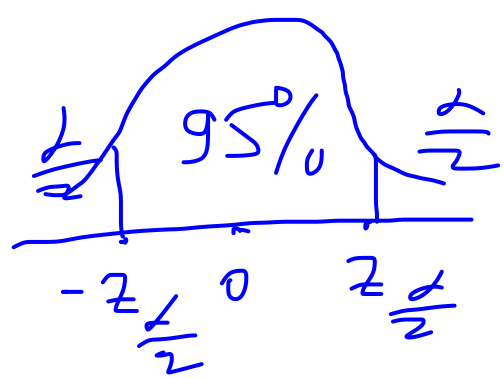
Example: The mean outer diameter of a skateboard bearing is supposed to be 22.000 millimeters (mm). The outer diameters vary Normally with standard deviation $\sigma = 0.010$ mm. When a lot of bearings arrives, the skateboard manufacturer takes an SRS of 5 bearings from the lot and measures their outer diameters. The manufacturer rejects the bearings if the sample mean diameter is significantly different from 22 at the 5% significance level.

H_0: lot meets standards or $\mu = 22$

H_a: does not $\mu \neq 22$

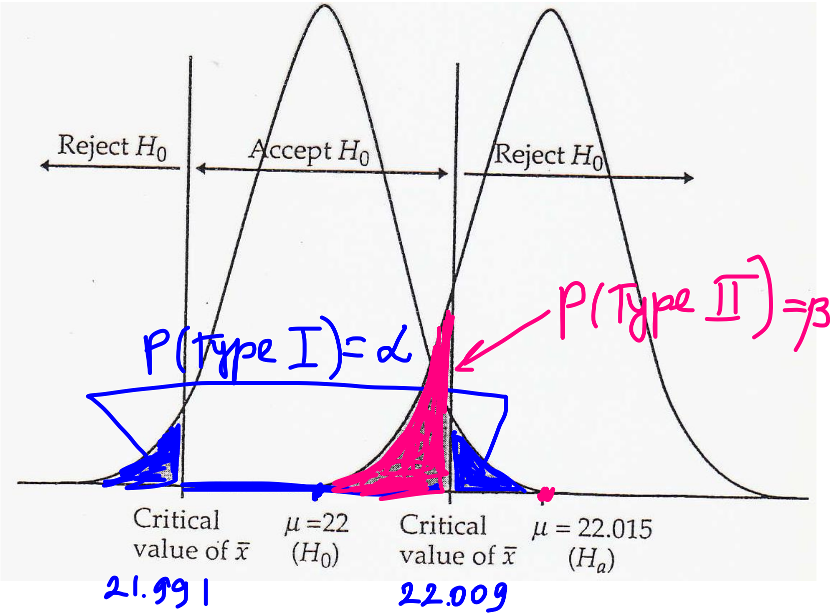
$$\frac{\bar{x} - 22}{0.01/\sqrt{5}} \geq 1.96 \Rightarrow \bar{x} \geq 22.009$$

$$\leq -1.96 \Rightarrow \bar{x} \leq 21.991$$



Suppose the producer and the manufacturer agree that a lot of bearings with mean 0.015 mm away from 22 should be rejected.

$\mu_a = 22.015$



$P(\text{Type I error})$

$= P(\text{reject } H_0 \text{ when } \mu = 22)$

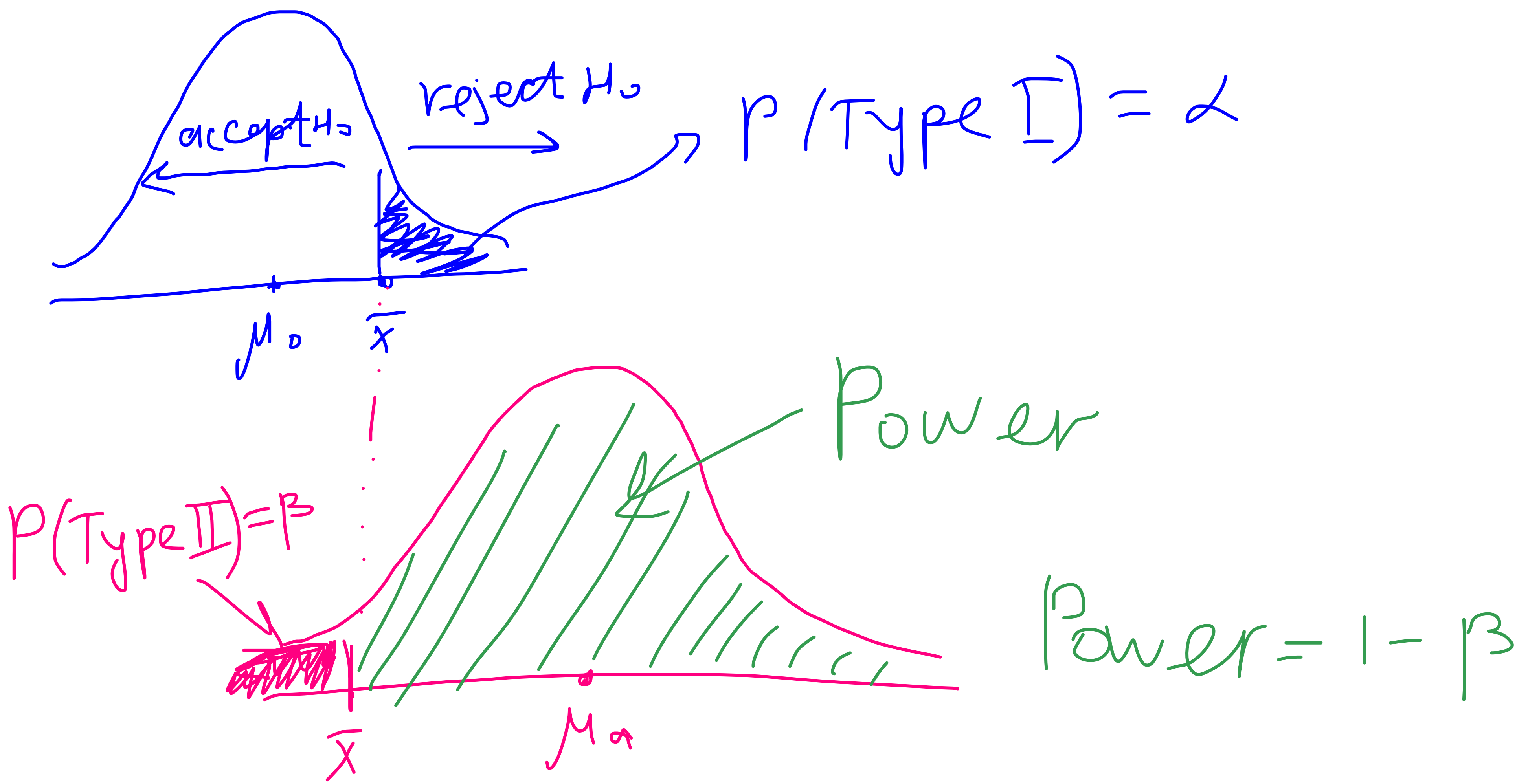
$= P(\bar{X} \geq 22.009$ or $\bar{X} \leq 21.991$ when $\mu = 22)$

$= \alpha$

$P(\text{Type II error}) = P(\text{accept } H_0 \text{ when } \mu = 22.015)$

$= P(21.991 \leq \bar{X} \leq 22.009 \text{ when } \mu = 22.015)$

Significance and Type I error: The significance level α of any fixed level test is the probability of a Type I error. That is, α is the probability that the test will reject H_0 when H_0 is in fact true.



Power and Type II error: The power of a fixed level test to detect a particular alternative is 1 minus the probability of a Type II error for that alternative (denoted by β).

Tests for a Population Mean (σ is unknown)

t distribution

Let $X_1, X_2, \ldots, X_n$ be an SRS from $N(\mu, \sigma)$, μ, σ are both unknown.

Then $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

$$s = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

We use *s* (sample standard deviation) to estimate σ.

Definition: When the standard deviation of a statistic is estimated from the data, the result is called the **standard error** of the statistic. The standard error of the sample mean is

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

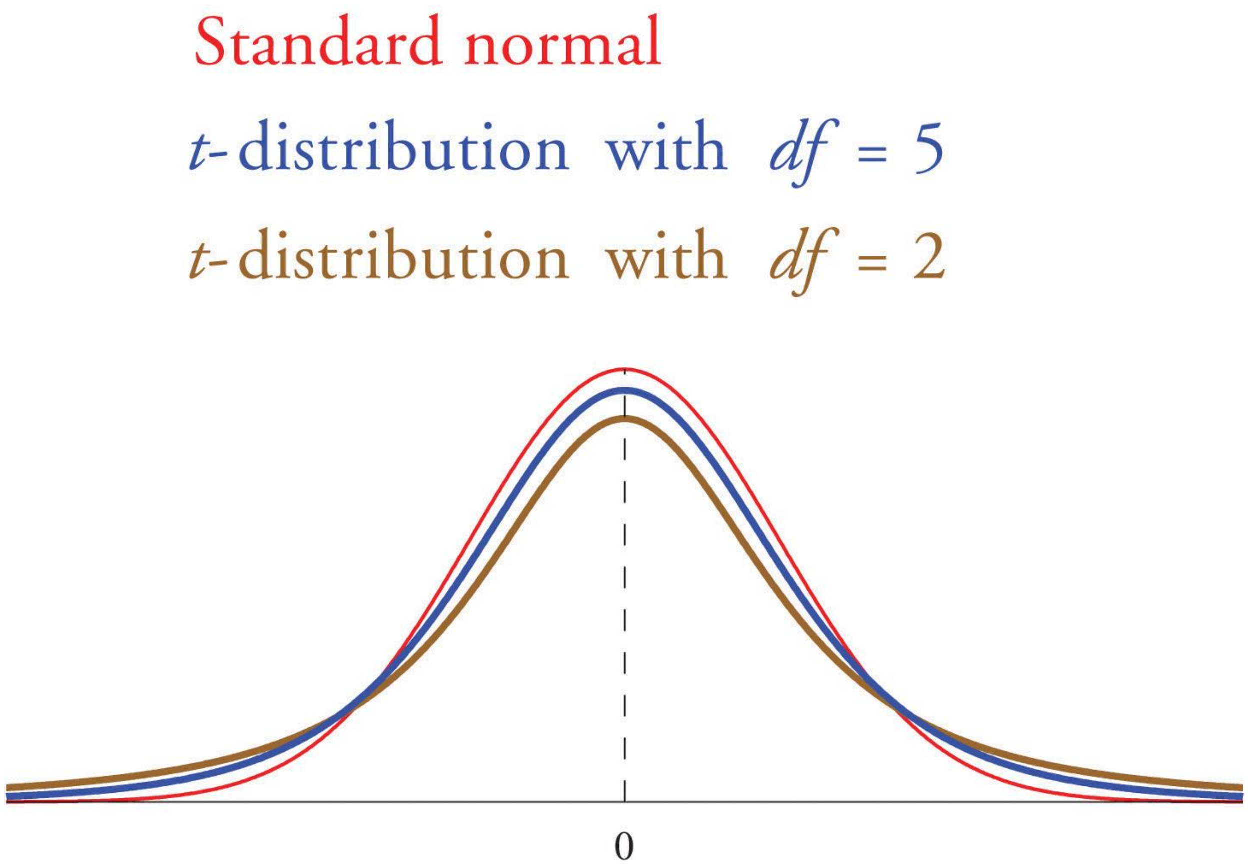
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \nsim N(0,1) \sim t_{n-1}$$

Definition: Suppose that an SRS of size n is drawn from an $N(\mu, \sigma)$ population. Then the **one-sample *t* statistic**

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

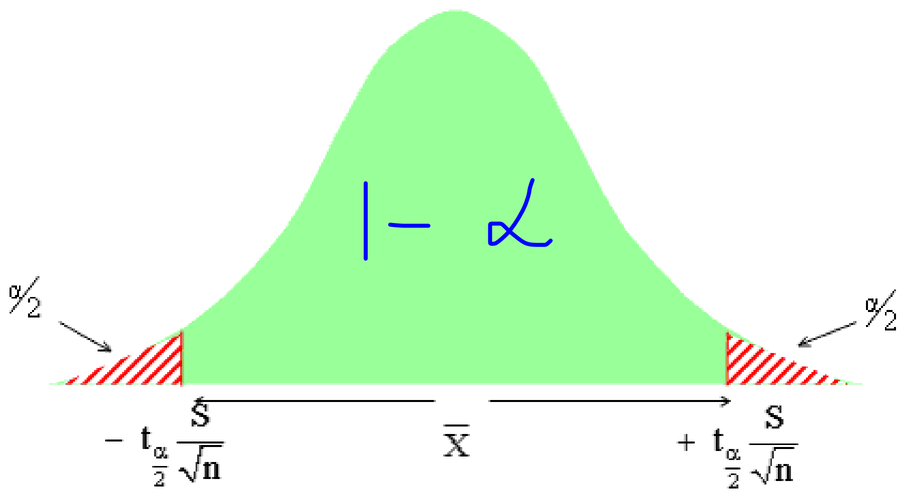
has the ***t* distribution (Student's *t*-distribution)** with n - 1 **degrees of freedom**.



<u>One-Sample *t* CI</u>: Suppose that an SRS of size *n* is drawn from a population having unknown mean μ. A 100(1-α)% confidence interval for μ is

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \qquad \bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Where $t_{\alpha/2}$ is the value for the t_{n-1} density curve with area 1-α between $-t_{\alpha/2}$ and $t_{\alpha/2}$.



The quantity

$$ME = t_{\alpha/2} \frac{s}{\sqrt{n}}$$

is the **margin of error**. This interval is exact when the population distribution is Normal and is approximately correct for large *n* in other cases.

Example: Founded in 1998, Telephia provides a wide variety of information on cellular phone use.

TELEPHIA®

In 2006, Telephia reported that, on average, United Kingdom (U.K.) subscribers with third-generation technology (3G) phones spent an average of 8.3 hours per month listening to full-track music on their cell phones.

Suppose we want to determine a 95% CI for the U.S. average and draw the following random sample of size 8 from the U.S. population of 3G subscribers:

5 6 0 4 11 9 2 3 (need to check Normality

Q-Q plot)

The sample mean is $\bar{x} = 5$ and the standard deviation $s = 3.63$ with degrees of freedom $n - 1 = 7$.

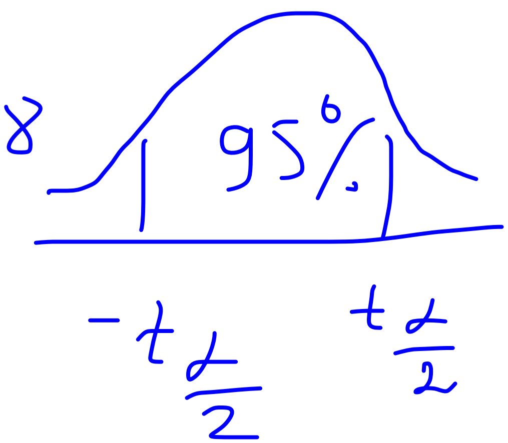
5   6   0   4   11   9   2   3

$\bar{x} = 5$ $s = 3.63$ $n - 1 = 7$

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{3.63}{\sqrt{8}} = 1.28$$

95% CI: $t_{\frac{\alpha}{2}} = 2.365$

$$\bar{x} \pm t_{\frac{\alpha}{2}} SE_{\bar{x}} = 5 \pm 2.365 \cdot 1.28$$



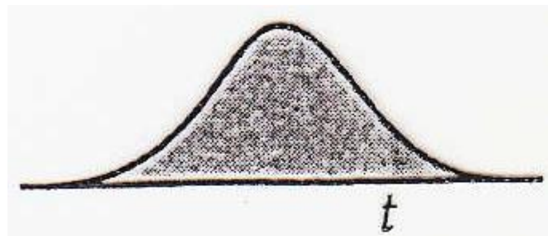
$$= (2, 8)$$

<u>Conclusion</u>: We are 95% confident that U.S. population spends on average between 2 and 8 hours listening music on cell phones.

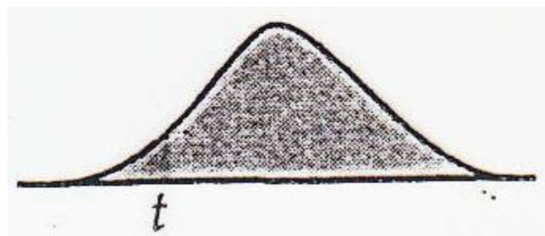
One-Sample t Test: Suppose that an SRS of size n is drawn from a population having unknown mean μ. To test the hypothesis H_0: $\mu = \mu_0$ based on an SRS of size n, compute the one-sample t statistic

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

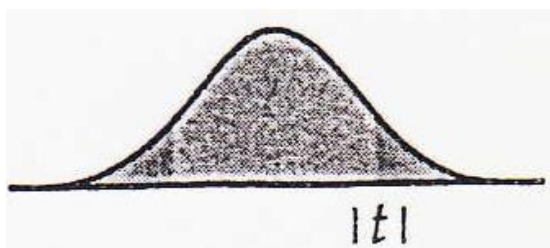
In terms of a random variable T having t_{n-1} distribution, the P-value for a test of H_0 against



H_a: $\mu > \mu_0$ is $P(T \geq t)$



H_a: $\mu < \mu_0$ is $P(T \leq t)$



H_a: $\mu \neq \mu_0$ is $2P(T \geq |t|)$

These P-values are exact if the population distribution is Normal and are approximately correct for large n in other cases.

Example: Suppose that, for the U.S. data in example before we want to test whether the U.S. average is different from the reported U.K. average.

$$H_0: \mu = 8.3$$
$$H_a: \mu \neq 8.3$$

$$n = 8, \qquad \bar{x} = 5, \qquad s = 3.63$$

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{5 - 8.3}{3.63/\sqrt{8}} = -2.57$$

$$\text{P-value} = 2P(T \geq 2.57), \quad T \sim t_7$$

$$2 \cdot 0.01 \leq \text{P-value} \leq 2 \cdot 0.02$$
$$0.02 \leq \text{P-value} \leq 0.04 < \alpha$$

p	0.02	0.01
t_7	2.517	2.998

Software P-value = 0.037

reject H_0 at $\alpha = 5\%$

What if we want to test whether the U.S. average is smaller than the UK average?

$$H_0: \mu = 8.3$$
$$H_a: \mu < 8.3$$

$$\text{P-value} = P(T \le -2.57)$$
$$= P(T \ge 2.57)$$

$$0.01 \le \text{P-value} \le 0.02$$

Software P-value = 0.0185

$\Rightarrow$ reject H_0 at $\alpha = 5\%$

At $\alpha = 0.05$ level, we conclude that the U.S. average is smaller than the U.K. average.

Matched Pairs *t* Procedures

In a matched pairs study, subjects are matched in pairs and the outcomes are compared within each pair.

Example: Many people believe that the moon influences the action of some individuals.

A study of dementia patients in nursing homes recorded various types of disruptive behaviors every day for 12 weeks.

Days were classified as moon days if they were in a three-day period centered at the day of the full moon.

For each patient the average number of disruptive behaviors was computed for moon days and for all other days.

The data for 15 subjects whose behaviors were classified as aggressive are presented in the table below.

Patient	Moon days	Other days	Difference
1	3.33	0.27	3.06
2	3.67	0.59	3.08
3	2.67	0.32	2.35
4	3.33	0.19	3.14
5	3.33	1.26	2.07
6	3.67	0.11	3.56
7	4.67	0.30	4.37
8	2.67	0.40	2.27
9	6.00	1.59	4.41
10	4.33	0.60	3.73
11	3.33	0.65	2.68
12	0.67	0.69	-0.02
13	1.33	1.26	0.07
14	0.33	0.23	0.10
15	2.00	0.38	1.62

For the differences:

$$n = 15, \qquad \bar{x} = 2.433, \qquad s = 1.46$$

$$n = 15, \qquad \bar{x} = 2.433, \qquad s = 1.46$$

We want to test

$$H_0 : \mu = 0$$
$$H_a : \mu \neq 0$$

$$t = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{2.433 - 0}{1.46/\sqrt{15}} = 6.45$$

p	0.001	0.0005
t_{14}	3.787	4.140

$$\text{P-value} = 2 \cdot P(t \geq 6.45) \leq 2 \cdot 0.0005$$
$$t \sim t_{14} \qquad = 0.001$$
$$< 0.01$$

$\Rightarrow$ reject H_0 at $\alpha = 0.01$

95% CI : $\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = 2.433 \pm 2.145 \cdot \frac{1.46}{\sqrt{15}}$

$= (1.62, \ 3.24) \not\ni 0$

The following are key points to remember concerning matched pairs:

- A matched pairs analysis is called for when subjects are matched in pairs or there are two measurements or observations on each individual and we want to examine the difference.

- For each pair or individual, use the difference between the two measurements as the data for your analysis.

- Use the one-sample confidence interval and significance-testing procedures that we learned earlier.

Robustness of the *t* procedure

The results of one-sample *t* procedures are exactly correct only when the population is Normal. In practice, the usefulness of the *t* procedures depends on how strongly they are affected by non-Normality.

Definition: A statistical inference procedure is called **robust** if the required probability calculations are insensitive to violations of the assumptions made.

Larger samples improve the accuracy of P-values and critical values from the t distributions when the population is not Normal. This is true for two reasons:

1. The sampling distribution of the sample mean $\bar{x}$ from a large sample is close to Normal (CLT). Normality of the individual observations is of little concern when the sample is large.

2. As the sample size n grows, the sample standard deviation s will be an accurate estimate of σ whether or not the population has a Normal distribution. This fact is closely related to the law of large numbers.

Here are practical guidelines for inference on a single mean:

- Sample size is less than 15: Use t procedures if data are close to Normal. If data are clearly non-Normal or if outliers are present, do not use t.

- Sample size at least 15: The t procedure can be used except in the presence of outliers or strong skewness.

- Large samples: The t procedures can be used even for clearly skewed distributions when the sample is large, roughly $n \geq 40$.