

# Inference III

Michael C. Hackett  
Assistant Professor, Computer Science

Community  
College  
*of* Philadelphia

# One-sample mean

- We've seen how to model a sample proportion,  $\hat{p}$ , using a normal distribution.
- We can also model the sample mean,  $\bar{x}$ , using a normal distribution when certain conditions are met.

# One-sample mean

- The sample mean tends to follow a normal distribution centered at the population mean,  $\mu$ , under certain conditions.
- The standard error for the sample mean can be computed using the population standard deviation,  $\sigma$ , and the sample size  $n$ .
- A sufficiently large sample of  $n$  independent observations from a population with mean  $\mu$  and standard deviation  $\sigma$ , the sampling distribution of  $\bar{x}$  will be nearly normal:

$$\text{Mean} = \mu$$

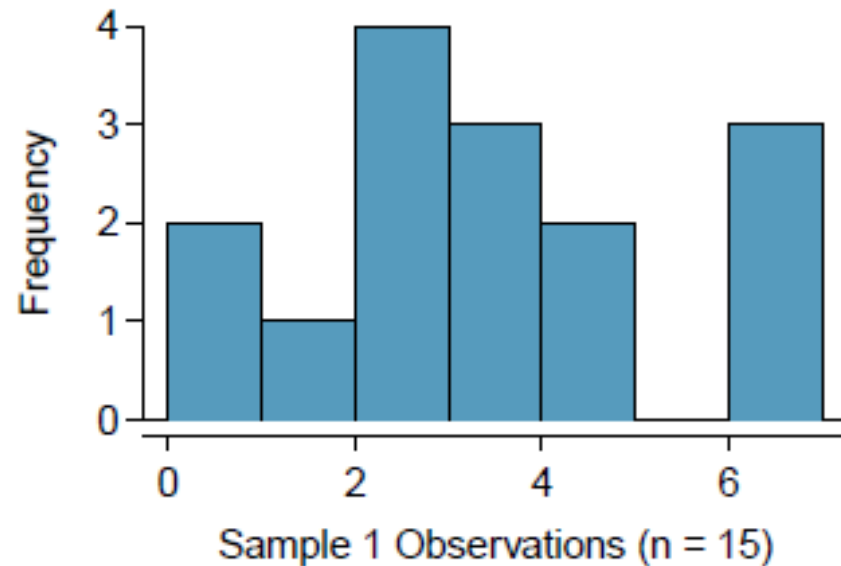
$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}}$$

# One-sample mean

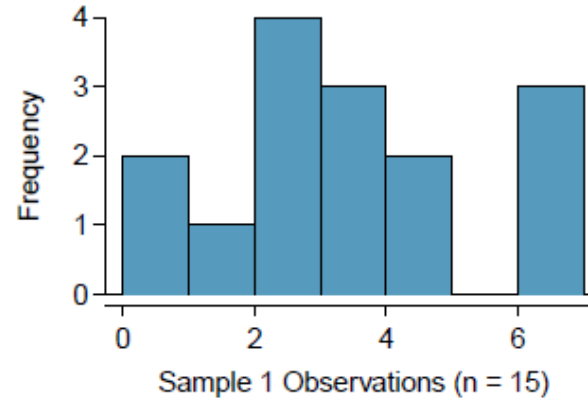
- Applying the Central Limit Theorem for a sample mean  $\bar{x}$  :
- The sample observations must be independent.
- When a sample is small, the sample observations must come from a normally distributed population.
- Normality checks:
  - $n < 30$ : With no clear outliers, the data can be assumed to be normal
  - $n \geq 30$ : With no particularly extreme outliers, the sampling distribution of  $\bar{x}$  can be assumed to be normal
- No perfect way to check for normality, but these are rules of thumb

# One-sample mean

- Is the independence and normality conditions met for the distribution below?
  - Assuming the data comes from simple random samples.



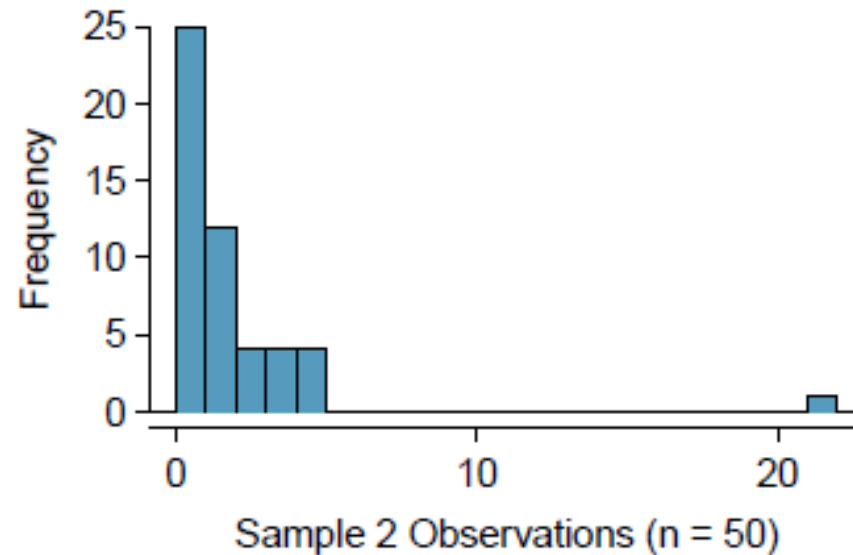
# One-sample mean



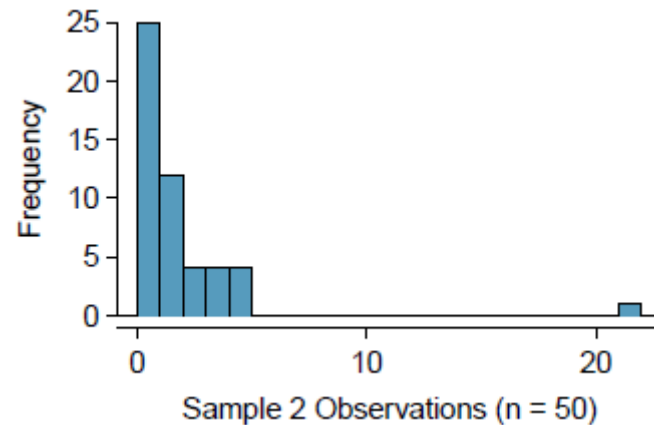
- Independence: Yes, since randomly sampled
- Normality: Fewer than 30 observations with no clear outliers
- Condition is reasonably met.

# One-sample mean

- Is the independence and normality conditions met for the distribution below?
  - Assuming the data comes from simple random samples.



# One-sample mean



- Independence: Yes, since randomly sampled
- Normality: More than 30 observations. One particularly extreme outlier.
- Condition is not reasonably met.



# One-sample mean

- When computing the standard error for a sample proportion, it relied on the population proportion,  $p$ , which is generally not known.
- Similarly, we cannot compute the standard error for  $\bar{x}$  since we do not know the population standard deviation,  $\sigma$ .
- The solution was to use the sample value in place of the population value when computing the standard error.

# One-sample mean

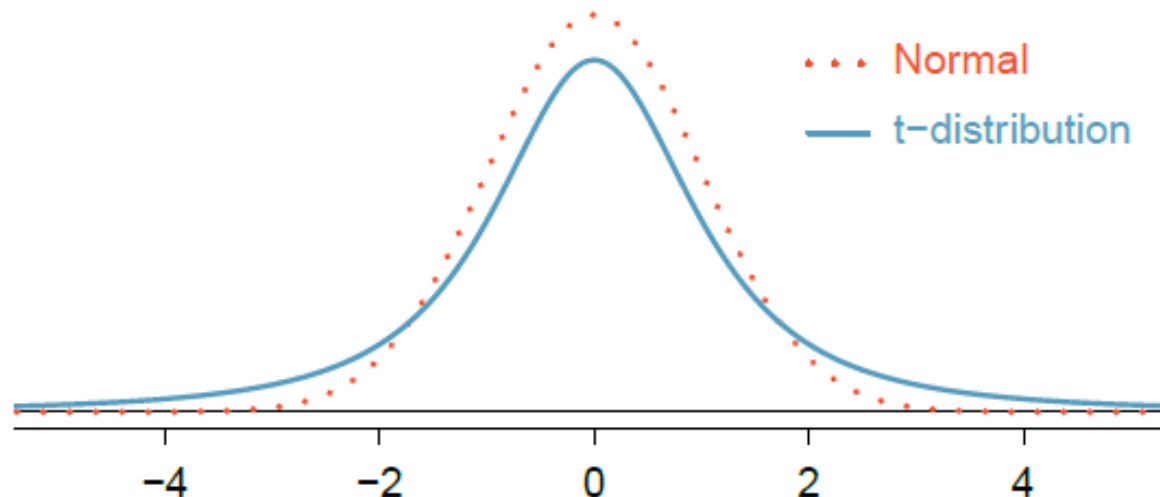
- A similar solution is used for computing the standard error of  $\bar{x}$ , using the sample standard deviation  $s$  in place of  $\sigma$

$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

- The estimate is less precise with smaller samples, which leads to problems using the normal distribution to model  $\bar{x}$ .

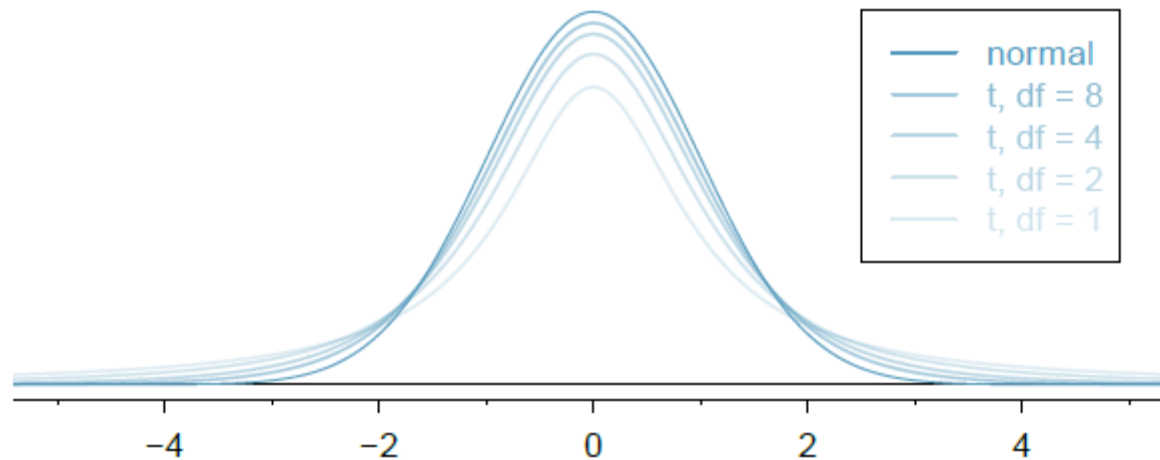
# The $t$ -distribution

- The  **$t$ -distribution**, while similar in shape to a normal distribution, has thicker tails so that observations are more likely to fall beyond two standard deviations from the mean.
  - This is correction is needed to account for using  $s$  in place of  $\sigma$  when calculating the SE.



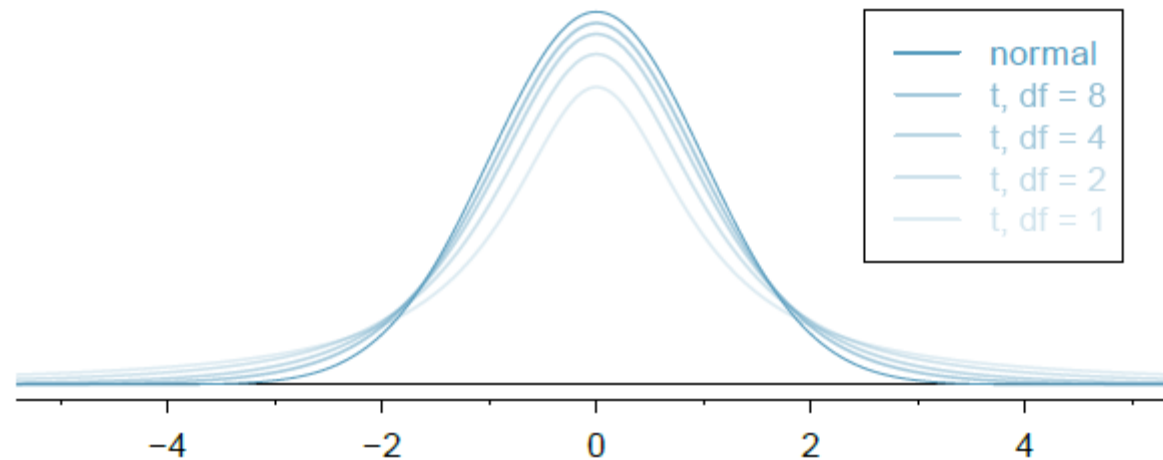
# The $t$ -distribution

- The  $t$ -distribution is always centered at zero and has one parameter, the degrees of freedom.
  - Degrees of freedom ( $df$ ) describes the form of the  $t$ -distribution.



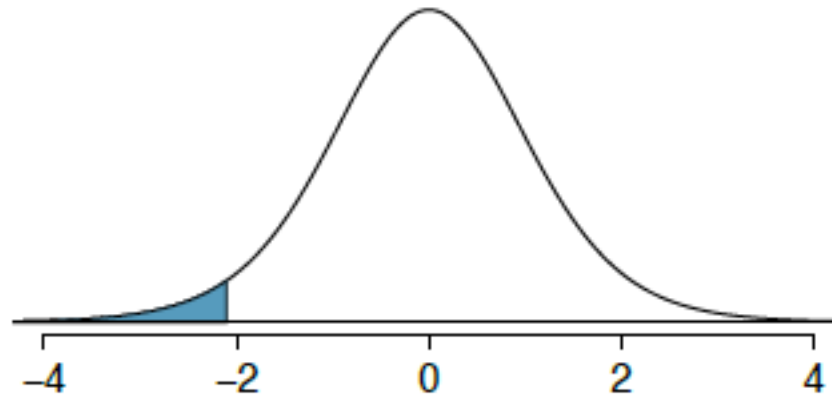
# The $t$ -distribution

- The chosen degrees of freedom is generally  $n - 1$ 
  - The more observations, the larger the degrees of freedom, the more normal the distribution



# The $t$ -distribution

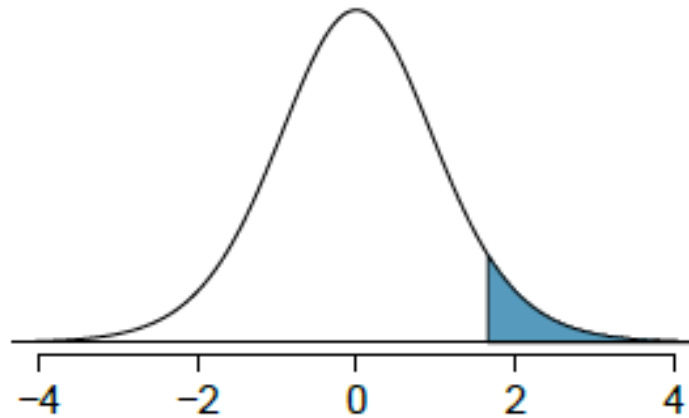
- The  $t$ -distribution below has 18 degrees of freedom.
  - What proportion falls below -2.10 (the shaded area)?



```
> pt(c(-2.10), 18, lower.tail=TRUE)
[1] 0.0250452
```

# The $t$ -distribution

- The  $t$ -distribution below has 20 degrees of freedom.
  - What proportion falls above 1.65 (the shaded area)?



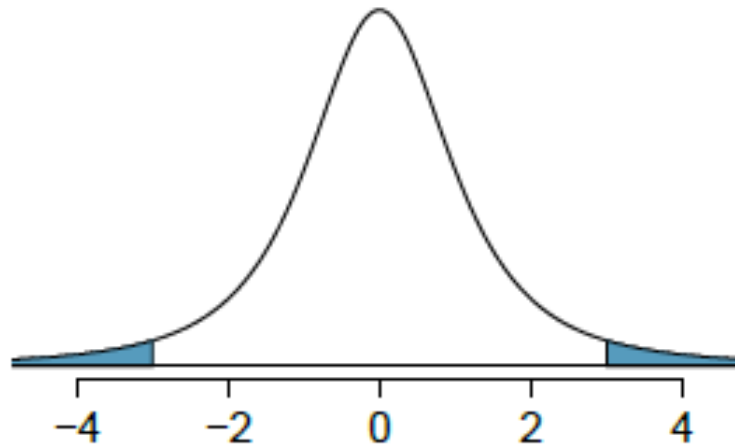
```
> pt(c(1.65), 20, lower.tail=FALSE)  
[1] 0.05728041
```

# The $t$ -distribution

Normal Distribution:

```
> pnorm(c(-3), lower.tail=TRUE) + pnorm(c(3), lower.tail=FALSE)
[1] 0.002699796
```

- The  $t$ -distribution below has 2 degrees of freedom.
  - What proportion is more than three units from the mean (the shaded area)?



```
> pt(c(-3), 2, lower.tail=TRUE) + pt(c(3), 2, lower.tail=FALSE)
[1] 0.09546597
```



# Confidence intervals for $t$ -distributions


- Using 19 randomly sampled dolphins, we will identify a confidence interval for the average mercury content in dolphin muscle.
- The minimum and maximum observed values below are used to evaluate if there are clear outliers.

$n$	$\bar{x}$	$s$	minimum	maximum
19	4.4	2.3	1.7	9.2

- Are the independence and normality conditions satisfied?
  - Independence: Random sampling
  - Normality: No clear outliers; All observations are within 2.5 standard deviations

# Confidence intervals for $t$ -distributions

- The normal model used  $z^*$  and the standard error to determine the width of a confidence interval.
- The confidence interval formula changes slightly when using the  $t$ -distribution:

$$\hat{p} \pm z^* \times SE_{\hat{p}}$$

$$\bar{x} \pm t^*_{df} \times \frac{s}{\sqrt{n}}$$

$n$	$\bar{x}$	$s$	minimum	maximum
19	4.4	2.3	1.7	9.2

# Confidence intervals for $t$ -distributions

- The standard error for the average mercury content in the  $n = 19$  dolphins:

$$SE = \frac{s}{\sqrt{n}} = \frac{2.3}{\sqrt{19}} = 0.528$$

- Appropriate degrees of freedom:

$$df = n - 1 = 19 - 1 = 18$$

$n$	$\bar{x}$	$s$	minimum	maximum
19	4.4	2.3	1.7	9.2

# Confidence intervals for $t$ -distributions

- Find  $t^*_{df}$  with 18 degrees of freedom and 95% confidence:
  - Proportion under tails = 5%
  - 2.5% under left tail, 2.5% under right tail

```
> qt(c(0.025), 18, lower.tail=FALSE)  
[1] 2.100922
```

- $t^*_{18} = 2.10$
- The cutoff for the upper tail is 2.10
  - Meaning the cutoff for the lower tail is -2.10
- 95% of the  $t$ -distribution with  $df=18$  lies within 2.10 units from the mean

# Confidence intervals for $t$ -distributions

- Computing the 95% confidence interval:

$$\bar{x} \pm t^*_{18} \times \frac{s}{\sqrt{n}} = 4.4 \pm 2.10 \times \frac{2.3}{\sqrt{19}} = (3.29, 5.51)$$

- 95% confident that the average mercury content of muscles in the dolphins is between 3.29 and 5.51.

$n$	$\bar{x}$	$s$	minimum	maximum
19	4.4	2.3	1.7	9.2

# Confidence intervals for $t$ -distributions

- *The FDA's webpage provides some data on mercury content of fish. Based on a sample of 15 croaker white fish, a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. We will assume these observations are independent and there exist no clear outliers.*
- Estimate the standard error:

$$SE = \frac{s}{\sqrt{n}} = \frac{0.069}{\sqrt{15}} = 0.0178$$

# Confidence intervals for $t$ -distributions

- *The FDA's webpage provides some data on mercury content of fish. Based on a sample of 15 croaker white fish, a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. We will assume these observations are independent and there exist no clear outliers.*
- Degrees of freedom:

$$df = n - 1 = 15 - 1 = 14$$

# Confidence intervals for $t$ -distributions

- *The FDA's webpage provides some data on mercury content of fish. Based on a sample of 15 croaker white fish, a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. We will assume these observations are independent and there exist no clear outliers.*
- $t^*_{df}$  with 14 degrees of freedom and 90% confidence:
  - Proportion under tails = 10%
  - 5% under left tail, 5% under right tail

```
> qt(c(0.05), 14, lower.tail=FALSE)
[1] 1.76131
```

- $t^*_{14} = 1.761$



# Confidence intervals for $t$ -distributions

- *The FDA's webpage provides some data on mercury content of fish. Based on a sample of 15 croaker white fish, a sample mean and standard deviation were computed as 0.287 and 0.069 ppm (parts per million), respectively. The 15 observations ranged from 0.18 to 0.41 ppm. We will assume these observations are independent and there exist no clear outliers.*
- 90% confidence interval:

$$\bar{x} \pm t^*_{14} \times \frac{s}{\sqrt{n}} = 0.287 \pm 1.761 \times \frac{0.069}{\sqrt{15}} = (0.256, 0.318)$$

- 90% the actual population mean is between 0.256 ppm and 0.318 ppm.

# One-sample $t$ -tests

- A  **$t$ -test** is a hypothesis test for a one-sample mean.

*The average time for all runners who finished the Cherry Blossom Race in 2006 was 93.29 minutes (93 minutes and about 17 seconds). We want to determine (using data from 100 randomly sampled participants in the 2017 Cherry Blossom Race) whether runners in this race are getting faster or slower, versus the possibility that there has been no change.*

$H_0$ : The average time was the same in 2017 as it was in 2006 ( $\mu = 93.29$  minutes)

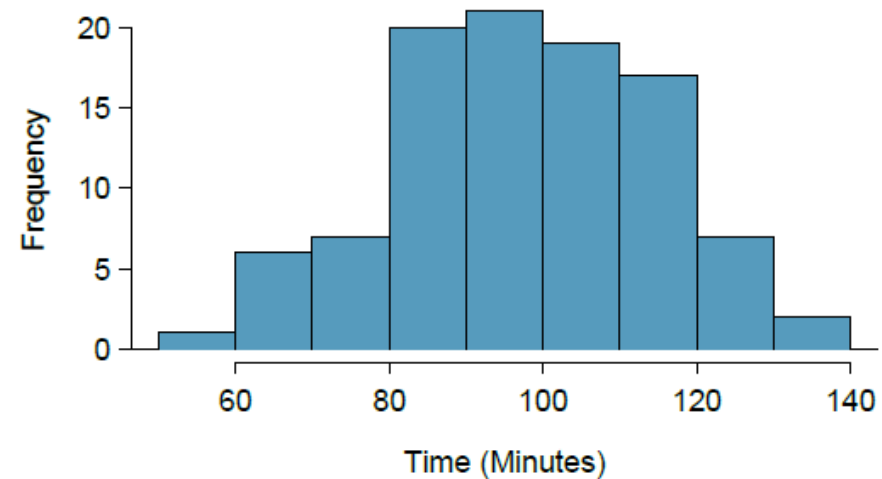
$H_A$ : The average time increased or decreased in 2017 compared to 2006 ( $\mu \neq 93.29$  minutes)

# One-sample $t$ -tests

- The process for completing a hypothesis test for the one-sample mean is nearly identical to completing a hypothesis test for a single proportion.
- First, find the Z-score using the observed value, null value, and standard error.
  - It is called a *T-score* since we use a  $t$ -distribution.
- Then, find the p-value using the same ideas we used previously
  - Find one-tail area under the sampling distribution and double it.

# One-sample $t$ -tests

- With a sample size of 100 ( $n \geq 30$ ),  $\bar{x}$  will be nearly normal if there are no extreme outliers.



- No apparent outliers

# One-sample $t$ -tests

- We'll say the sample mean of the 100 racers from 2017 is calculated to be 97.32 minutes and the sample standard deviation is 16.98 minutes.
  - The average run time in 2006 was said to be 93.29 minutes
- To find the T-score, first calculate the standard error

$$SE = \frac{s}{\sqrt{n}} = \frac{16.98}{\sqrt{100}} = 1.70$$

# One-sample $t$ -tests

- The T-score, can now be computed using the sample mean (97.32), null value (93.29), and SE (1.70)

$$T = \frac{97.32 - 93.29}{1.70} = 2.37$$

- $df = 100 - 1 = 99$

# One-sample $t$ -tests

- The p-value is calculated to be 0.02

```
> 2*pt(c(2.37), 99, lower.tail=FALSE)  
[1] 0.01972642
```

- Less than 0.05 significance, the null hypothesis is rejected.
  - Runners in the 2017 race were slower than runners in the 2006 race

# One-sample $t$ -tests

## 1. Prepare

- Identify the parameter of interest, list hypotheses, identify the significance level, and identify  $\bar{x}$ ,  $s$  and  $n$ .

## 2. Check

- Verify conditions to ensure  $\bar{x}$  is nearly normal.

## 3. Calculate

- If the conditions hold, compute the standard error, compute the T-score, and identify the p-value.

## 4. Conclude

- Evaluate the hypothesis test by comparing the p-value to  $\alpha$  and provide a conclusion in the context of the problem.



# Paired Data

- Two sets of observations are *paired* if each observation in one set has a special connection with exactly one observation in the other data set.
  - The textbook illustrates this by comparing the cost of other textbooks between two vendors (the UCLA bookstore and Amazon)

	subject	course_number	bookstore	amazon	price_difference
1	American Indian Studies	M10	47.97	47.45	0.52
2	Anthropology	2	14.26	13.55	0.71
3	Arts and Architecture	10	13.50	12.53	0.97
⋮	⋮	⋮	⋮	⋮	⋮
68	Jewish Studies	M10	35.96	32.40	3.56

# Paired Data

- When analyzing paired data, it is often useful to look at the difference in outcomes.
  - bookstore – amazon = price\_difference

	subject	course_number	bookstore	amazon	price_difference
1	American Indian Studies	M10	47.97	47.45	0.52
2	Anthropology	2	14.26	13.55	0.71
3	Arts and Architecture	10	13.50	12.53	0.97
⋮	⋮	⋮	⋮	⋮	⋮
68	Jewish Studies	M10	35.96	32.40	3.56

# Paired Data

- Histogram of price differences:



- To analyze a paired data set, analyze the differences in  $n$ ,  $\bar{x}$ , and  $s$

$n_{diff}$	$\bar{x}_{diff}$	$s_{diff}$
68	3.58	13.42

# Paired Data

## 1. Prepare

- $n_{diff} = 68$
- $\bar{x}_{diff} = 3.58$
- $s_{diff} = 13.42$
- Parameter of interest: price difference

$H_0$ : There is no difference in the average textbook price ( $\mu_{diff} = 0$ )

$H_A$ : There is a difference in the average textbook price ( $\mu_{diff} \neq 0$ )

# Paired Data

## 2. Check

- Verify conditions to ensure  $\bar{x}_{diff}$  is nearly normal.
- Observations are random samples
- There are some outliers, but none are particularly extreme.
- Normality of  $\bar{x}_{diff}$  is satisfied.



# Paired Data

## 3. Calculate

- Compute the standard error, the T-score, and identify the p-value.

- $SE = \frac{s_{diff}}{\sqrt{n_{diff}}} = \frac{13.42}{\sqrt{68}} = 1.63$

- $T = \frac{3.58 - 0}{1.63} = 2.20$

- p-value = 0.031

```
> 2*pt(c(2.20), 67, lower.tail=FALSE)
[1] 0.03125996
```

# Paired Data

## 4. Conclude

- Evaluate the hypothesis test by comparing the p-value to  $\alpha$  and provide a conclusion in the context of the problem.
- 0.031 is less than 0.05 significance; reject the null hypothesis.
- Amazon prices were, on average, lower than the UCLA bookstore's prices for UCLA courses

# Paired Data

```
> qt(c(0.025), 67, lower.tail=FALSE)
[1] 1.996008
```

- Computing the 95% confidence interval:

$$\bar{x}_{diff} \pm t^*_{67} \times \frac{s_{diff}}{\sqrt{n_{diff}}} = 3.58 \pm 2.00 \times 1.63 = (0.32, 6.84)$$

- 95% confident that Amazon is between \$0.32 and \$6.84 less expensive than the UCLA bookstore for UCLA course books.