STA258H5
University of Toronto Mississauga

Al Nosedal and Omid Jazi

Winter 2023

ONE SAMPLE CONFIDENCE INTERVALS ON A MEAN WHEN THE POPULATION VARIANCE IS UNKNOWN

(more realistic)

Large-Sample Confidence interval for μ

Parameter : μ.

Confidence interval :

$$\bar{Y} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right).$$

Valid if

1. Random sample
2. Independent and identically distributed observations
3. n is large enough for CLT to apply

Small-Sample Confidence interval for μ

Parameter : μ.

Confidence interval ($\nu =$ df) :

$$\bar{Y} \pm t_{\alpha/2} \left(\frac{S}{\sqrt{n}} \right), \qquad \nu = n - 1.$$

Valid if:

1. independent and identically distributed observations
2. Random sample
3. population **must** have Normal distribution (CLT does not apply)

σ known

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

One-Sample CI on pop. mean μ

2. When σ is <u>not</u> known

$$\bar{x} \pm t_{(n-1,\alpha/2)} \frac{s}{\sqrt{n}}$$

$\frac{s}{\sqrt{n}}$ ← standard error

sample st dev

$t_{(n-1,\alpha/2)} \frac{s}{\sqrt{n}}$ ← margin of error

(CI's usually wider when σ unknown)

CASES WHERE VALID

$$\bar{X} \pm t_{(n-1,\alpha/2)} \cdot \frac{S}{\sqrt{n}}$$

Valid for

1. large samples where population is normal
2. " " " " " not normal
   (By CLT)
3. small " " " " normal

large : $n \geq 30$

Independence Assumption

The data values should be independent. There's really no way to check independence of the data by looking at the sample, but we should think about whether the assumption is reasonable.

Randomization condition

The data arise from a random sample or suitably randomized experiment. Randomly sampled data - especially data from a Simple Random Sample - are ideal.

Normal Population assumption

- For very small samples ($n < 15$ or so), the data should follow a Normal model pretty closely. If you do find outliers or strong skewness, don't use this method.
- For moderate samples (n between 15 and 40 or so), the t-method will work well as long as the data is unimodal and reasonably symmetric. Make a histogram, boxplot, or Q-Q plot to check.
- When the sample size is larger than 40 or 50, the t method is safe to use unless the data are extremely skewed. Make a histogram, boxplot, or Q-Q plot to check.

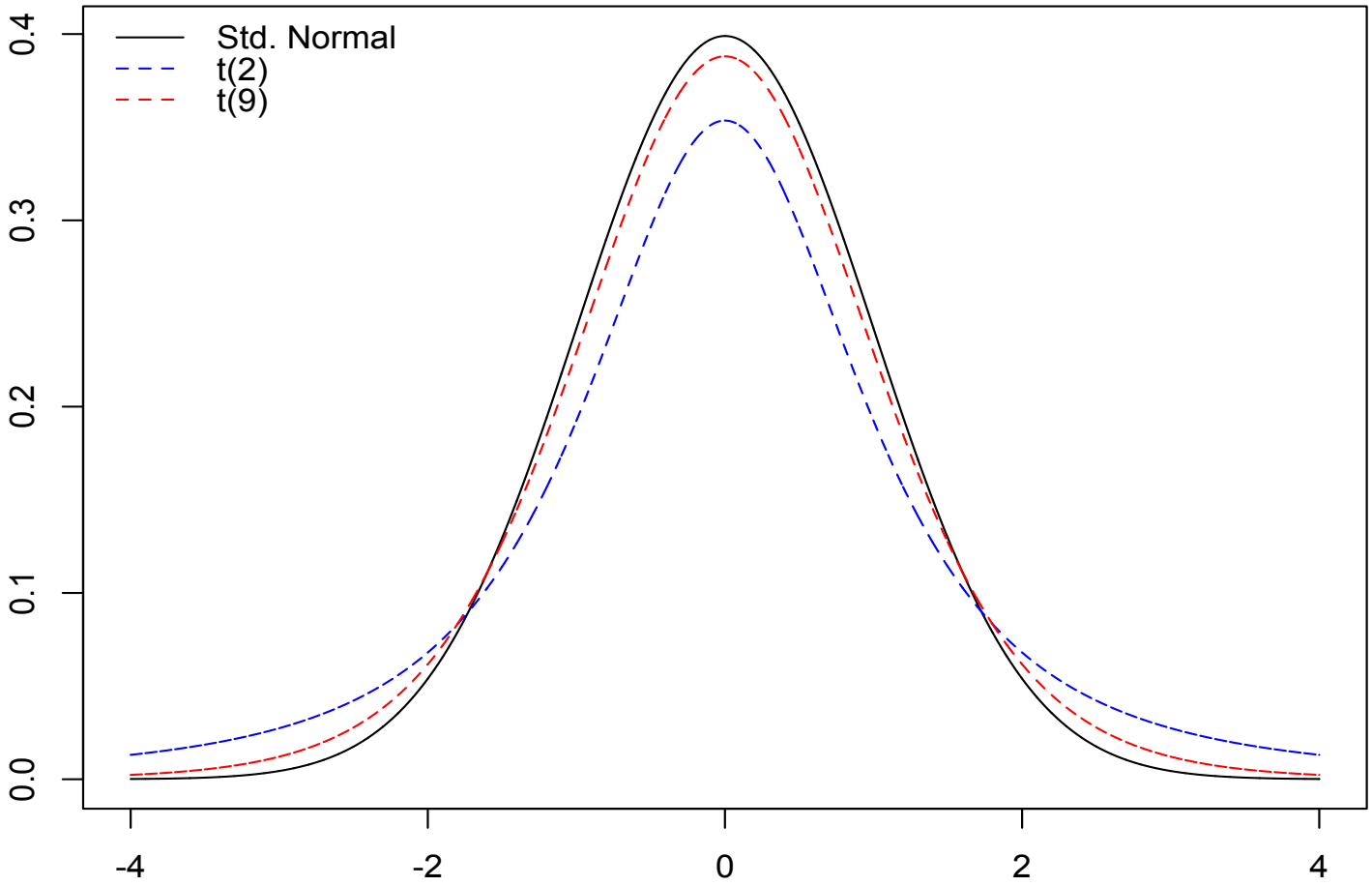
Standard Error

When the standard deviation of a statistic is estimated from data, the result is called the *standard error* of the statistic. The standard error of the sample mean $\bar{x}$ is $\frac{s}{\sqrt{n}}$.

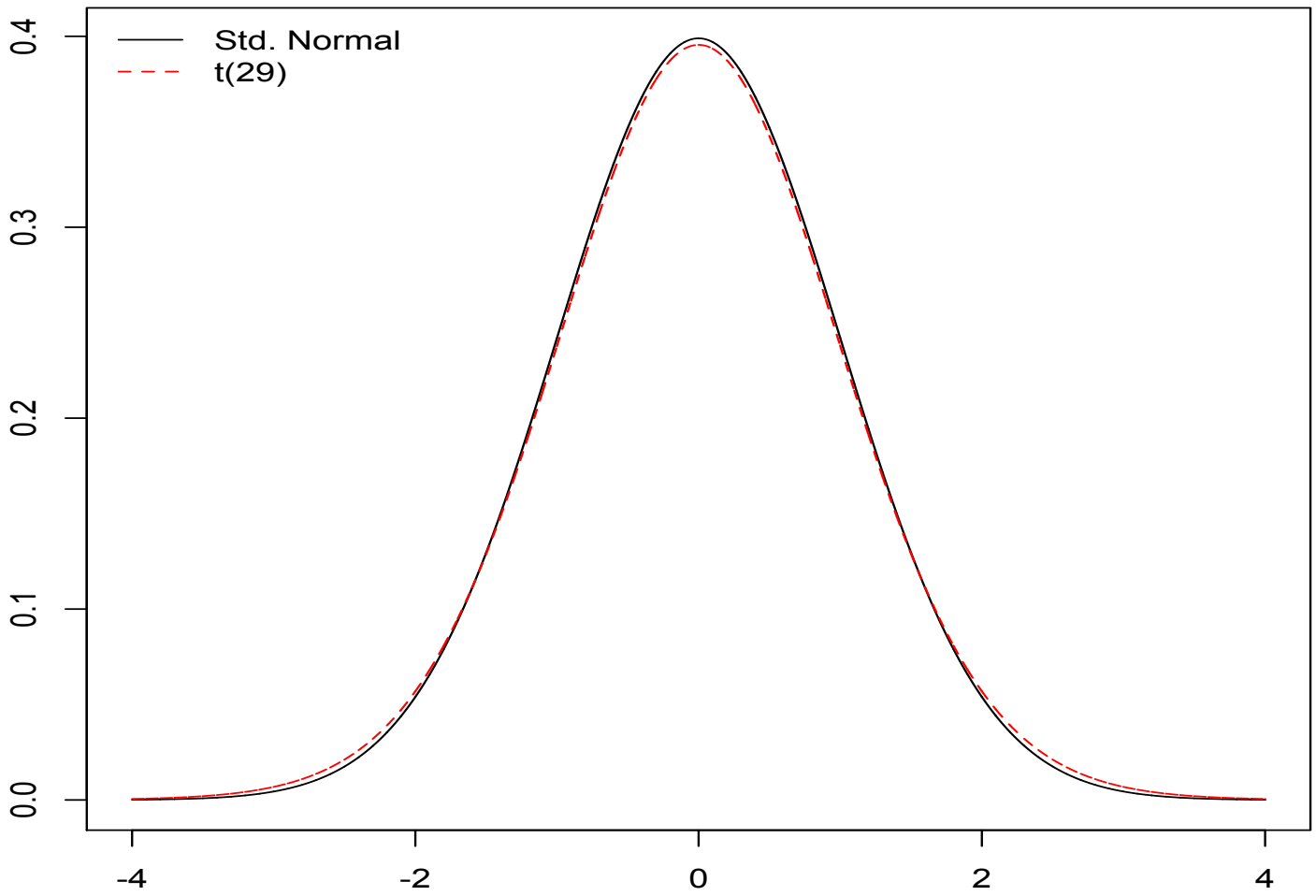
The t distributions

- The density curves of the t distributions are similar in shape to the Standard Normal curve. They are symmetric about 0, single-peaked, and bell-shaped.
- The spread of the t distributions is a bit greater than of the Standard Normal distribution. The t distributions have more probability in the tails and less in the center than does the Standard Normal. This is true because substituting the estimate s for the fixed parameter σ introduces more variation into the statistic.
- As the degrees of freedom increase, the t density curve approaches the $N(0, 1)$ curve ever more closely. This happens because s estimates σ more accurately as the sample size increases. So using s in place of σ causes little extra variation when the sample is large.

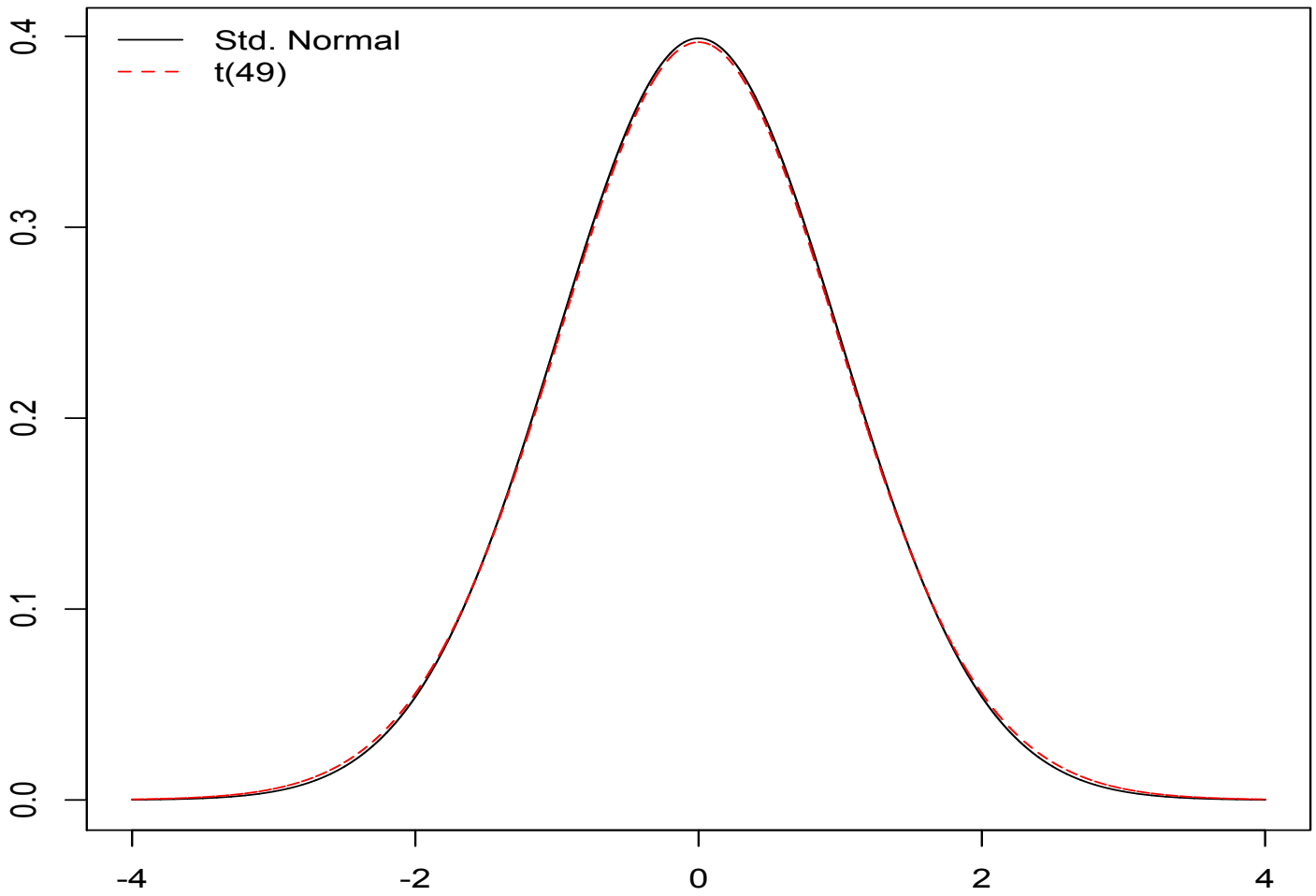
Density curves



Density curves



Density curves



Example: Ancient air

The composition of the earth's atmosphere may have changed over time. To try to discover the nature of the atmosphere long ago, we can examine the gas in bubbles inside ancient amber. Amber is tree resin that has hardened and been trapped in rocks. The gas in bubbles within amber should be a sample of the atmosphere at the time the amber was formed. Measurements on specimens of amber from the late Cretaceous era (75 to 95 million years ago) give these percents of nitrogen:
63.4 65 64.4 63.3 54.8 64.5 60.8 49.1 51.0
Assume (this is not yet agreed on by experts) that these observations are an SRS from the late Cretaceous atmosphere. Use a 90% confidence interval to estimate the mean percent of nitrogen in ancient air (Our present-day atmosphere is about 78.1% nitrogen).

Assume normality.

Want 90% CI for nitrogen

Data given : 63.4, 65, — — —, 51.0

($n = 9$ (small))

No info regarding σ known

↳ calculate $\bar{x}$, s

$\bar{x} = 59.589$, $s = 6.255$, $n = 9$ (normal)

Since σ unknown

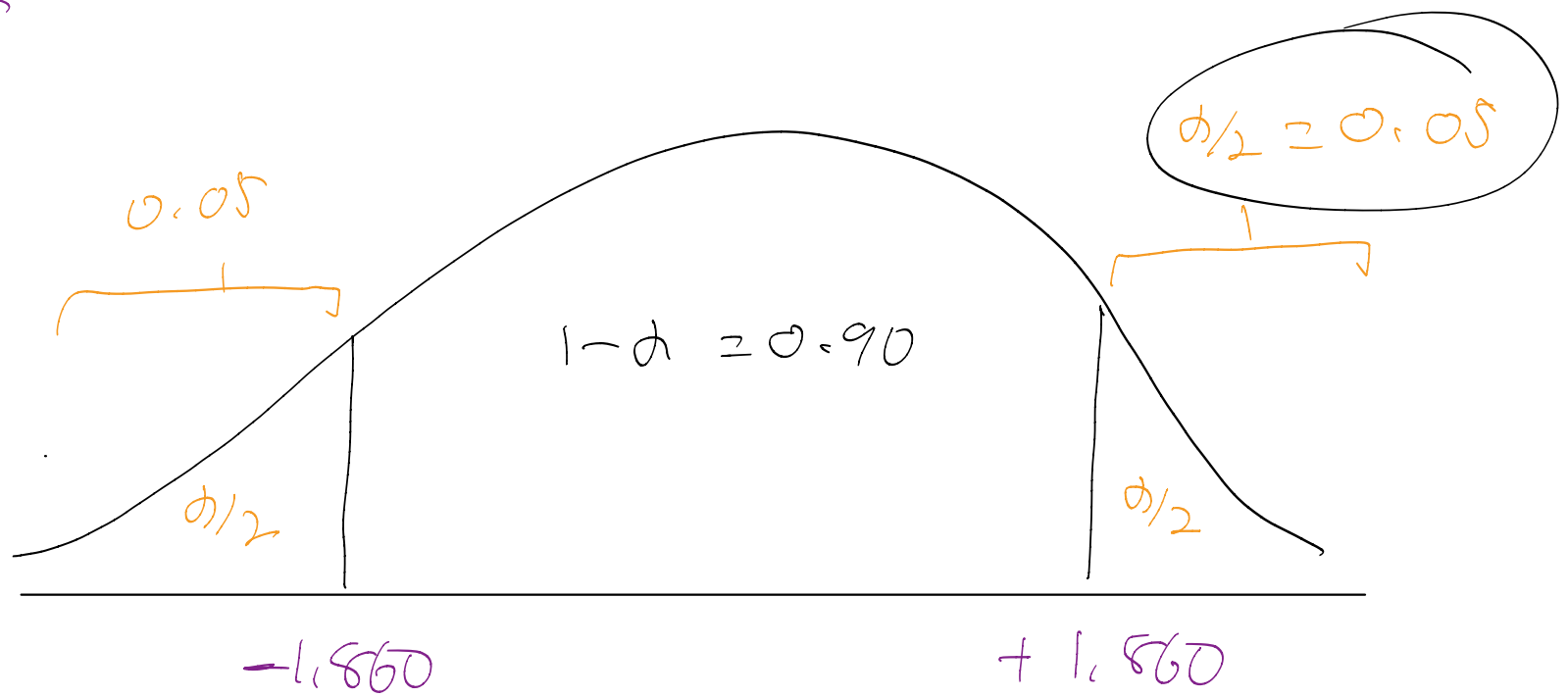
$$\bar{x} \pm t_{\alpha/2,\, n-1} \cdot \frac{s}{\sqrt{n}}$$

?

$t_{(n-1,\alpha/2)}$ for 90% CI with $n=9$

$n-1 = 9-1 = 8$

t at 8 df



$t_{(n-1,\alpha/2)} = t_{(8,0.05)} = 1.860$

$\bar{x} = 59.589$, $s = 6.255$, $n = 9$ $t = 1.860$

Since σ unknown

$$\bar{x} \pm t_{(\alpha/2,\, n-1)} \cdot \frac{s}{\sqrt{n}}$$

$$= 59.589 \pm (1.860) \cdot \left(\frac{6.255}{\sqrt{9}}\right)$$

$$= 59.589 \pm 3.8782$$

$$= (55.711,\ 63.467)$$

Interp : exercise

Compare to today's level

Solution

$\mu =$ mean percent of nitrogen in ancient air. We will estimate μ with a 90% confidence interval.
With $\bar{x} = 59.5888$, $s = 6.2552$, and $t^* = 1.860$ ($df = 9 - 1 = 8$), the 90% confidence interval for μ is

$$59.5888 \pm 1.860 \left(\frac{6.2552}{\sqrt{9}} \right)$$

$$59.5888 \pm 3.8782$$

$$55.7106 \text{ to } 63.4670$$

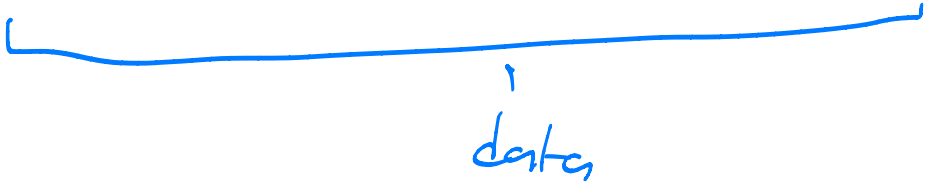
```r
# Step 1. Entering data;

nitrogen=c(63.4 ,65,64.4,63.3,54.8,
64.5,60.8,49.1,51.0);

# Step 2. Constructing CI;

t.test(nitrogen,conf.level=0.90);
```

```
##
##  One Sample t-test
##
## data:  nitrogen
## t = 28.578, df = 8, p-value = 2.43e-09
## alternative hypothesis: true mean is not equal to 0
## 90 percent confidence interval:
##  55.71155 63.46622
## sample estimates:
## mean of x
##  59.58889
```

Exercise

Most owners of digital cameras store their pictures on the camera. Some will eventually download these to a computer or print them using their own printers or a commercial printer. A film-processing company wanted to know how many pictures were stored on computers. A random sample of 10 digital camera owners produced the data given here. Estimate with 95% confidence the mean number of pictures stored on digital cameras.

25 6 22 26 31 18 13 20 14 2.



Example (Slide 35)

$n = 10$. Have rain data. Want a 95% CI on pop mean

Is σ known or unknown?

No indication σ known (assume σ not known)

Data: 25, 6, 22, 26, 31, 18, 13, 20, 14, 2

CI's on pop mean

σ known $\rightarrow$ Z

σ unknown $\rightarrow$ t

CI on the mean (σ is not known)

$$\bar{X} \pm t_{(n-1,\alpha/2)} \cdot \frac{s}{\sqrt{n}}$$

(? ? ?)

use data to find $\bar{x}$ and s.

Sample mean ($\bar{x}$)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{25+6+22+\ldots+14+2}{10} = \frac{177}{10} = 17.7$$

Sample variance (s^2)

① Direct

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{(25-17.7)^2 + (6-17.7)^2 + \ldots + (2-17.7)^2}{10-1}$$

$$= \frac{742}{9} = 82.4556$$

② Indirectly (several steps)

$$\sum_{i=1}^{10} x_i = 25 + 6 + 22 + \cdots + 14 + 2 = 177$$

$$\sum_{i=1}^{10} x_i^2 = 25^2 + 6^2 + 22^2 + \cdots + 14^2 + 2^2 = 3875$$

$$s^2 = \frac{\left(\sum x_i^2\right) - \frac{\left(\sum x_i\right)^2}{n}}{n-1} = \frac{3875 - \frac{(177)^2}{10}}{10-1}$$

$$s^2 = 82.4556$$

Sample St. dev (s)

$$s = \sqrt{s^2} = \sqrt{82.4556} = 9.081$$

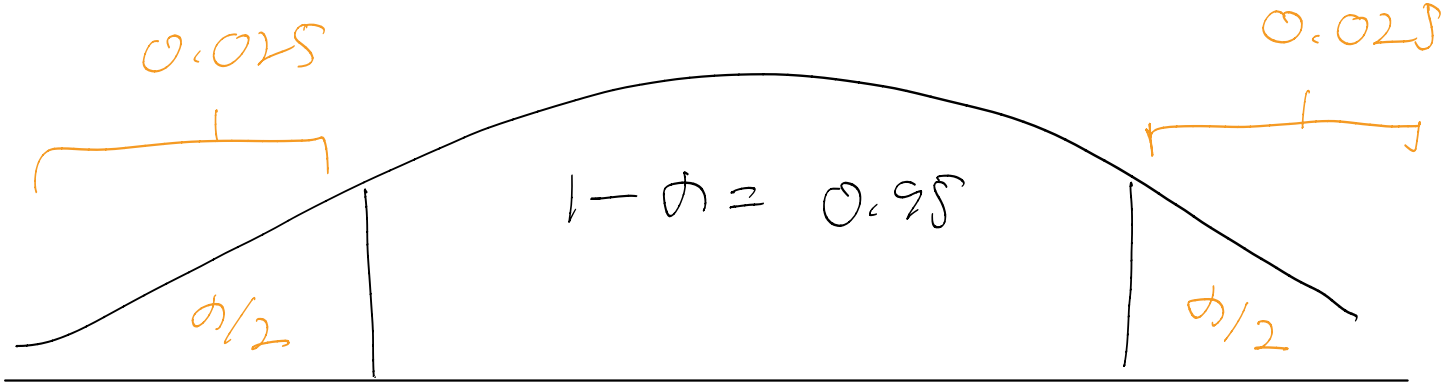
$\bar{x} = 17.7, \quad s = 9.0815, \quad n = 10$

$$\bar{x} \pm \underbrace{t_{(n-1,\,\alpha/2)}}_{?} \cdot \frac{s}{\sqrt{n}}$$

Find $t_{(n-1,\,\alpha/2)}$ for a 95% CI with $n = 10$

$df = n - 1$
$= 10 - 1$
$= 9$

t distrib at 9 df



$1 - \alpha = 0.95$
$\alpha = 0.05$
$\frac{\alpha}{2} = 0.025$

$t_{(n-1,\,\alpha/2)} = t_{(9,\,0.025)} = 2.262$

$\bar{x} = 17.7$, $s = 9.081$, $n = 10$, $t_{(n-1, \alpha/2)} = t_{(9,\ 0.025)} = 2.262$

$$\bar{x} \pm t_{(n-1, \alpha/2)} \cdot \frac{s}{\sqrt{n}} = 17.7 \pm 2.262 \left(\frac{9.081}{\sqrt{10}} \right)$$

$$= 17.7 \pm 6.495$$

$$= (11.205,\ 24.195)$$

Interp: (give in context)

We are 95% confident the mean number of images stored is between 11.205 and 24.195

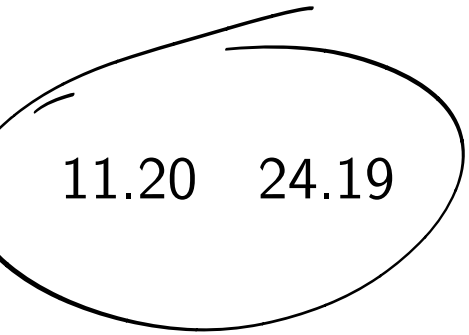
- How does conf level affect width of a CI
- For example above suppose pop. st. dev $\sigma = 9.081$. How would this affect the CI?

Solution

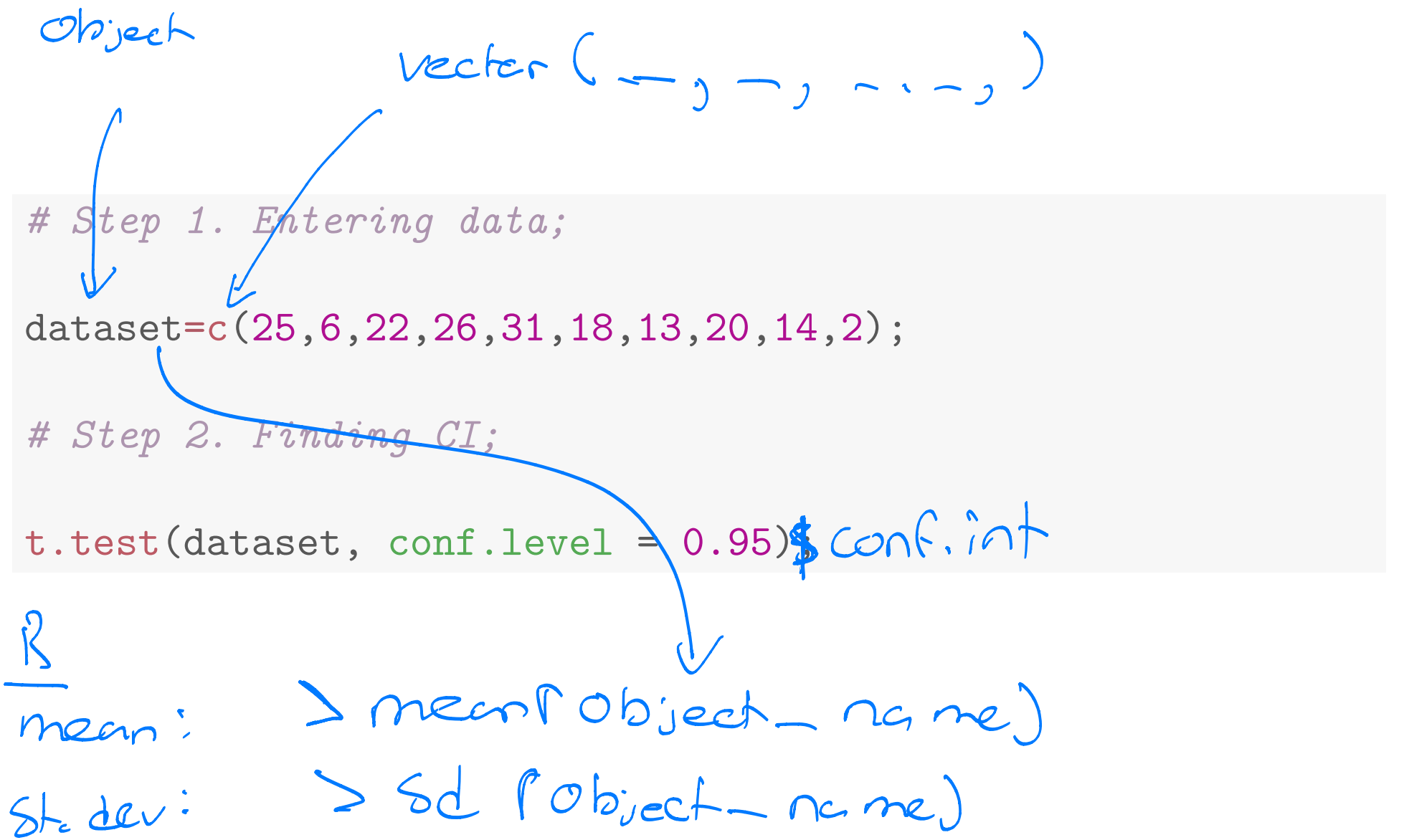
μ = mean number of pictures stored on digital cameras. We will estimate μ with a 95% confidence interval.
$\bar{x} = 17.7$, $s = 9.080504$, and $df = 10 - 1 = 9$. From our table, $t^* = 2.262$ (or $t(9) = 2.262$).
margin of error $= 2.262\left(\frac{9.080504}{\sqrt{10}}\right) = 6.49535$.
$LCL = \bar{x} -$ margin of error $= 17.7 - 6.49535 = 11.20465$
$UCL = \bar{x} -$ margin of error $= 17.7 + 6.49535 = 24.19535$

Solution

95 percent confidence interval:



R Code


object
vector(_, _, ..., _)
# Step 1. Entering data;
dataset=c(25,6,22,26,31,18,13,20,14,2);
# Step 2. Finding CI;
t.test(dataset, conf.level = 0.95)$conf.int
R
mean: > mean(object_name)
St. dev: > sd(object_name)

R Code

```
##
##  One Sample t-test
##
## data:  dataset
## t = 6.164, df = 9, p-value = 0.0001659
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  11.2042 24.1958
## sample estimates:
## mean of x
##      17.7
```

Electric Insulators

A manufacturing company produces electric insulators. If the insulators break when in use, a short circuit is likely. To test the strength of the insulators, you carry out destructive testing to determine how much force is required to break the insulators. You measure force by observing how many pounds are applied to the insulator before it breaks. The table on the next slide lists 30 values from this experiment. Construct a 95% confidence interval estimate for the population mean force required to break the insulator.

Data

1870 1728 1656 1610 1634 1784 1522 1696 1592 1662
1866 1764 1734 1662 1734 1774 1550 1756 1762 1866
1820 1744 1788 1688 1810 1752 1680 1810 1652 1736

R Code

μ = population mean force required to break electric insulators.

```
# Step 1. Entering data;

dataset=c(1870,1728,1656,1610,1634,1784,1522,1696,
1592,1662,1866,1764,1734,1662,1734,1774,1550,1756,
1762,1866,1820,1744,1788,1688,1810,1752,1680,1810,
1652,1736);

# Step 2. Finding CI;

t.test(dataset, conf.level = 0.95);
```

R Code

```
##
##  One Sample t-test
##
## data:  dataset
## t = 105.41, df = 29, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  1689.961 1756.839
## sample estimates:
## mean of x
##    1723.4
```

Solution

$\mu =$ population mean force required to break electric insulators. We will estimate μ with a 95% confidence interval.
With $\bar{x} = 1723.4$, $s = 89.5508$, and $t^* = 2.045$ ($df = 30 - 1 = 29$), the 95% confidence interval for μ is

$$1723.4 \pm 2.045 \left(\frac{89.5508}{\sqrt{30}} \right)$$

Homework?

The operations manager of a production plant would like to estimate the mean amount of time a worker takes to assemble a new electronic component. After observing 120 workers assembling similar devices, she noticed that their average time was 16.2 minutes (with standard deviation 3.6 minutes). Construct a 92% confidence interval for the mean assembly time. State all necessary assumptions.

Tax Collected from Audited Returns

In 2010, 142,823,000 tax returns were filed in the United States. The Internal Revenue Service (IRS) examined 1.107%, or 1,581,000, of them to determine if they were correctly done. To determine how well the auditors are performing, a random sample of these returns was drawn and the additional tax was reported, see file taxes.txt. Estimate with 95% confidence the mean additional income tax collected from the 1,581,000 files audited.

Reading our data

```
#Step 1. Importing data;

# url of taxes;
url="https://mcs.utm.utoronto.ca/~nosedal/data/taxes.txt"

taxes_data= read.table(url,header=TRUE);

names(taxes_data);

# first 6 observations from file
head(taxes_data);

taxes = taxes_data$Taxes;
```

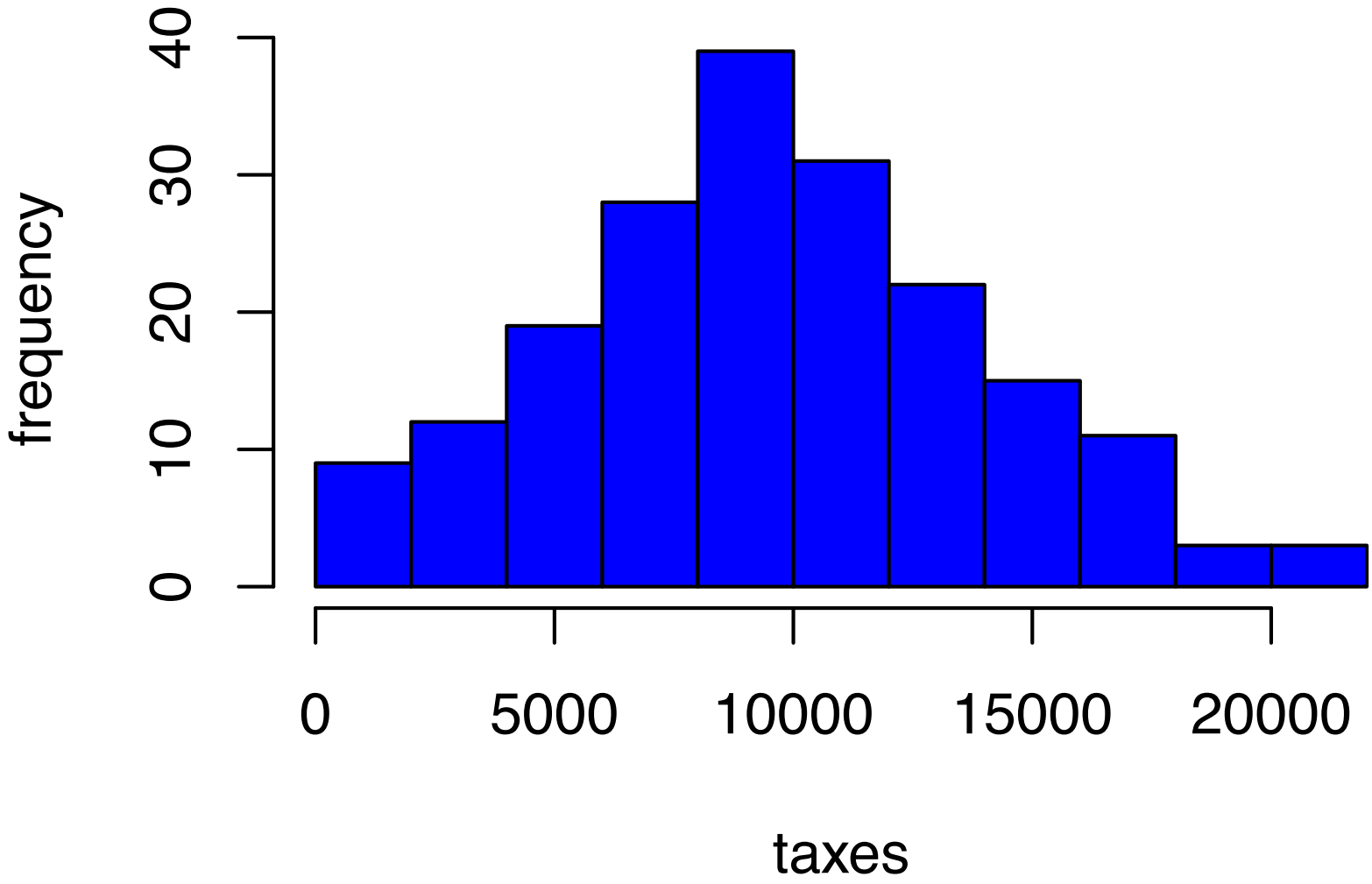
Reading our data

```
## [1] "Taxes"
##      Taxes
## 1 13069.55
## 2  6915.39
## 3 16103.88
## 4  1088.93
## 5  2895.24
## 6  2365.28
```

Histogram

```
hist(taxes,
main="Histogram for our example ",
xlab="taxes", ylab="frequency",
col="blue");
```

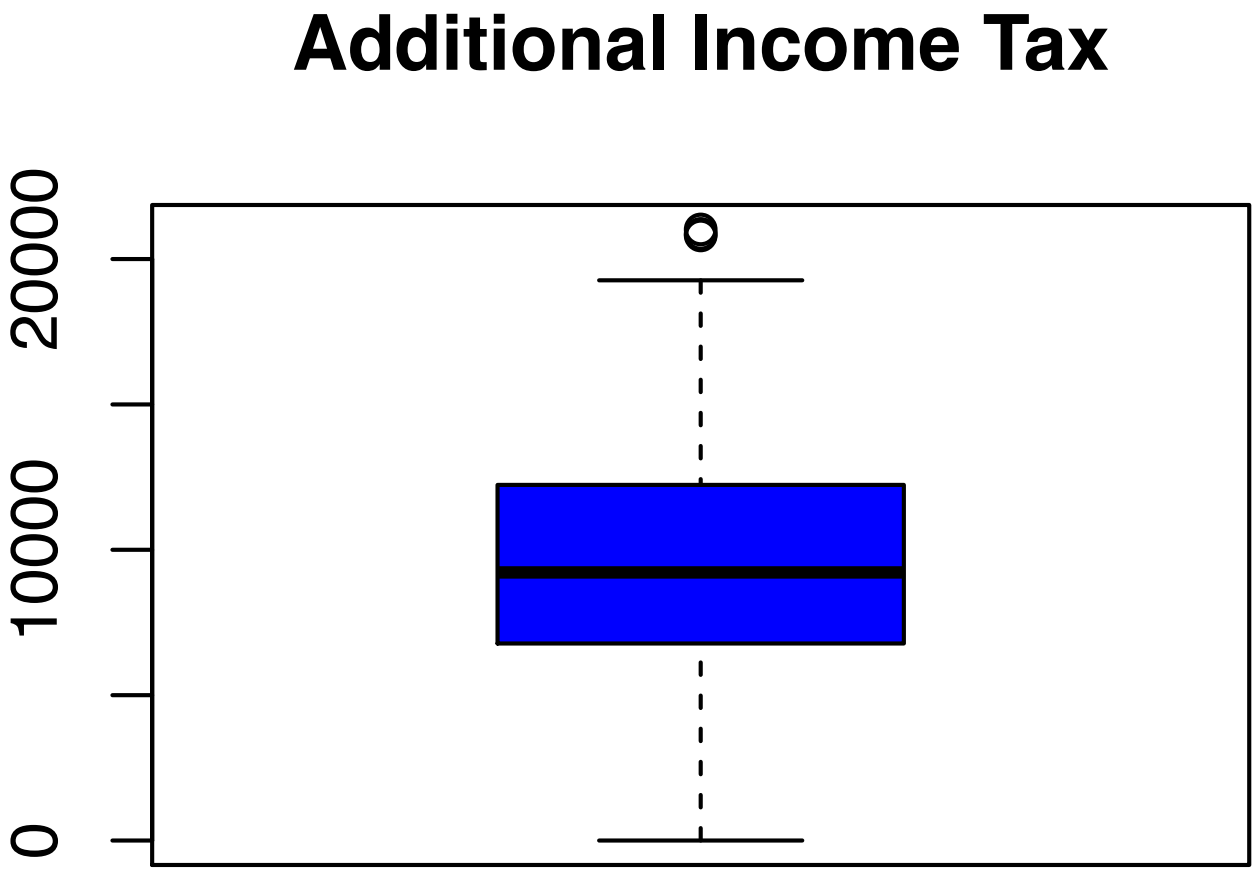
Histogram for our example



Boxplot

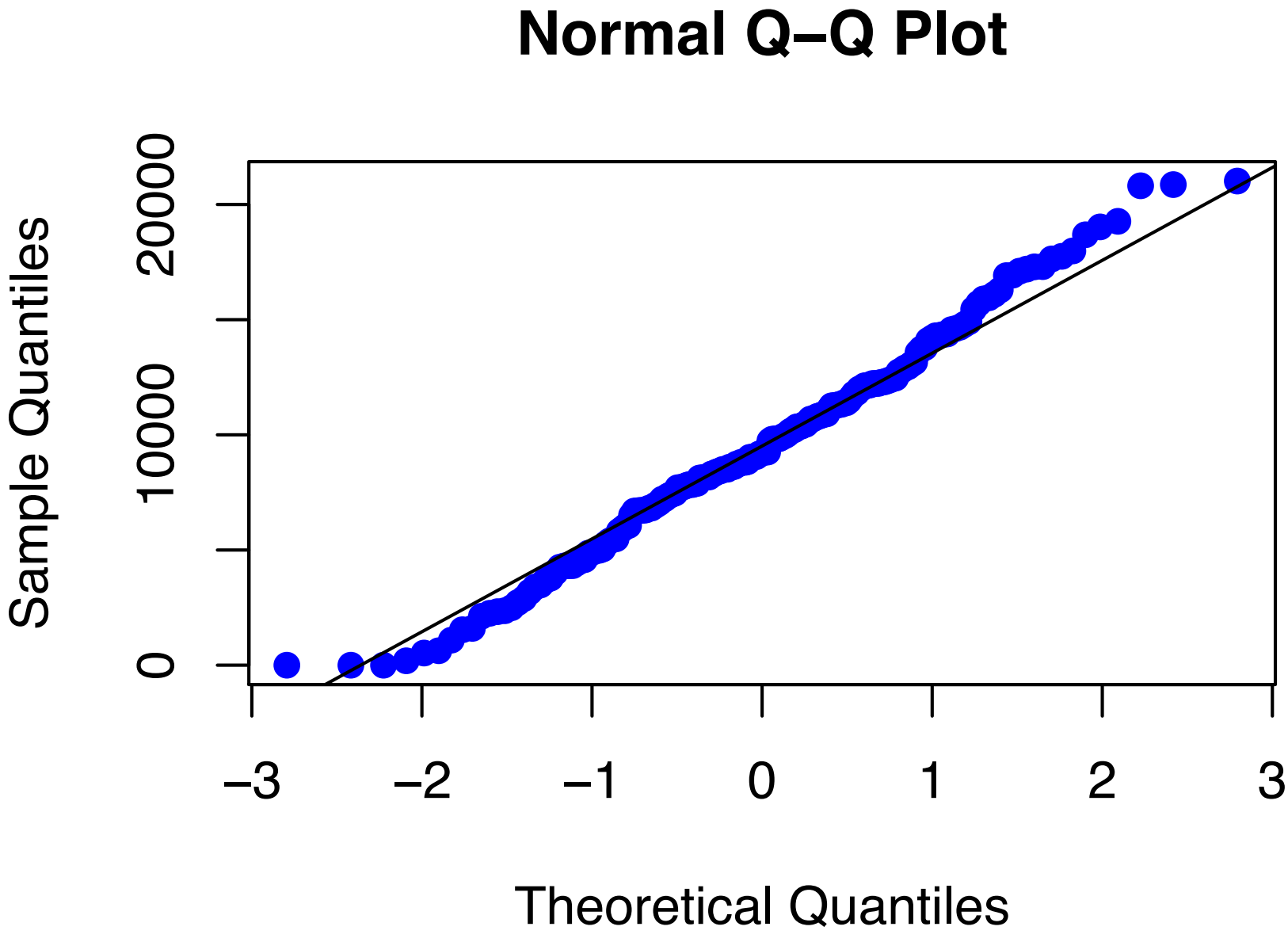
```
boxplot(taxes,
main="Additional Income Tax",
col="blue");
```

Boxplot



R Code

```
## Q-Q plot (using R function);

qqnorm(taxes,col="blue",pch=19);
qqline(taxes);
```


Normal Q–Q Plot
Sample Quantiles
0
10000
20000
−3
−2
−1
0
1
2
3
Theoretical Quantiles

Finding CI

```r
# Step 2. Constructing CI;

t.test(taxes,conf.level=0.95);
```

Finding CI

```
##
##  One Sample t-test
##
## data:  taxes
## t = 29.345, df = 191, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   8886.932 10167.721
## sample estimates:
## mean of x
##  9527.326
```

We estimate that the mean additional tax collected lies between \$8,887 and \$10,168 (with 95% confidence).

A few final comments

When we introduced the Student t-distribution, we pointed out that the t-statistic is Student t-distributed if the population from which we've sampled is Normal. However, statisticians have shown that the mathematical process that derived the Student t-distribution is **robust**, which means that if the population is non-Normal, the results of the confidence interval estimate are still valid provided that the population is **not extremely non-Normal**. Our histogram, boxplot, and Q-Q plot suggest that our variable of interest is not extremely non-Normal, and in fact, may be Normal.