Central Limit Theorem, I.I.D variables

$X_n, n = 1, \ldots, N$ i.i.d, $E(X_n) = \mu, Var(X_n) = \sigma^2$.

Define the sum - $S_N = \sum_{n=1}^{N} X_n$.

Standardize the variables X_n and the variable S_N:

$$Y_n = \frac{X_n - \mu}{\sigma}, \quad Z_N = \frac{S_N - E(S_N)}{SD(S_N)} = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} Y_n.$$

Central Limit Theorem, I.I.D variables

Since $E(Y_n) = 0, Var(Y_n) = 1$:

$$(*) \quad m_{Y_n}(t) \sim m_{Y_n}(0) + tm'_{Y_n}(0) + t^2/2m''_{Y_n}(0) = 1 + \frac{1}{2}t^2,$$

Central Limit Theorem, I.I.D variables

Since $E(Y_n) = 0, Var(Y_n) = 1$:

$$(*) \quad m_{Y_n}(t) \sim m_{Y_n}(0) + tm'_{Y_n}(0) + t^2/2m''_{Y_n}(0) = 1 + \frac{1}{2}t^2,$$

$$\text{So} \qquad m_{Z_N}(t) = Ee^{\frac{t}{\sqrt{N}} \sum_{n=1}^{N} Y_n}$$

Central Limit Theorem, I.I.D variables

Since $E(Y_n) = 0, Var(Y_n) = 1$:

$$(*) \quad m_{Y_n}(t) \sim m_{Y_n}(0) + tm'_{Y_n}(0) + t^2/2m''_{Y_n}(0) = 1 + \frac{1}{2}t^2,$$

$$\begin{aligned} \text{So} \qquad m_{Z_N}(t) =& Ee^{\frac{t}{\sqrt{N}}\sum_{n=1}^{N} Y_n} \\ \text{Independence} \qquad =& \prod_{n=1}^{N} Ee^{\frac{t}{\sqrt{N}}Y_n} = \prod_{n=1}^{N} m_{Y_n}(t/\sqrt{N}) \end{aligned}$$

Central Limit Theorem, I.I.D variables

Since $E(Y_n) = 0, Var(Y_n) = 1$:

$$(*) \quad m_{Y_n}(t) \sim m_{Y_n}(0) + tm'_{Y_n}(0) + t^2/2m''_{Y_n}(0) = 1 + \frac{1}{2}t^2,$$

$$\begin{aligned}
\text{So} \qquad m_{Z_N}(t) &= Ee^{\frac{t}{\sqrt{N}}\sum_{n=1}^{N} Y_n} \\
\text{Independence} \qquad &= \prod_{n=1}^{N} Ee^{\frac{t}{\sqrt{N}}Y_n} = \prod_{n=1}^{N} m_{Y_n}(t/\sqrt{N}) \\
\text{By *} \qquad &\sim \left[1 + \frac{1}{2}\left(\frac{t}{\sqrt{N}}\right)^2\right]^N
\end{aligned}$$

Central Limit Theorem, I.I.D variables

Since $E(Y_n) = 0, Var(Y_n) = 1$:

$$(*) \quad m_{Y_n}(t) \sim m_{Y_n}(0) + tm'_{Y_n}(0) + t^2/2m''_{Y_n}(0) = 1 + \tfrac{1}{2}t^2,$$

$$\begin{aligned}
\text{So} \qquad m_{Z_N}(t) &= Ee^{\frac{t}{\sqrt{N}}\sum_{n=1}^{N} Y_n} \\
\text{Independence} \qquad &= \prod_{n=1}^{N} Ee^{\frac{t}{\sqrt{N}}Y_n} = \prod_{n=1}^{N} m_{Y_n}(t/\sqrt{N}) \\
\text{By *} \qquad &\sim \left[1 + \frac{1}{2}\left(\frac{t}{\sqrt{N}}\right)^2\right]^N \\
&= \left[1 + \frac{1}{2}\frac{t^2}{N}\right]^N \underset{N\to\infty}{\longrightarrow} e^{\frac{1}{2}t^2}.
\end{aligned}$$

Central Limit Theorem, I.I.D variables

Since $E(Y_n) = 0, Var(Y_n) = 1$:

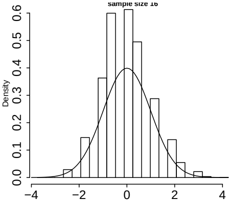
$$(*) \quad m_{Y_n}(t) \sim m_{Y_n}(0) + tm'_{Y_n}(0) + t^2/2m''_{Y_n}(0) = 1 + \tfrac{1}{2}t^2,$$

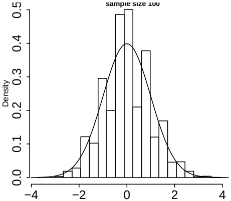
$$\begin{aligned}
\text{So} \qquad m_{Z_N}(t) =& Ee^{\frac{t}{\sqrt{N}} \sum_{n=1}^{N} Y_n} \\
\text{Independence} \qquad =& \prod_{n=1}^{N} Ee^{\frac{t}{\sqrt{N}} Y_n} = \prod_{n=1}^{N} m_{Y_n}(t/\sqrt{N}) \\
\text{By } * \qquad \sim& \left[1 + \frac{1}{2}\left(\frac{t}{\sqrt{N}}\right)^2\right]^N \\
=& \left[1 + \frac{1}{2}\frac{t^2}{N}\right]^N \xrightarrow[N\to\infty]{} e^{\frac{1}{2}t^2}.
\end{aligned}$$

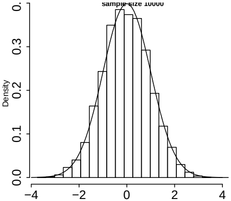
Which is the moment generating function of the standard normal distribution, i.e. with mean 0 and variance 1.

Central Limit Theorem, I.I.D variables, Version 1

Conclusion, version 1: As N gets larger the distribution of Z_N - the standardized sum S_N of N i.i.d random variables with mean μ and variance σ^2 gets closer and closer to $N(0,1)$.



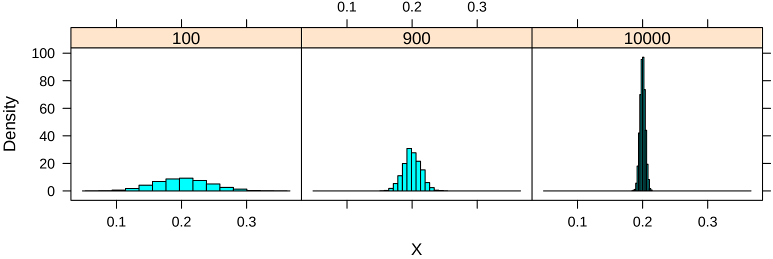




Central Limit Theorem, I.I.D variables, Version 2

Version 2: The distribution of the sample average $\overline{X_N}$ is approximately $N(\mu, \frac{\sigma}{\sqrt{N}})$.

$$\frac{S_N - N\mu}{\sqrt{N}\sigma} = Z_N \sim N(0,1) \rightarrow \frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}} \sim N(0,1) \rightarrow \bar{X}_N \sim N(\mu, \sigma/\sqrt{N}).$$



Central Limit Theorem, I.I.D Normal variables

Special case: X_n are draws from a population that has a normal distribution (in the first place).

If $f_{X_n} = N(\mu, \sigma)$, then $f_{\overline{X}} = N(\mu, \frac{\sigma}{\sqrt{N}})$.

This is exact, it is not an approximation.
Why? Use moment generating functions:
The MGF of X_n is $m_{X_n}(t) = e^{t\mu + t^2\sigma^2/2}$

Inference for the mean I

A Typical Inference Problem

Suppose we want to find out about the mean lifetime μ of a certain brand of light bulbs.

Suppose that the true mean μ is unknown, but we know (from previous studies) that the SD σ of light bulb lifetime is 30 hours.

In order to estimate the population mean μ we:

- Take a SRS of 100 light bulbs.
- Calculate the mean lifetime in the sample to be 1100 hours.

What can we say about the population mean?

- $E(\bar{X}) = \mu, \qquad SD(\bar{X}) = 30/\sqrt{100} = 3$
- $\bar{X} \to \mu$ (Law of Large Numbers)
- $\bar{X} \dot{\sim} N(\mu, 3)$ (CLT)

Inference for the mean II

Recall from the previous lecture that

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

The distribution is exact if the population distribution is normal, and approximately correct for large n in other cases, by the CLT. Thus,

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Rearranging terms, we have

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Inference for the mean III

In other words, there is 95% probability that the *random interval*

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

will cover μ.

In our example, $\bar{x} = 1100$, $\sigma = 30$, and $n = 100$. Therefore, the **95% confidence interval** for μ is

$$(1100 - 1.96 \times 3, 1100 + 1.96 \times 3) = (1094.12, 1105.88)$$

Calculating a confidence interval I

For the time being, we'll continue to assume that σ is known. To calculate a 95% **confidence interval** for the population mean μ

1. Take a random sample of size n and calculate the sample mean $\bar{x}$.
2. If n is large enough, $\bar{x} \dot{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ (by the CLT).
3. The confidence interval is given by

$$\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \quad \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Meaning of a confidence interval I

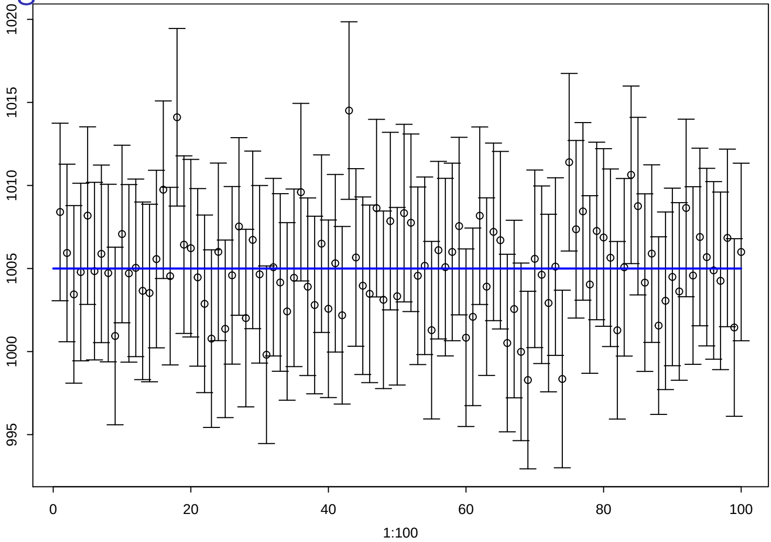
Suppose we repeat the following procedure multiple times:

1. Draw a random sample of size n
2. Calculate a 95% confidence interval for the sample

95% of the intervals thus constructed will cover the true (unknown) population mean.

```
sigma=30;mu=1005;cf=.95;n=121
lb=replicate(100,rnorm(n,mu,sigma))
aves=colSums(lb)/n
err95=sigma*qnorm(1-(1-cf)/2)/sqrt(n)
plotCI(1:100,y=aves,err95,ylab="")
lines(c(0,100),c(mu,mu),lwd=2,col="blue")
```

Meaning of a confidence interval II



Meaning of a confidence interval III

Consider estimating the speed of light using 64 measurements with sample mean $\bar{x} = 298,054\,km/s$.

Assume we know (from previous experience) that the SD of measurements made using the same procedure is $60\,km/s$.

What is a 95% CI for the true speed of light?

Incorrect:

- There is a 95% probability that the true speed of light lies in the interval (298,039.3, 298,068.7).
- In 95% of all possible samples, the true speed of light lies in the interval (298,039.3, 298,068.7).

Meaning of a confidence interval IV

Correct:

- There is 95% probability that the true speed of light lies in the random interval $(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$.
- If we repeatedly draw samples and calculate confidence intervals using this procedure, 95% of these intervals will cover the true speed of light.
- We are 95% confidents that the true speed of light lies in the interval (298,039.3, 298,068.7).

General form of a confidence interval I

In general, a CI for a parameter has the form

$$\text{estimate} \pm \text{margin of error}$$

where the margin of error is determined by the confidence level $(1-\alpha)$, the population SD σ, and the sample size n.

A $(1-\alpha)$ confidence interval for a parameter θ is an interval computed from a SRS by a method with probability $(1-\alpha)$ of containing the true θ.

For a random sample of size n drawn from a population of unknown mean μ and known SD σ, a $(1-\alpha)$ CI for μ is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

General form of a confidence interval II

Here z^* is the **critical value**, selected so that a standard Normal density has area $(1-\alpha)$ between $-z^*$ and z^*.

The quantity $z^*\sigma/\sqrt{n}$, then, is the **margin error**.

If the population distribution is normal, the interval is *exact*. Otherwise, it is *approximately correct for large n*.

General form of a confidence interval III

Some cautions on using the formula

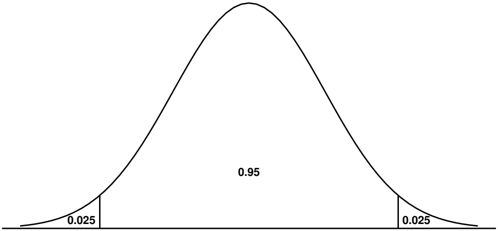
- Any formula for inference is correct only in specific circumstances.
- The data must be a SRS from the population.
- Because $\bar{x}$ is not resistant, outliers can have a large effect on the confidence interval.
- If the sample size is small and the population is not Normal, the true confidence level will be different.
- You need to know the standard deviation σ of the population.

General form of a confidence interval IV

Finding z^*

For a given confidence level $(1-\alpha)$, how do we find z^*?

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



$$P(-z^* \le Z \le z^*) = (1-\alpha) \Leftrightarrow P(Z < -z^*) = \frac{\alpha}{2}$$

General form of a confidence interval V

Thus, for a given confidence level $(1 - \alpha)$, we can look up the corresponding z^* value on the Normal table.

Common z^* values:

Confidence Level	90%	95%	99%
z^*	1.645	1.96	2.576

Hypothesis testing I

A **hypothesis test** is an assessment of the evidence provided by the data in favor of (or against) some claim about the population.

For example, suppose we perform a randomized experiment or take a random sample and calculate some sample statistic, say the sample mean.

We want to decide if the *observed* value of the *sample* statistic is consistent with some *hypothesized* value of the corresponding *population* parameter.
If the observed and hypothesized value differ (as they almost certainly will), is the difference due to an incorrect hypothesis or merely due to chance variation?

Hypothesis testing II

Example: Filling Coke Bottles

A machine at a Coke production plant is designed to fill bottles with 16oz of Coke. The actual amount varies slightly from bottle to bottle. From past experience, it is known that the SD 0.2oz. A SRS of 100 bottles filled by the machine has a mean 15.94oz per bottle. Is this evidence that the machine needs to be recalibrated, or could this difference be a result of random variation?

Example: GRE Scores

The mean score of all examinees on the Verbal and Quantitative sections of the GRE is about 1040. Suppose 14 randomly sampled U of C graduate students had a mean GRE V+Q score of 1310. Does this indicate that, as a whole, U of C graduate students have a higher mean GRE score than the national average?

Hypothesis testing III

Step 1. Formulate the null hypothesis and the alternative hypothesis

- The **null hypothesis** H_0 is the statement being tested. Usually it states that the difference between the observed value and the hypothesized value is only due to chance variation. For example, $\mu = 16$ oz.
- The **alternative hypothesis** H_a is the statement we will favor if we find evidence that the null hypothesis is false. It usually states that there is a real difference between the observed and hypothesized values.

  Usually the alternative is: $\mu \neq 16$ (Rarely: $\mu > 16$, or $\mu < 16$.)

A test is called

- **two-sided** if H_a is of the form $\mu \neq 16$.
- **one-sided** if H_a is of the form $\mu > 16$, or $\mu < 16$.

Hypothesis testing IV

Step 2. Calculate the **test statistic** on which the test will be based.

The test statistic measures the difference between the observed data and what would be expected *if* the null hypothesis were true.

Our goal is to answer the question, "How many standard errors is the observed sample value from the hypothesized value (under the null hypothesis)?"

For the Coke example, the test statistic is

$$z = \frac{15.94 - 16}{0.2/\sqrt{100}} = -3$$

Hypothesis testing V

Step 3. Find the **p-value** of the observed result

- The p-value is the probability of observing a test statistic *as extreme or more extreme than actually observed*, assuming the null hypothesis H_0 is true.
- The smaller the p-value, the stronger the evidence *against* the null hypothesis.
- if the p-value is as small or smaller than some number α (e.g. 0.01, 0.05), we say that the result is **statistically significant** at level α.
- α is called the **significance level** of the test.

In the case of the Coke example, $p = 0.0013$ for a one-sided test or $p = 0.0026$ for a two-sided test.

Hypothesis testing VI

Suppose H_0 is actually true. If we draw many samples, and perform a test for each one, α of these tests will (incorrectly) reject H_0. In other words, α is the probability that we will make a **Type I error**. Type II error occurs when we do not reject H_0 despite the fact that the alternative is true.

Hypothesis testing VII

Example: Filling Coke Bottles (cont.)
We are interested in assessing whether or not the machine needs to be recalibrated, which will be the case if it is systematically over- or under-filling bottles. Thus, we will use the hypotheses

$$H_0 : \mu = 16$$
$$H_a : \mu \neq 16$$

Recall that $\bar{x} = 15.94$, $\sigma = 0.2$, and $n = 100$. Thus,

$z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}} = -3$

The p-value for a two-sided test is $p = 2P(Z \geq 3) = 0.0026$.
If $\alpha = 0.01$, we reject H_0.
If $\alpha = 0.05$, we also reject H_0.

Hypothesis testing VIII

Example: TV Tubes

TV tubes are taken at random and the lifetime measured. $n = 100, \sigma = 300$ and $\bar{x} = 1265$ (days). Test whether the population mean is 1200, or greater than 1200.

$$H_0 : \mu = 1200$$
$$H_a : \mu > 1200$$

Under $H_0, \bar{x} \sim N(1200, 30)$.
$\therefore z = \frac{\bar{x}-1200}{30} \sim N(0, 1)$ under H_0
The test statistic is $z = \frac{1265-1200}{30} = 2.17$, and the p-value is $P(Z \geq 2.17 | H_0) = 0.015$
This is evidence against H_0 at significance level 0.05, so we reject H_0. That is, we conclude that the average lifetime of TV tubes is greater than 1200 days.

Hypothesis testing IX

Critical Value z_α

If the P-value is less than α we reject H_0.
For a two sided test This requires computing
$P(|Z| \geq z) = 2P(Z \geq z)$, for the observed test statistic z, and comparing it to α.
Alternatively we can find the critical value z_α such that
$P(|Z| \geq z_\alpha) = \alpha$ and check if $|z| > z_\alpha$.
For a one-sided test we find z_α such that $P(Z > z_\alpha) = \alpha$ and check if $z > z_\alpha$.
In the previous example $z_{.05} = 1.64$. Since $2.17 > 1.64$ the null hypothesis is rejected.

Tests for population mean I

Suppose we want to test the hypothesis that μ has a specific value:

$$H_0 : \mu = \mu_0$$

Since $\bar{x}$ estimates μ, the test is based on $\bar{x}$, which has a (perhaps approximately) Normal distribution. Thus,

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

is a standard normal random variable, *under the null hypothesis*.
p-values for different alternative hypotheses:

- $H_a : \mu > \mu_0$ – p-value is $P(Z \geq z)$ (area of right-hand tail)
- $H_a : \mu < \mu_0$ – p-value is $P(Z \leq z)$ (area of left-hand tail)
- $H_a : \mu \neq \mu_0$ – p-value is $2P(Z \geq |z|)$ (area of both tails)

Test for population proportion I

When population is just 0/1's population mean is simply proportion of 1's. One option is to use the same kind of test as above. (We will get back to that)

However in this case, once you make a Null hypothesis on the population proportion $H_0 : p = p_0$, you have fully determined the distribution of the number of 1's in the sample: $B(N, p_0)$.

So with a one sided alternative, say $H_a : p > p_0$, you can compute $s = \sum_{n=1}^{N} x_n$ and get the p-value: $P(S > s)$ for $S \dot{\sim} B(N, p_0)$.

Test for population proportion II

Example: You want to test Trump's hypothesis that there were many illegitimate voters who voted against him. I.e. he should have larger percent of the votes.
Your Null hypthesis is that that the proportion of Trump supporters was indeed $p = p_0 = .461$, against the alternative that $p > p_0$
You take a random sample of 1000 people, you keep only those who voted and were legitimate registered voters, say 534 are left. You find that 225 voted for Trump.

```
Pv=1-pbinom(225,534,.461)
Pv
```

```
[1] 0.9639
```

Test for population proportion III

So the P-value is .964! If the Null is true there is 96.4% chance of getting 225 or more.
You therefore reject Trumps Hypothesis...

But then you notice that there is a 3.6% chance of getting 225 or less. So if we had made the other one-sided alternative $H_a : p < p_0$ - Trump actually got less than .461, (maybe his votes were fraudulent) we would reject the Null at the 5% confidence level.

But you can't do that! That's called data snooping. Formulating the hypothesis after you've seen the data.

Test Interpretations I

p-value	Interpretation
$p > 0.10$	no evidence against H_0
$0.05 < p \leq 0.10$	weak evidence against H_0
$0.01 < p \leq 0.05$	evidence against H_0
$p \leq 0.01$	strong evidence against H_0

Test Interpretations II

Saying that a result is *statistically significant* does not signify that it is large or necessarily important. That decision depends on the particulars of the problem. A statistically significant result only says that there is substantial evidence that H_0 is false.

Failure to reject H_0 does not imply that H_0 is correct. It only implies that *we have insufficient evidence to conclude that H_0 is incorrect*.

Hypothesis tests and CI's I

A level α two-sided test rejects a hypothesis $H_0 : \mu = \mu_0$ exactly when the value of μ_0 falls outside a $(1 - \alpha)$ confidence interval for μ.

For example, consider a two-sided test of the following hypotheses

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

at the significance level $\alpha = .05$.

Assume the test statistic is z and $2P(Z > |z|) = 2P(Z > z) = p < \alpha$. Let z_α be the critical value for level α. Assume the population SD is σ_0.

Hypothesis tests and CI's II

$$p < \alpha$$

$$\Updownarrow$$

$$z > z_\alpha \quad \text{or} \quad z < -z_\alpha$$

$$\Updownarrow$$

$$\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} > z_\alpha \quad \text{or} \quad \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} < -z_\alpha$$

$$\Updownarrow$$

$$\mu_0 < \bar{x} - z_\alpha \cdot \frac{\sigma_0}{\sqrt{n}} \quad \text{or} \quad \mu_0 > \bar{x} + z_\alpha \cdot \frac{\sigma_0}{\sqrt{n}}$$

$$\Updownarrow$$

$$\mu_0 \notin [\bar{x} - z_\alpha \cdot \frac{\sigma_0}{\sqrt{n}}, \bar{x} + z_\alpha \cdot \frac{\sigma_0}{\sqrt{n}}]$$

μ_0 is not in the α confidence interval if and only if the null hypothesis is rejected at the α level.

Hypothesis tests and CI's III

- If μ_0 is a value inside the 95% confidence interval for μ, then this test will have a p-value greater than .05, and therefore will not reject H_0.
- If μ_0 is a value outside the 95% confidence interval for μ, then this test will have a p-value smaller than .05, and therefore will reject H_0.