# Chapter 9: Hypothesis Tests

Yu Yang

School of Statistics
University of Minnesota

October 24, 2022

# Motivating Example

A diet pill company advertises that at least 75% of its customers lose 10 pounds or more within 2 weeks. You suspect the company of falsely advertising the benefits of taking the pills. Suppose you take a sample of 100 product users and find that only 5% have lost at least 10 pounds. Is this enough to prove your claim? What about if 72% has lost at least 10 pounds?

**Goal:** Use sample data to formally test hypotheses about unknown population parameters (such as $\mu$ or $p$).

# Elements of a Hypothesis Test

1. **Assumptions:** The reliability of any hypothesis test relies on a certain set of assumptions being satisfied.

2. **Hypotheses:** Each hypothesis test has two hypotheses about the **population:**
   (a) **Null Hypothesis ($H_0$):** A statement about the population we want to **disprove**. It often represents **no effect**.

   (b) **Alternative Hypothesis ($H_a$):** What we hope to find evidence for. It is an *alternative* to the null hypothesis and should be stated before looking at the data (to avoid bias)!

**Diet Pill Example:** Let $p$ = true proportion of diet pill customers that lose at least 10 points. State the null and alterntive hypotheses:
$H_0 : p = .75$
$H_a : p < .75$

# Elements of a Hypothesis Test

3. **Test Statistic:** A test statistic is a measure of how compatible the data are with the null hypothesis. The **larger** the test statistic, the **less** compatible the data are with the null hypothesis.
   **General form:**

$$T = \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} = \frac{\text{estimate - hypothesized value}}{\text{standard error of estimate when } H_0 \text{ is true}}$$

   If $|T|$ is large, then
   - the sample estimate is many standard errors away from the hypothesized value
   - the data don't agree with the hypothesized value

# Elements of a Hypothesis Test

**NOTE:**

- $T$ is a function of the sample (changes with every sample)
- $T$ is a statistic
- $T$ has a sampling distribution under $H_0$ (that is, assuming $H_0$ is true)

The sampling distribution of $T$ helps us measure how likely our sample data are when $H_0$ is true.

# Elements of a Hypothesis Test

4. ***p*-value:** The *p*-value helps us interpret the test statistic.

   **Definition:** Assume $H_0$ is true. Then the *p*-value is the probability that the test statistic $T$ takes a value (in support of $H_a$) as or more extreme than the one we observed.

**Diet Pill Example:** Suppose that in your sample of 100 customers, 65% had lost at least 10 pounds in 2 weeks. Recall our hypotheses:

$H_0 : p = .75$
$H_a : p < .75$

# Elements of a Hypothesis Test

(a) Assuming $H_0$ is true, what is the sampling distribution of $\hat{p}$?

# Elements of a Hypothesis Test

(a) Assuming $H_0$ is true, what is the sampling distribution of $\hat{p}$?
If $H_0$ is true, then $p = p_0 = .75$

$$\hat{p} \dot{\sim} N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right) = N\left(.75, \sqrt{\frac{.75(1-.75)}{100}}\right) = N(.75, .04)$$

(b) Use part (a) to determine the sampling distribution of the test statistic $T$.

$$T = \frac{\text{estimate - hypothesized value}}{\text{standard error of estimiate when } H_0 \text{ is true}}$$

# Elements of a Hypothesis Test

(a) Assuming $H_0$ is true, what is the sampling distribution of $\hat{p}$?
If $H_0$ is true, then $p = p_0 = .75$

$$\hat{p} \dot{\sim} N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right) = N\left(.75, \sqrt{\frac{.75(1-.75)}{100}}\right) = N(.75, .04)$$

(b) Use part (a) to determine the sampling distribution of the test statistic $T$.

$$T = \frac{\text{estimate - hypothesized value}}{\text{standard error of estimiate when } H_0 \text{ is true}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \dot{\sim} N(0,1)$$

(c) Calculate the test statistic for this hypothesis test.

# Elements of a Hypothesis Test

(a) Assuming $H_0$ is true, what is the sampling distribution of $\hat{p}$?
If $H_0$ is true, then $p = p_0 = .75$

$$\hat{p} \dot{\sim} N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right) = N\left(.75, \sqrt{\frac{.75(1-.75)}{100}}\right) = N(.75, .04)$$

(b) Use part (a) to determine the sampling distribution of the test statistic $T$.

$$T = \frac{\text{estimate - hypothesized value}}{\text{standard error of estimiate when } H_0 \text{ is true}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \dot{\sim} N(0,1)$$

(c) Calculate the test statistic for this hypothesis test.

$$T = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{.65 - .75}{.04} = -2.5$$

# Elements of a Hypothesis Test

(d) Calculate and interpret the $p$-value for this test statistic.

# Elements of a Hypothesis Test

(d) Calculate and interpret the $p$-value for this test statistic.

$$\begin{aligned} p\text{-value} &= P(T < -2.5) \\ &= P(Z < -2.5) \\ &= .0062 \end{aligned}$$

Interpretation: If the proportion of customers that lose at least 10 pounds is truly .75, it is very unlikely that we would choose a sample with so few people achieving this weight loss (prob = .0062). Thus our data provides evidence that $H_0$ is false.

In general, **the smaller the $p$-value, the more evidence we have against $H_0$.**

# Elements of a Hypothesis Test

5. **Conclusion and Interpretation:** We use the $p$-value to decide whether or not the data provides sufficient evidence to reject $H_0$ in favor of $H_a$.

   **The Idea:** Recall that a small p-value indicates that our observation based on the sample is unlikely to occur under $H_0$. How small is small enough to reject $H_0$?

# Statistical Significance

**Rejection Rule:** We determine "statistical significance" by comparing a *p*-value to a *significance level* $\alpha$.

- *p*-value $< \alpha \Rightarrow$ We have strong evidence against $H_0$. We conclude that the results are **statistically significant** at level $\alpha$ and reject $H_0$.
- *p*-value $\geq \alpha \Rightarrow$ We fail to reject $H_0$ in favor of $H_a$. We do **not** have enough evidence to conclude $H_0$ is false.

Oct 24 Lecture Stopped Here

# Interpreting the Significance Level

If, for example, we choose $\alpha = .05$, we require strong enough evidence **against** $H_0$ that when $H_0$ is true, there is only a 5% chance that we mistakenly reject it.

Common choices of $\alpha$:

- $\alpha = .01 \Rightarrow$ need to be 99% confident that $H_0$ is false in order to reject
- $\alpha = .05 \Rightarrow$ need to be 95% confident that $H_0$ is false in order to reject
- $\alpha = .10 \Rightarrow$ need to be 90% confident that $H_0$ is fase in order to reject.

$\alpha$ is chosen by the researcher *prior to* collecting data.m

# Steps for Hypothesis Testing

**Recap:** Steps for hypothesis testing

1. Check assumptions
2. Formulate hypotheses (null and alternative)
3. Determine test statistic (and its distribution)
4. Compute $p$-value (based on the distribution of the test statistic)
5. Draw conclusions

# Normal Hypothesis Test for Population Proportion $p$

**Assumptions:**

1. Random sample

2. $n$ is large enough that $\hat{p}$ is approximately normal if $H_0$ is true.

   Recall our rule of thumb for applying the CLT: at least 15 expected successes ($np_0 \geq 15$) and 15 expected failures ($n(1 - p_0) \geq 15$.

# Hypotheses

1. $H_0 : p = p_0$
   $H_a : p < p_0$
2. $H_0 : p = p_0$
   $H_a : p > p_0$
3. $H_0 : p = p_0$
   $H_a : p \neq p_0$

- Note: the value of $p_0$ is the same in both $H_0$ and $H_a$.
- Hypotheses 1 and 2 are called "**one-sided tests**" and hypothesis 3 is called a "**two-sided test**".

# Test Statistic

The test statistic is one number that summarizes the data (assuming $H_0$ is true). In this case, our test statistic is

$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \dot{\sim} N(0,1) \text{ under } H_0(p = p_0)$$

Why does $z^*$ follow this distribution?

# Test Statistic

The test statistic is one number that summarizes the data (assuming $H_0$ is true). In this case, our test statistic is

$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \dot{\sim} N(0,1) \text{ under } H_0(p = p_0)$$

Why does $z^*$ follow this distribution?

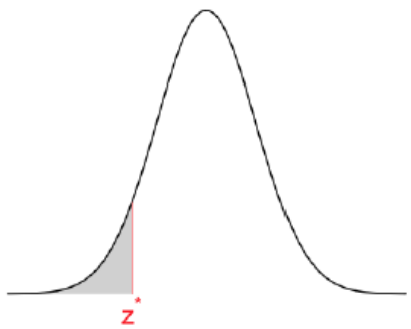
$\hat{p} \dot{\sim} N\left(p_0, \sqrt{\frac{p_0(1-p_0)}{n}}\right)$ under the null hypothesis.

So, when we standardize, we get $N(0,1)$!

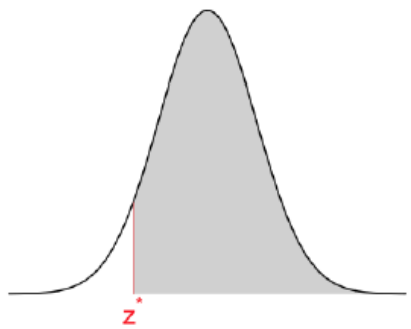
# *p*-value

*p*-value = P(observe a value as or more extreme than $z^*$ in favor of $H_a|H_0$ is true)

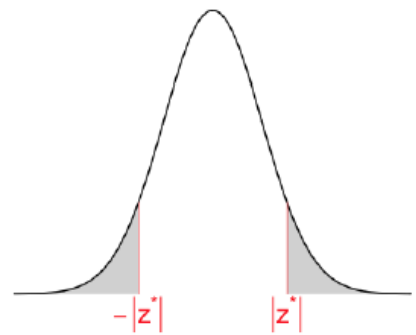
Hypothesis 1:
**p-val** $= P(Z < z^*)$




Hypothesis 2:
**p-val** $= P(Z > z^*)$




Hypothesis 3:
**p-val** $= 2P(Z > |z^*|)$




**Conclusion:**

*p*-value $< \alpha \Rightarrow$ reject $H_0$

*p*-value $\geq \alpha \Rightarrow$ fail to reject $H_0$

# Example 9.1

The manufacturer of a new breast cancer screening method claims that it detects cancer in more than 83% women who have it. To investigate, they apply their screening method to a sample of 203 randomly selected women known to have breast cancer. The test detects cancer in 184 of these women. Use this information to test their claim at the .01 level.

# Example 9.1: Assumptions

Our testing assumptions are satisfied:

- This is a random sample
- There are at least 15 expected successes and failures (under $H_0$).

  Expected successes $= 203\ (.83) = 168 \geq 15$

  Expected failures $= 203(1 - .83) = 35 \geq 15$

# Example 9.1: Hypotheses

Read the question carefully. They want to test their claim that the machine detects cancer correctly more than 83% of the time.

$H_0$ :

# Example 9.1: Hypotheses

Read the question carefully. They want to test their claim that the machine detects cancer correctly more than 83% of the time.

$H_0$ : p = .83
$H_a$ :

# Example 9.1: Hypotheses

Read the question carefully. They want to test their claim that the machine detects cancer correctly more than 83% of the time.

$H_0 : p = .83$
$H_a : p > .83$

# Test Statistic

Remember the test statistic and its distribution

$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1) \text{ under } H_0$$

# Test Statistic

Remember the test statistic and its distribution

$$z^* = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1) \text{ under } H_0$$

In this case

$$z^* = \frac{184/203 - .83}{\sqrt{\frac{.83(1-.83)}{203}}} = 2.88$$

# $p$-value

$p$-value $= P(Z > 2.88) =$ `pnorm(2.88, lower.tail = FALSE)` $= .002$.

There is only a .2% chance that our test performs this well in a sample of size 203 under $H_0$.

# Conclusion:

Significance level is .01 (given in the problem), so we compare the $p$-value to that number!

$p$-value $= .002 < .01 \Rightarrow$ reject $H_0$ at the .01 level.

**Interpretation:** We **have enough evidence** to conclude that the true proportion of breast cancers detected by the new screening method is significantly greater than 83%.

# Example 9.2

Among the employees eligible for management training at a large supermarket chain in Florida, 40% are women. However, since management training began, only 12 of the 40 employees (30%) chosen for the training were women. At the .05 level, test the claim of a women's group that women are being passed over for management training in favor of their male colleagues. That is, test the claim that a disproportionate number of people selected for the training are men.

Let $p$ = proportion of employees selected for training that are women.

Hypotheses:
$H_0$:

# Example 9.2

Among the employees eligible for management training at a large supermarket chain in Florida, 40% are women. However, since management training began, only 12 of the 40 employees (30%) chosen for the training were women. At the .05 level, test the claim of a women's group that women are being passed over for management training in favor of their male colleagues. That is, test the claim that a disproportionate number of people selected for the training are men.

Let $p$ = proportion of employees selected for training that are women.

Hypotheses:
$H_0$: $p = .40$
$H_a$:

# Example 9.2

Among the employees eligible for management training at a large supermarket chain in Florida, 40% are women. However, since management training began, only 12 of the 40 employees (30%) chosen for the training were women. At the .05 level, test the claim of a women's group that women are being passed over for management training in favor of their male colleagues. That is, test the claim that a disproportionate number of people selected for the training are men.

Let $p$ = proportion of employees selected for training that are women.

Hypotheses:
$H_0$: $p = .40$
$H_a$: $p < .40$

Do the assumptions hold?

# Example 9.2

Among the employees eligible for management training at a large supermarket chain in Florida, 40% are women. However, since management training began, only 12 of the 40 employees (30%) chosen for the training were women. At the .05 level, test the claim of a women's group that women are being passed over for management training in favor of their male colleagues. That is, test the claim that a disproportionate number of people selected for the training are men.

Let $p$ = proportion of employees selected for training that are women.

Hypotheses:
$H_0$: $p = .40$
$H_a$: $p < .40$

Do the assumptions hold? Yes! Expected successes $= 40(.40) = 16 \geq 15$ and expected failures $= 40(.60) = 24 \geq 15$

# Example 9.2 in R

```
> prop.test(x = 12, n = 40, conf.level = 0.95, p = 0.40,
+           alternative = "less", correct = FALSE)

1-sample proportions test without continuity correction

data:  12 out of 40, null probability 0.4
X-squared = 1.6667, df = 1, p-value = 0.09835
alternative hypothesis: true p is less than 0.4
95 percent confidence interval:
 0.0000000 0.4287085
sample estimates:
  p
0.3
```

Notes:

1. If $H_a : p > .40$, then `alternative = "greater"`
2. If $H_a : p \neq .40$, then `alternative = "two.sided"`

# Example 9.2 - Interpretation

The $p$-value is 0.09835.
This means that under $H_0$, there is a nearly 9.8% probability that we observe a sample proportion of .3 or lower. This is not very unlikely, so it is reasonable that this sample is drawn from a population with proportion .4

Conclusion:

# Example 9.2 - Interpretation

The $p$-value is 0.09835.
This means that under $H_0$, there is a nearly 9.8% probability that we observe a sample proportion of .3 or lower. This is not very unlikely, so it is reasonable that this sample is drawn from a population with proportion .4

Conclusion:

We failed to reject $H_0$ at the significance level .05.
We do not have enough evidence to conclude that the true proportion of employees selected for training that are women is significantly less than 40%.

# The $t$-test: Hypothesis Testing for Population Mean $\mu$

The basic structure of hypothesis tests regarding population means is the same, only the details change.

**Assumptions**:

1. Random sample
2. Normality and/or large $n$: Population distribution is approximately normal and/or sample size is large.

**Hypotheses**:

1. $H_0 : \mu = \mu_0$
   $H_a : \mu < \mu_0$
2. $H_0 : \mu = \mu_0$
   $H_a : \mu > \mu_0$
3. $H_0 : \mu = \mu_0$
   $H_a : \mu \neq \mu_0$

# The $t$-Test: Hypothesis Testing for Population Mean $\mu$

**Test Statistic**:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \text{ (from Chapter 8)}$$

***p*-value**:
*p*-value = P(observe a value as or more extreme than $t|H_0$ is true)

| Hypothesis 1 | Hypothesis 2 | Hypothesis 3 |
|---|---|---|
| p-val = $P(T_{n-1} < t)$ | p-val = $P(T_{n-1} > t)$ | p-val = $2P(T_{n-1} > \|t\|)$ |

**Conclusion**:
*p*-value $< \alpha \Rightarrow$ reject $H_0$
*p*-value $\geq \alpha \Rightarrow$ fail to reject $H_0$

# Example 9.3

In the 2004 General Social Survey, men were asked how many hours they worked in the previous week. For the random sample of 895 male workers, the mean was 45.3 hours with a standard deviation of 14.8 hours. Does this data support the claim that the true average number of hours that men work each week ($\mu$) exceeds the standard 40 hour work week? (Test this claim at the 0.05 level).

**Assumptions:**

# Example 9.3

In the 2004 General Social Survey, men were asked how many hours they worked in the previous week. For the random sample of 895 male workers, the mean was 45.3 hours with a standard deviation of 14.8 hours. Does this data support the claim that the true average number of hours that men work each week ($\mu$) exceeds the standard 40 hour work week? (Test this claim at the 0.05 level).

**Assumptions:**
Random sample and $n$ is large enough so that the $t$ procedure will be robust to any skewness in the distribution of incomes.

**Hypothesis:**
$H_0$ :

# Example 9.3

In the 2004 General Social Survey, men were asked how many hours they worked in the previous week. For the random sample of 895 male workers, the mean was 45.3 hours with a standard deviation of 14.8 hours. Does this data support the claim that the true average number of hours that men work each week ($\mu$) exceeds the standard 40 hour work week? (Test this claim at the 0.05 level).

**Assumptions:**
Random sample and $n$ is large enough so that the $t$ procedure will be robust to any skewness in the distribution of incomes.

**Hypothesis:**
$H_0 : \mu = 40$
$H_a :$

# Example 9.3

In the 2004 General Social Survey, men were asked how many hours they worked in the previous week. For the random sample of 895 male workers, the mean was 45.3 hours with a standard deviation of 14.8 hours. Does this data support the claim that the true average number of hours that men work each week ($\mu$) exceeds the standard 40 hour work week? (Test this claim at the 0.05 level).

**Assumptions:**
Random sample and $n$ is large enough so that the $t$ procedure will be robust to any skewness in the distribution of incomes.

**Hypothesis:**
$H_0 : \mu = 40$
$H_a : \mu > 40$

# Example 9.3

**Test Statistic:**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{45.3 - 40}{14.8/\sqrt{895}} = 10.7$$

# Example 9.3

**Test Statistic:**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{45.3 - 40}{14.8/\sqrt{895}} = 10.7$$

***p*-value:**

$p$-val $= P(t_{894} > 10.7) < .001$

```
> pt(10.7, df = 894, lower.tail = FALSE)
[1] 1.586841e-25
```

# Example 9.3

**Interpretation:**
Under $H_0$, there is a less than .001 probability to observe a $t$-statistic of 10.7 or greater. This is very unlikely, so it is highly unlikely that this sample is drawn from a population with mean 40.

# Example 9.3

**Interpretation:**
Under $H_0$, there is a less than .001 probability to observe a $t$-statistic of 10.7 or greater. This is very unlikely, so it is highly unlikely that this sample is drawn from a population with mean 40.

**Conclusion:**
$p$-val $< .001 < .05 \Rightarrow$ Reject $H_0$ at the .05 level.

We have enough evidence to conclude that the average work week for men is significantly larger than 40 hours.

# Hypothesis Test and Confidence Intervals: Example 9.4

It is important for nutritional information on food packaging to be accurate. A random sample of $n = 10$ frozen dinners of a certain brand was selected from a production line. The mean calorie content for these dinners was 246.6 with a standard deviation of 10.803 calories. The data is approximately normal with no standard outliers.

(a) The packaging for this dinner lists the calorie content as 240 calories. Let $\mu$ = true mean calorie content of the frozen dinner and test whether or not the information on the package is accurate at the 0.05 level.

**Assumption:**

# Hypothesis Test and Confidence Intervals: Example 9.4

It is important for nutritional information on food packaging to be accurate. A random sample of $n = 10$ frozen dinners of a certain brand was selected from a production line. The mean calorie content for these dinners was 246.6 with a standard deviation of 10.803 calories. The data is approximately normal with no standard outliers.

(a) The packaging for this dinner lists the calorie content as 240 calories. Let $\mu$ = true mean calorie content of the frozen dinner and test whether or not the information on the package is accurate at the 0.05 level.

**Assumption:**
Random sample and though $n$ is small, the population from which the data are sample is approximately normal according to the problem description

# Example 9.4

**Hypotheses:**

$H_0$ :

# Example 9.4

**Hypotheses:**

$H_0 : \mu = 240$

$H_a :$

# Example 9.4

**Hypotheses:**

$H_0 : \mu = 240$

$H_a : \mu \neq 240$

**Test Statistic:**

# Example 9.4

**Hypotheses:**

$H_0 : \mu = 240$

$H_a : \mu \neq 240$

**Test Statistic:**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{246.6 - 240}{10.803/\sqrt{10}} = 1.93$$

***p*-value:**

# Example 9.4

**Hypotheses:**

$H_0 : \mu = 240$

$H_a : \mu \neq 240$

**Test Statistic:**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{246.6 - 240}{10.803/\sqrt{10}} = 1.93$$

***p*-value:**

$p\text{-val} = 2P(t_9 > 1.93) = 2(.0428) = .086$

```
> 2 * pt(1.93, df = 9, lower.tail = FALSE)
[1] 0.08567249
```

**Conclusion:**

# Example 9.4

**Hypotheses:**
$H_0 : \mu = 240$
$H_a : \mu \neq 240$

**Test Statistic:**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{246.6 - 240}{10.803/\sqrt{10}} = 1.93$$

***p*-value:**
$p\text{-val} = 2P(t_9 > 1.93) = 2(.0428) = .086$

```
> 2 * pt(1.93, df = 9, lower.tail = FALSE)
[1] 0.08567249
```

**Conclusion:** $p$-val $= .0856 > .05 \Rightarrow$ Fail to reject $H_0$ at the .05 level. There is not enough evidence to conclude that true mean content is significantly different from what the label says.

# Example 9.4

(b) Calculate and interpret a 95% confidence interval for $\mu$.

# Example 9.4

(b) Calculate and interpret a 95% confidence interval for $\mu$.

$$\begin{aligned}
\bar{x} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} &= 246.6 \pm t_{.025,9}\frac{10.803}{\sqrt{10}} \\
&= 246.6 \pm 2.262\frac{10.803}{\sqrt{10}} \\
&= 246.6 \pm 7.73 \\
&= (238.87, 254.33)
\end{aligned}$$

We are 95% confident that the true mean calorie content is between 238.87 and 254.33 calories.

# Example 9.4

## (c) Repeat these analyses in R.

```
> cal <- c(240, 253, 243, 267, 258, 239, 235, 252, 246, 233)
> t.test(x = cal, conf.level = .95,
alternative = "two.sided", mu = 240)

One Sample t-test
data:  cal
t = 1.9319, df = 9, p-value = 0.08541
alternative hypothesis: true mean is not equal to 240
95 percent confidence interval:
238.8718 254.3282
sample estimates:
mean of x
246.6
```

- `mu` is hypothesized value of $\mu$
- If $H_a : \mu < 240$, then we would do `alternative = "less"`
- If $H_a : \mu > 240$, then we would do `alternative = "greater"`
- Always need `alternative = "two.sided"` to get the correct CI.

# Example 9.4

(d) What conclusions can you make about the package's claim based on the hypothesis test and the confidence interval?

Notice that 240 calories is in the 95% CI for $\mu$. Therefore, we can use both the hypothesis text and CI to conclude that 240 calories is a plausible value of $\mu$.

# Equivalence Between CI and 2-sided Hypothesis Test for $\mu$

Inference about $\mu$ has an exact equivalence between the **two-sided** hypothesis test at level $\alpha$ and the confidence interval with confidence level $1 - \alpha$.

# Equivalence Between CI and 2-sided Hypothesis Test for $\mu$

Inference about $\mu$ has an exact equivalence between the **two-sided** hypothesis test at level $\alpha$ and the confidence interval with confidence level $1 - \alpha$.

Specifically, suppose we wish to test $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ at level $\alpha$. Then,

1. If $\mu_0$ is in the 1 - $\alpha$ CI for $\mu$, then
   $\mu_0$ is compatible with the data, do not reject $H_0$ at level $\alpha$

2. If $\mu_0$ is not in the $1 - \alpha$ confidence level CI for $\mu$, then
   $\mu_0$ is not compatible with the data, reject $H_0$ at level $\alpha$.

# Equivalence Between CI and 2-sided Hypothesis Test for $\mu$

Inference about $\mu$ has an exact equivalence between the **two-sided** hypothesis test at level $\alpha$ and the confidence interval with confidence level $1 - \alpha$.

Specifically, suppose we wish to test $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ at level $\alpha$. Then,

1. If $\mu_0$ is in the 1 - $\alpha$ CI for $\mu$, then
   $\mu_0$ is compatible with the data, do not reject $H_0$ at level $\alpha$

2. If $\mu_0$ is not in the $1 - \alpha$ confidence level CI for $\mu$, then
   $\mu_0$ is not compatible with the data, reject $H_0$ at level $\alpha$.

For example, we can use a 95% CI for $\mu$ to make a conclusion for the two-sided test at the .05 level and can use a 99% CI to make a conclusion for the two-sided test at the .01 level (and so on).

# Equivalence Between CI and 2-sided Hypothesis Test for $\mu$

1. Confidence intervals and one-sided tests for $\mu$ are not compatible!
   Two-sided tests and CIs consider alternative values in both tails of the sampling distribution whereas, one-sided tests only consider alternative values in one tail of the sampling distribution.

2. Inference about **proportions** does **not** have an exact equivalence between the confidence interval and 2-sided hypothesis test.
   CI uses $se = \sqrt{\hat{p}(1-\hat{p})/n}$
   tests use $se = \sqrt{p_0(1-p_0)/n}$

# Type I and Type II Errors

Inference based on a hypothesis test may not always reflect the "truth"!

| | $H_0$ true | $H_a$ true |
|---|---|---|
| Do not reject $H_0$ | **correct** | **Type II error** |
| Reject $H_0$ | **Type I error** | **correct** |

**Note:** Error does not mean we did anything wrong - it just means that the conclusion based on our data does not reflect the true state of nature.

**Type I:** Incorrectly reject $H_0$ (false positive).
**Type II:** Incorrectly fail to reject $H_0$ (false negative).

# Example 9.5

According to the Journal of Psychology and Aging, older workers have an average job satisfaction rating of 4.3 (on a scale of 0 to 5). We are interested in knowing if the average satisfaction rating is lower among young workers. That is, we want to test

$$H_0 : \mu = 4.3$$
$$H_a : \mu < 4.3$$

where $\mu =$ the mean job satisfaction rate for younger workers. What are the Type I and Type II errors in the context of this problem?

Type I:

# Example 9.5

According to the Journal of Psychology and Aging, older workers have an average job satisfaction rating of 4.3 (on a scale of 0 to 5). We are interested in knowing if the average satisfaction rating is lower among young workers. That is, we want to test

$$H_0 : \mu = 4.3$$
$$H_a : \mu < 4.3$$

where $\mu =$ the mean job satisfaction rate for younger workers. What are the Type I and Type II errors in the context of this problem?

Type I: We conclude $\mu < 4.3$ when the true satisfaction rating for youngsters is not significantly less than for older workers.

Type II:

# Example 9.5

According to the Journal of Psychology and Aging, older workers have an average job satisfaction rating of 4.3 (on a scale of 0 to 5). We are interested in knowing if the average satisfaction rating is lower among young workers. That is, we want to test

$$H_0 : \mu = 4.3$$
$$H_a : \mu < 4.3$$

where $\mu =$ the mean job satisfaction rate for younger workers. What are the Type I and Type II errors in the context of this problem?

Type I: We conclude $\mu < 4.3$ when the true satisfaction rating for youngsters is not significantly less than for older workers.

Type II: We conclude that the true satisfaction rating for youngsters is not significantly less than for older workers when it actually is.

# Probability of a Type I error

Recall the interpretation of the significance level, $\alpha$, of a hypothesis test.

> *If we set significance level* $= \alpha$*, we require strong enough evidence against* $H_0$ *that the probability of mistakenly rejecting* $H_0$ *when it is actually true is only* $\alpha$*.*

Therefore, by definition

$$\boxed{\text{P(Type I error)} = \alpha}$$

**Controlling the Probability of a Type I error:** Our choice of $\alpha$ controls the chances of making a Type I error. How should we choose $\alpha$?

# Probability of a Type I error

Recall the interpretation of the significance level, $\alpha$, of a hypothesis test.

> *If we set significance level* $= \alpha$*, we require strong enough evidence against* $H_0$ *that the probability of mistakenly rejecting* $H_0$ *when it is actually true is only* $\alpha$*.*

Therefore, by definition

$$\boxed{\text{P(Type I error)} = \alpha}$$

**Controlling the Probability of a Type I error:** Our choice of $\alpha$ controls the chances of making a Type I error. How should we choose $\alpha$? $\alpha$ is set to reflect how bad a Type I error would be. The worse the consequences of a Type I error, the smaller one should set $\alpha$.

# Probability of a Type II Error

Calculating the probability of a Type II Error can be complex. It is not as easy to control as the probability if a Type I error. In general

P(Type II error) decreases as P(Type I error) increases.

Why?
As P(Type I error) increases, we require less evidence for rejecting $H_0$.
This means that we are less likely to make a Type II error.

# Limitations and Common Misinterpretations

1. Statistical significance does not mean practical significance. Statistical significance relates to the existence of an effect (or difference), whereas practical significance relates to the size of a possible effect (or difference).

   **Example:** Let $\mu =$ true IQ of children in a certain region of the U.S.. Based on a sample of 5000 children with an average IQ of 100.8 and standard deviation of 16.21, the p-value for the test of $H_0 : \mu = 100$ versus $H_a : \mu > 100$ is approximately .0002. Therefore, though there is not much of a practical difference between the sample mean (100.8) and the hypothesized mean (100), this difference is highly statistically significant. *That is, there is statistical significance, but not practical significance.*

# Limitations and Common Misinterpretations

2. There is no sharp distinction between "significant" and "not significant"(i.e. $\alpha$ is not a magic number).

   **Example:** $p$-values of .047 and .052 have almost no practical difference, but they will result in different conclusions if $\alpha = .05$. Therefore, always report the exact $p$-value when it is close to $\alpha$.

3. "Failing to reject" $H_0$ is not equivalent to "accepting" $H_0$.

   When we 'fail to reject' $H_0$, we are saying we do not have enough evidence to conclude that $H_0$ is false. We are not saying $H_0$ is true.

# Limitations and Common Misinterpretations

4. Recall the definition of the $p$-value:
   $p$-value = P(test statistic is more extreme than the one we observed $|H_0$ is true).
   Note that $p$-value $\neq$ P($H_0$ is true $|$ observed test statistic)

5. It is misleading to report results only if they are statistically significant.

   Suppose we run 20 similar tests, of which only one is statistically significant. If we report the one significance test, it is quite likely that we are reporting a Type I error.

6. Don't always believe what you read! Since researchers do often only report results that are significant and beneficial to their cause, be sure to keep in mind that these results may actually be the result of a Type I error!