Intro to Social Science Data Analysis

Lecture 7: Overview of Statistical Inference

Christopher Gandrud

October 16, 2012

Assignment 2

Due: Friday 19 October

Describe at least **3** variables in a data set.

You need to select a **range of descriptive statistical tools**. The tools should include both **numerical descriptive statistics** and **graphics**.

These tools should describe the variables':

- central tendency,
- variation,
- their relationships with the other variables.

The descriptions need to be discussed **in paragraph form**.

The description must be **reproducible**. So you should email me the link to a Dropbox folder with:

- the `.csv` data set,
- the `.Rmd` R markdown file,
- the final `.html` file.

Principles of Quantitative Graphics (1)

What are some of the goals that quantitative graphics should try to achieve?

Principles of Quantitative Graphics (2)

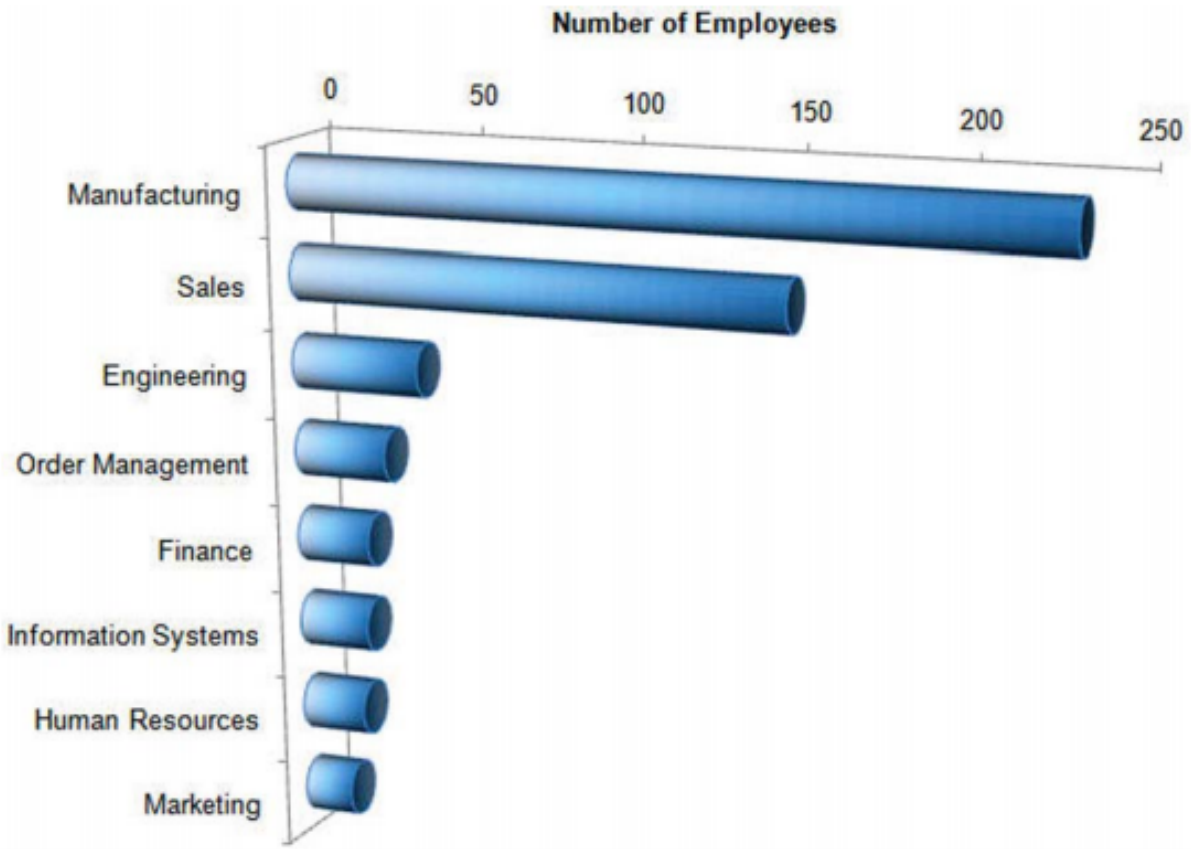
What are some things you should not do in quantitative graphics?

In particular, how does the **data-ink ratio** help us create effective graphics?

Remember:

$$\text{Data Ink Ratio} = \frac{\text{data} - \text{ink}}{\text{total ink}}$$

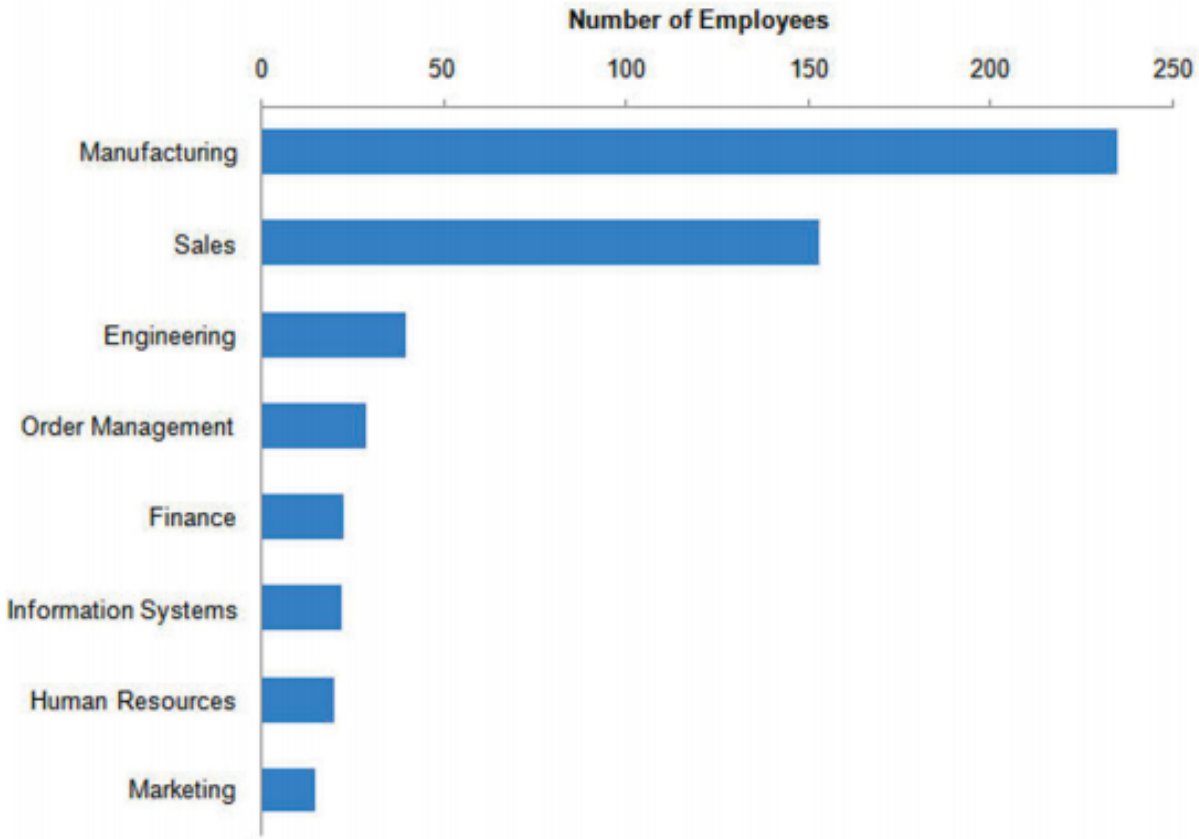
What is wrong with this graph?



Source:

`http://www.perceptualedge.com/articles/visual_business_intelligence/rules_for_using_color.pdf`

Improved



Source:

`http://www.perceptualedge.com/articles/visual_business_intelligence/rules_for_using_color.pdf`

Today's Data Download

```
# Load package RCurl
library(RCurl)

# Create URL object for OpenIntro data
url <- paste("https://raw.github.com/christopher",
             "gandrud/Introduction_to_Statistics_",
             "and_Data_Analysis",
             "_Yonsei/master/Lectures/Lecture7/",
             "OpenIntroData/run10.txt", sep = "")

# Download data
Run10 <- getURL(url, ssl.verifypeer = FALSE)

# Convert Run10 to data.frame
Run10 <- read.table(textConnection(Run10),
                    sep = "\t", header = TRUE)
```

Today's Variables

```
names(Run10)

## [1] "place"    "time"     "pace"     "age"
## [5] "gender"   "location" "state"    "divPlace"
## [9] "divTot"
```

Today we will use these variables (from Diaz et al. (2011, Ch. 4)):

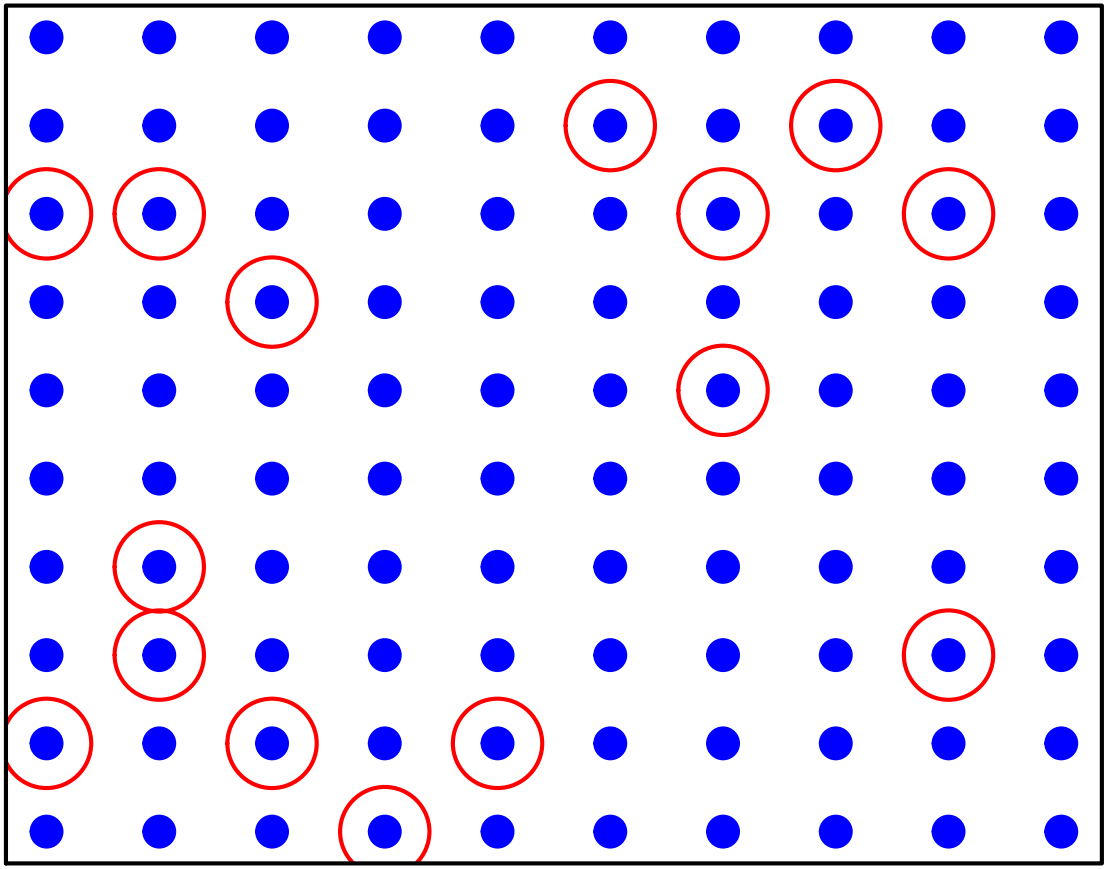
Variable	Description
`time`	ten mile run time, minutes
age	age of runner in years

Random Sample

Our data contains the **entire population** of 16,974 runners who finished the 2009 Cherry Blossom 10 Mile Run in Washington, DC.

But imagine, however, that we only have a **simple random sample** of 100 runners.

Illustrate Random Sample

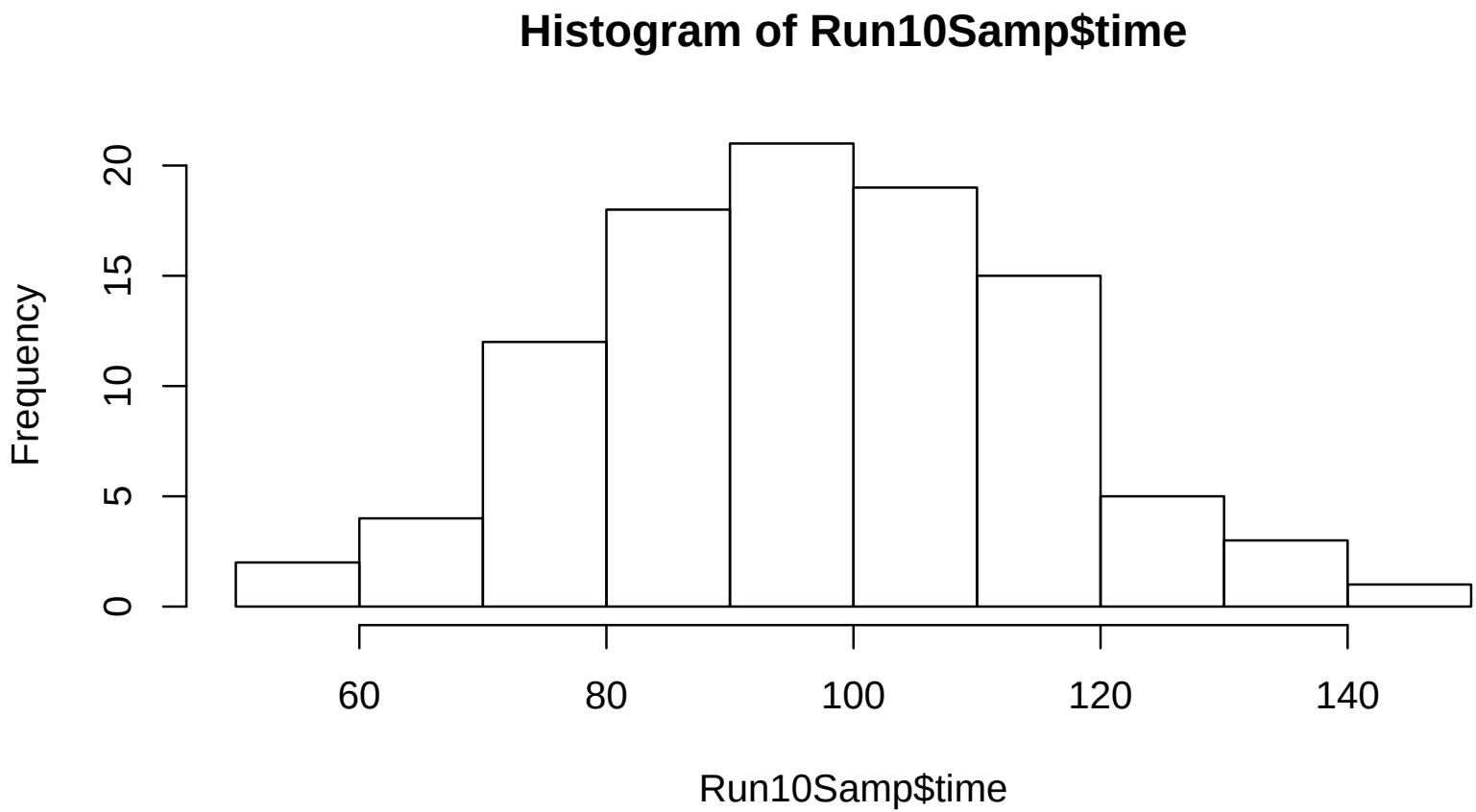


Take random sample

```
# Find number of runners in population
nrow(Run10)

## [1] 16924


# Take a random sample of 100 runners
Run10Samp <- Run10[sample(1:nrow(Run10), 100,
                          replace=FALSE),]

# Find number of runners in sample
nrow(Run10Samp)

## [1] 100
```

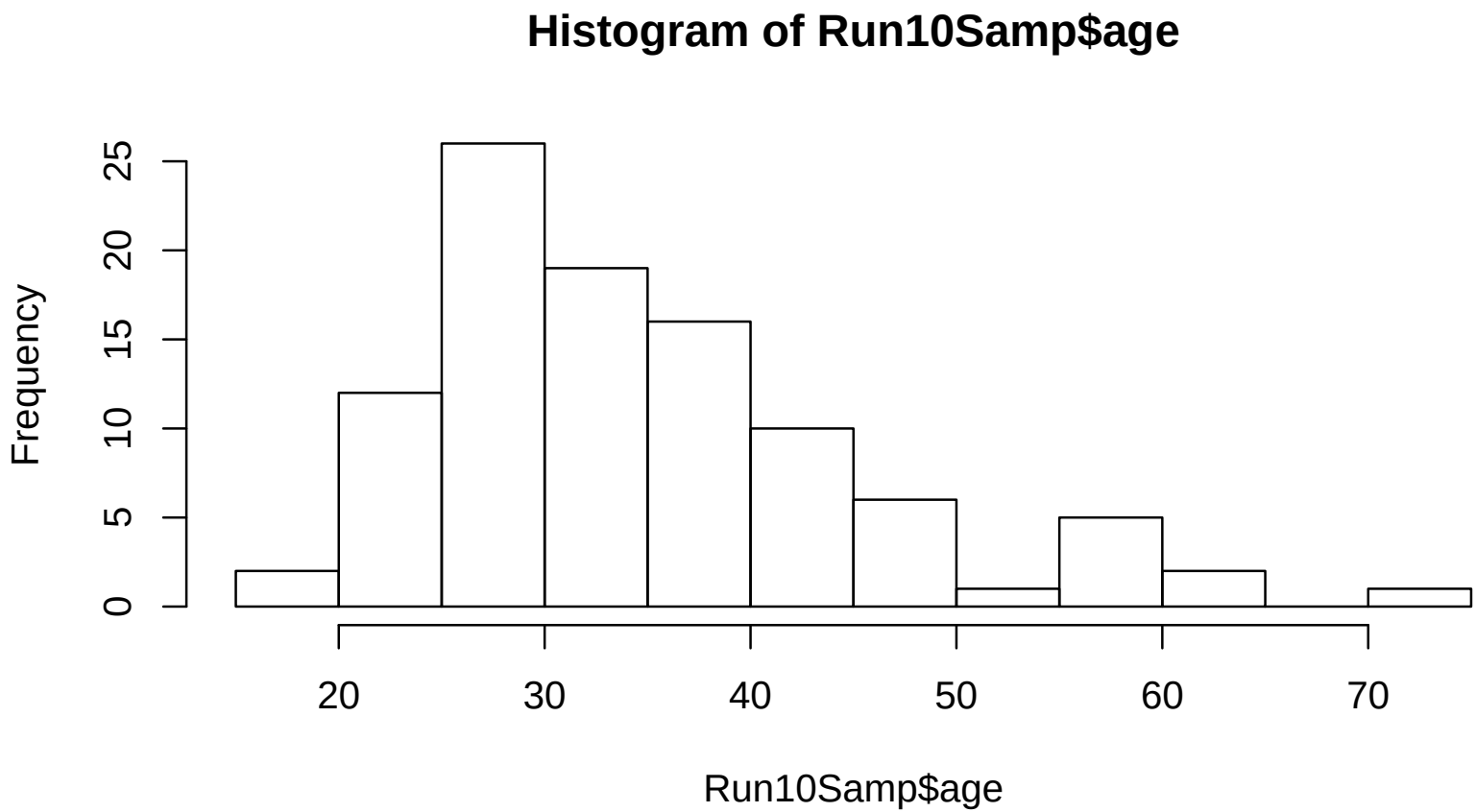
Sample Distribution of `time`

```
hist(Run10Samp$time)
```



Sample Distribution of age

```
hist(Run10Samp$age)
```



The Question

How similar is the sample to the population?

In other words, how **representative** is the sample?

In other other words, what can we **infer** about the population from the sample?

More specifically: Means

Today we are particularly interested in how similar the **sample means** $\bar{x}$ are to the **true population means** μ **for:**

- average `time` it took to complete the race,
- average age of the runners.

More specifically: Means

Today we are particularly interested in how similar the **sample means** $\bar{x}$ are to the **true population means** μ **for:**

- average `time` it took to complete the race,
- average age of the runners.

More specifically: Means

Today we are particularly interested in how similar the **sample means** $\bar{x}$ are to the **true population means** μ **for:**

- average `time` it took to complete the race,
- average age of the runners.

To answer these questions we will need tools of **statistical inference**.

Today we will learn the basics of statistical inference.

We will use these tools throughout the rest of the course and statistics.

Statistical inference is key for linking our (sampled) data to what we care about: questions about populations.

The Sample Means

Lets find our sample means.

Time:

```
mean(Run10Samp$time)

## [1] 97.01
```

Age:

```
mean(Run10Samp$age)

## [1] 35.53
```

Point Estimates (1)

These are **point estimates** of a **population parameter**, the population mean.

Point Estimates (2)

We can find point estimates of other parameters.

For example, lets find point estimates for the sample **standard deviation of the mean** (s). (The population parameter of the standard deviation is usually written σ).

Time:

```
sd(Run10Samp$time)

## [1] 17.91
```

Age:

```
sd(Run10Samp$age)

## [1] 10.67
```

Summary of our Random Sample

Variable	$\bar{x}$	μ	s	σ
`time`	97	?	17.9	?
`age`	35.5	?	10.7	?

But...

But, **how do we know** how **close** the point estimates are to the real population parameters?

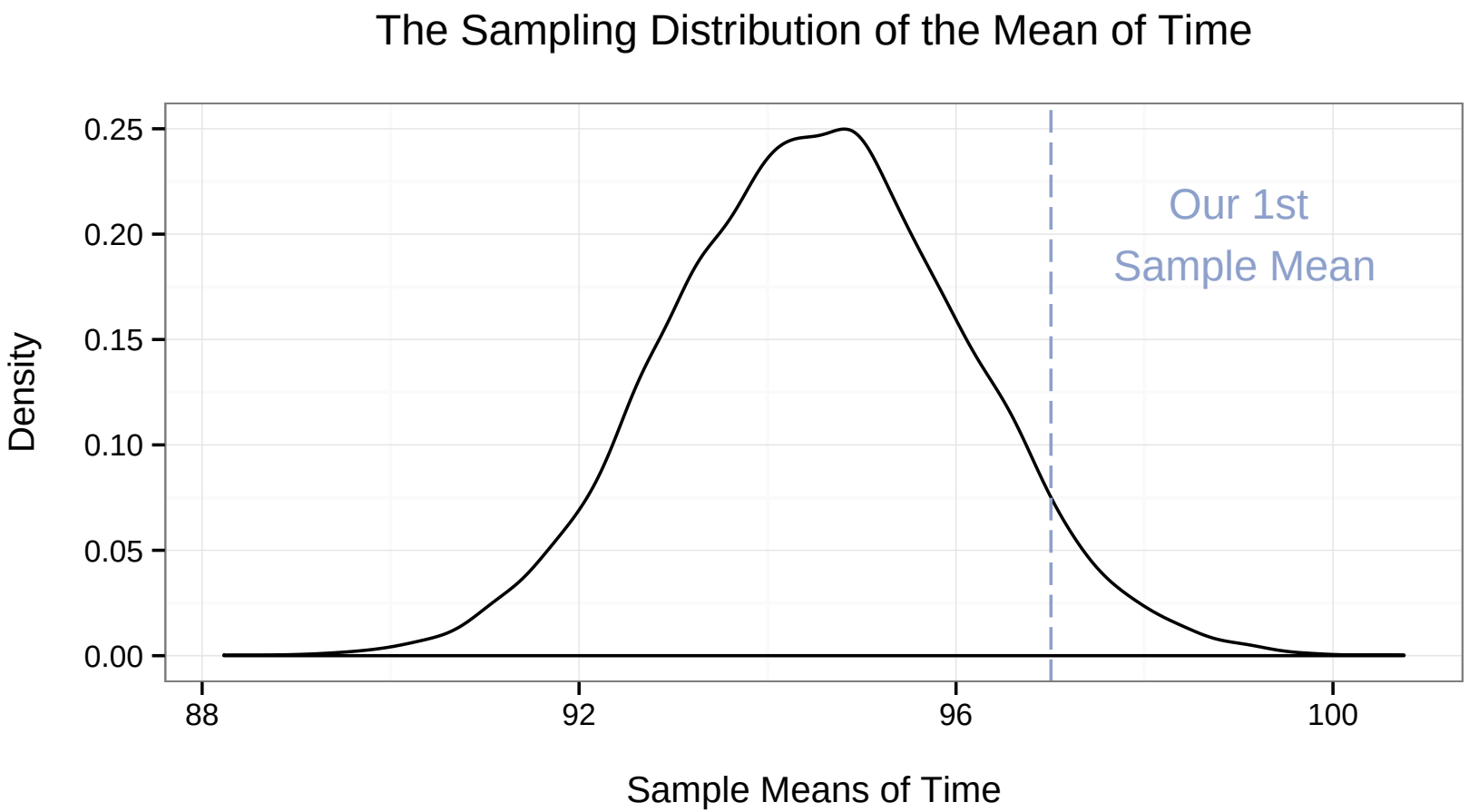
Especially since taking **other samples** we will create **different estimates**.

10000 Sample Means!

Lets find the sample mean of `time` ($\bar{x}_{time}$) in 10,000 samples!

```
# Find the sample means from 10,000 samples
TimeSampMeans <- rep(0, 10000)
  for (i in 1:10000) {
    Samp <- sample(Run10$time, 100)
    TimeSampMeans[i] <- mean(Samp)
    }

# Create a data frame object
TimeSampMeans <- data.frame(TimeSampMeans)
```

Distribution of the 10,000 Sample Means!

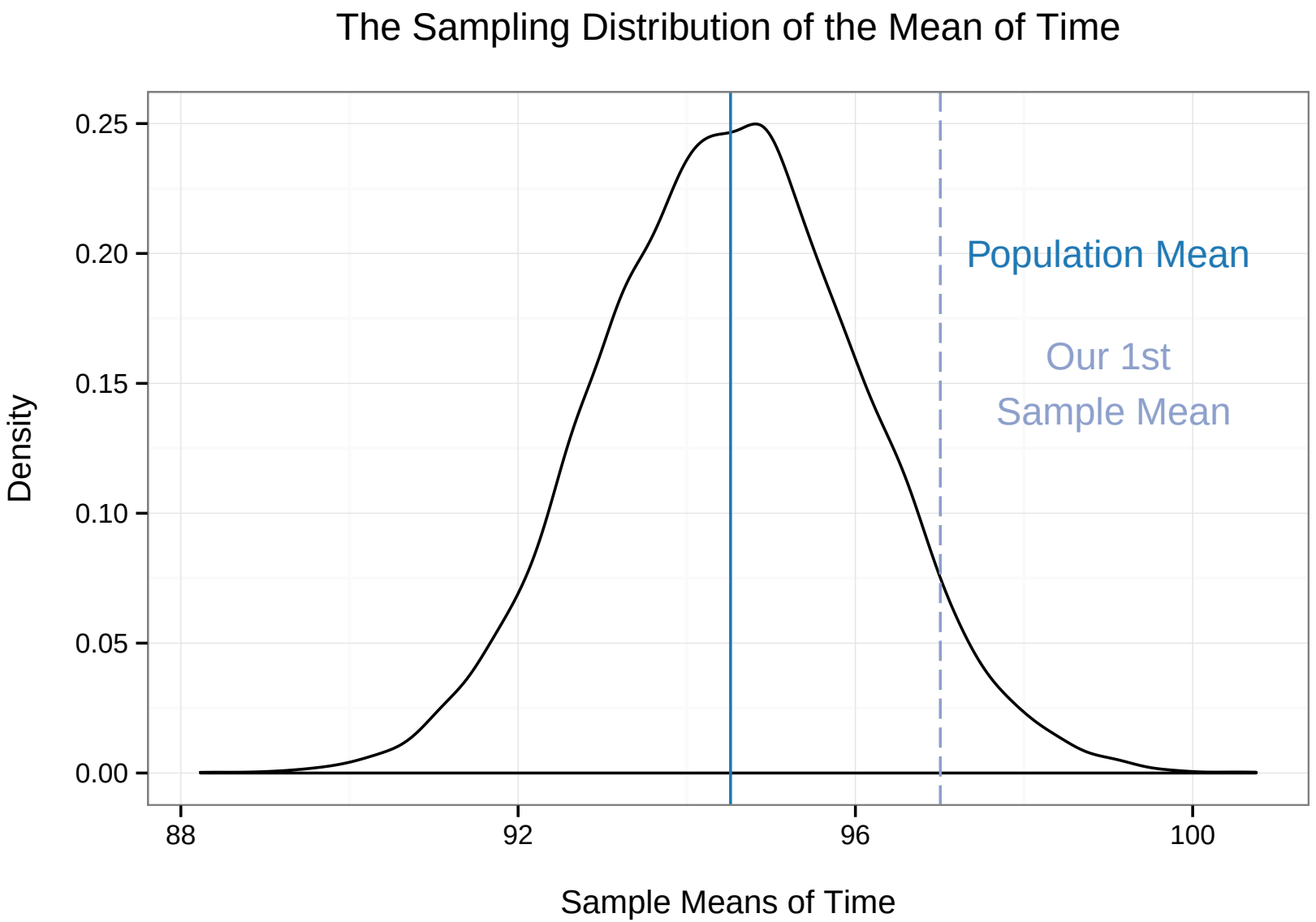


Sampling distribution and the population mean

The sampling distribution is **centered** on the population mean!

Stated much more confusingly: the mean of the sampling distribution of the sample means is the population mean.

Population Mean vs. Sample Means



Not solved yet.

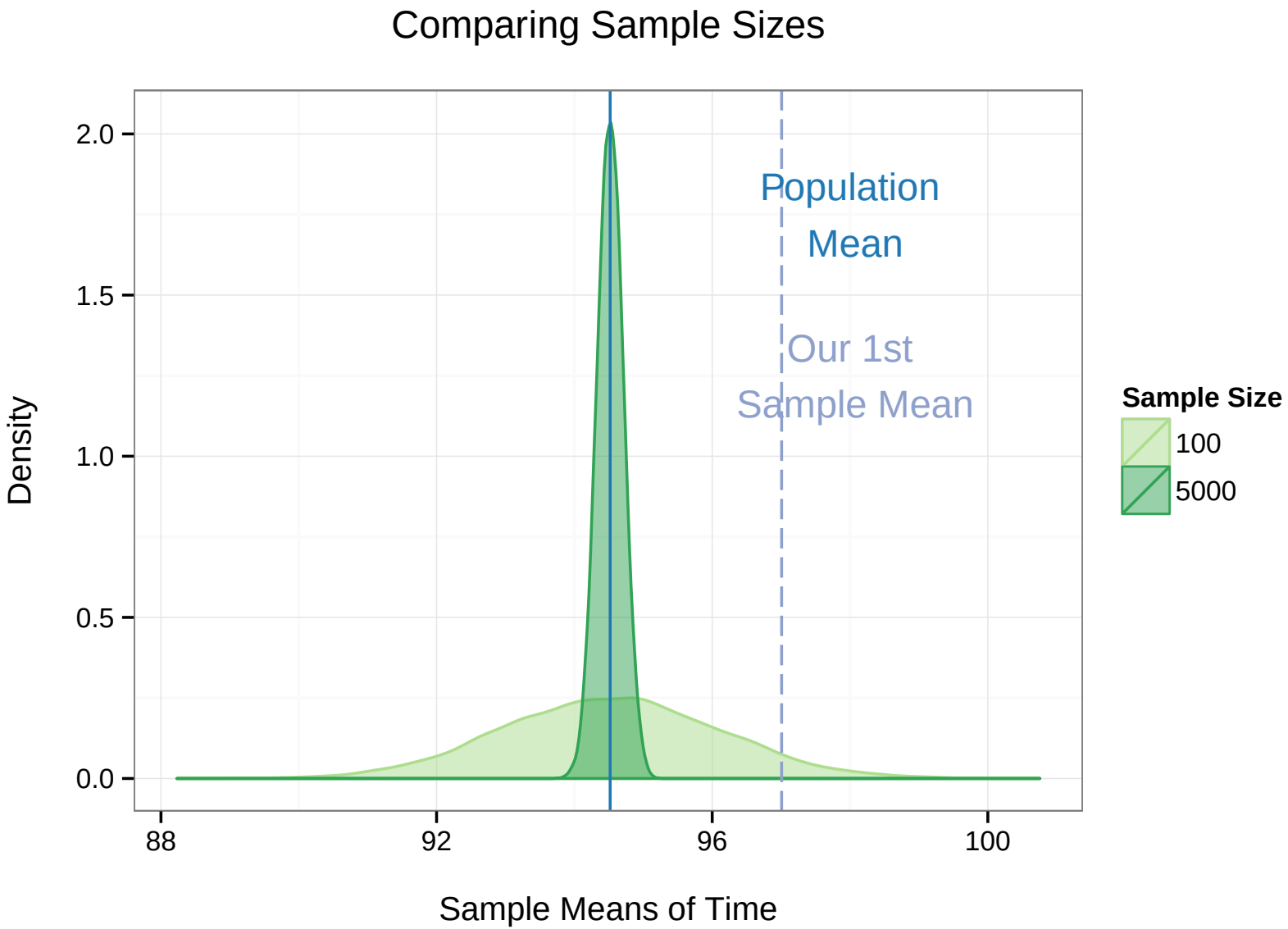
Ok, that is nice in theory, but what if we don't know the sampling distribution?

In real life we can't take 10,000 samples!

Question...

Question: What would the sampling distribution of the sample mean look like if the sample size was 5,000 rather than 100?

Comparing Sampling Distributions of $\bar{x}_{time}$ 100 vs. 5,000



The **likely distance** between the sample mean and the population mean is related to the **sample size** (n),

The **bigger** the sample size, the more likely it is that the point estimate is **close** to the population parameter.

Also, a **smaller sample standard deviation** (s) indicates less variability $\rightarrow$ more likely that the point estimate is **close** to the population parameter.

So...

Finally, the Standard Error!

Standard Error (SE):

$$SE = \frac{s}{\sqrt{n}}$$

Assuming that the observations are *independent*, i.e. the value of one observation does not depend on the value of another observation.

SE and Uncertainty

The **standard error** is a **measure** of our **uncertainty** about how well the point estimate represents the population parameter.

The Standard Error of the Mean of `time`

```
# Load plotrix
library(plotrix)

# Calculate standard error of the mean of time
std.error(Run10Samp$time)

## [1] 1.791
```

`Time` in our original sample with 100 observations had a sampling mean of 97 and a standard error of 1.8.

Um. . .

Um . . . a sampling mean of 97 and a standard error of 1.8?

That's not very easy to understand.

We need something more useful.

We need something that is more **intuitive** to interpret.

Maybe a **range** of likely population parameter values.

The Confidence Interval!

About **95%** of the time the population parameter will be within **about 2 standard errors** of the point estimate.

95% Confidence Interval (CI):

$$CI = \text{point estimate} \pm 2 * SE$$

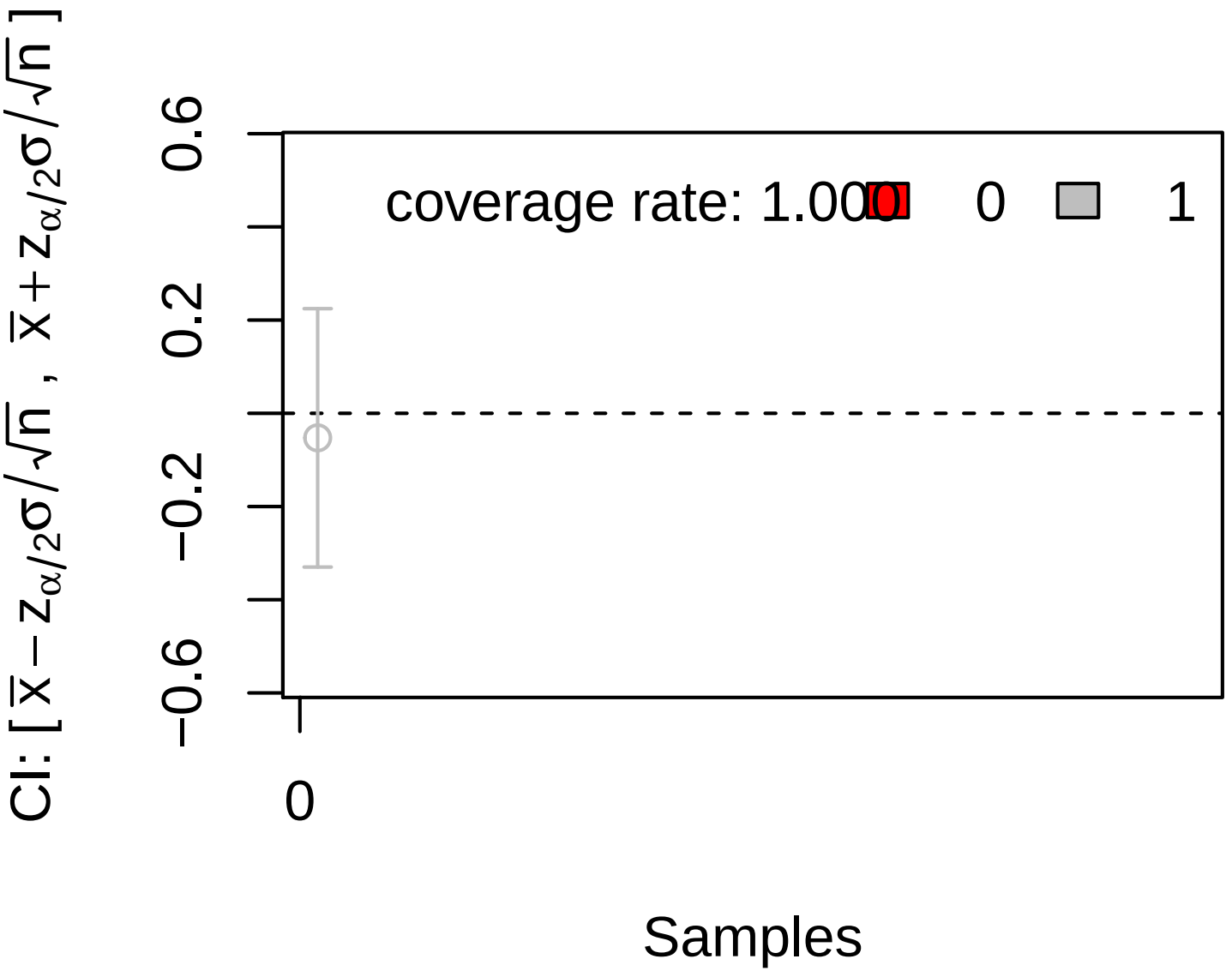
Central Limit Theorem (informal definition):

If a sample has at least 50 independent observations, and the data is not extremely skewed, the distribution of the sampling mean closely approximates the normal distribution (Diaz et al., 2011, 151))

So what? This lets our confidence interval be more **precise**.

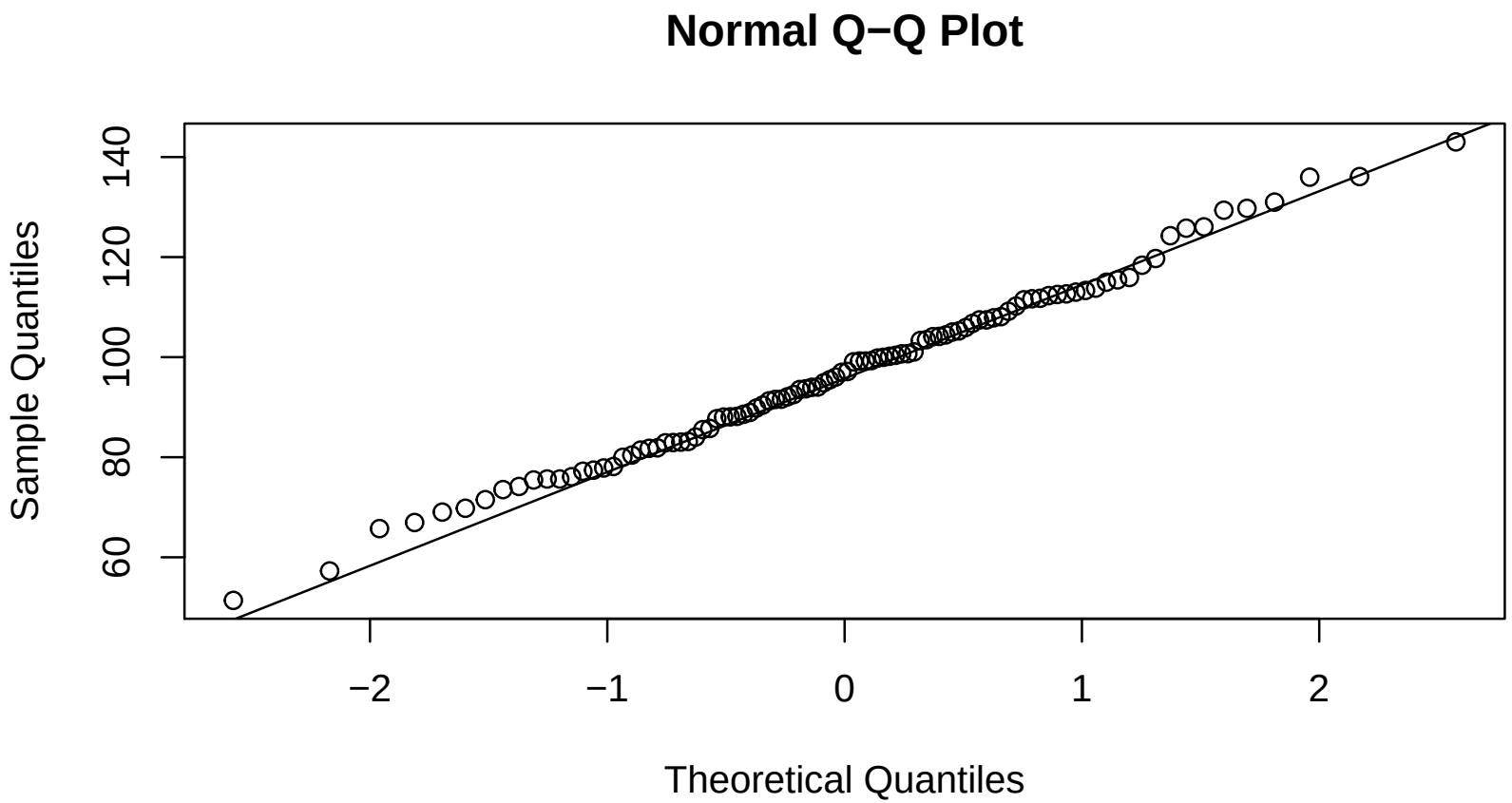
$$CI = \text{point estimate} \pm 1.96 * SE$$

Confidence Interval Simulation



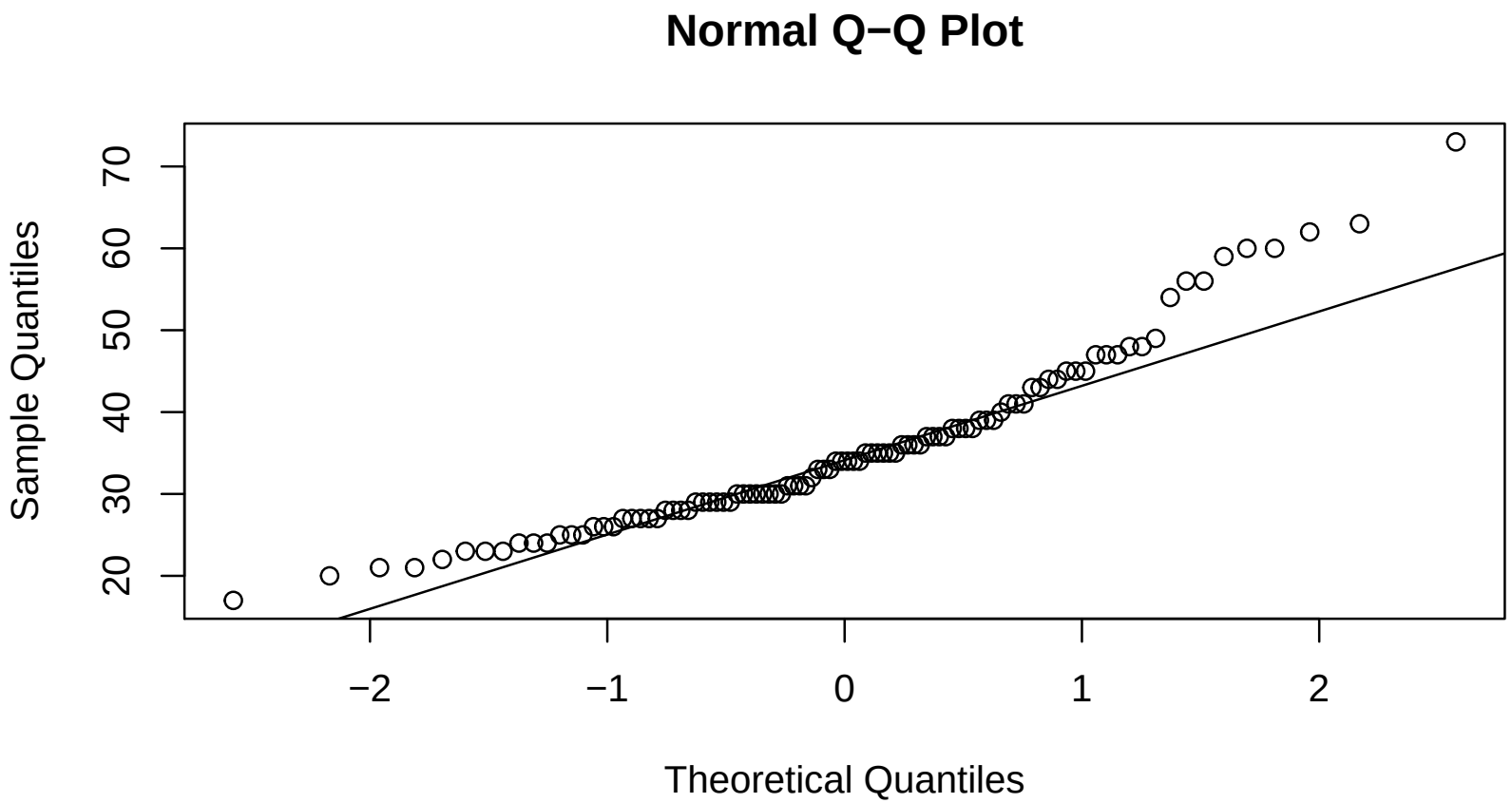
Visually Confirm Normally Distributed Time Data

```
# Create Quantile-Quantile Plots
qqnorm(Run10Samp$time); qqline(Run10Samp$time)
```



Visually Confirm Normally Distributed Age Data

```
# Create Quantile-Quantile Plots
qqnorm(Run10Samp$age); qqline(Run10Samp$age)
```



Independence

In general, observations from a simple random sample of **less than 10% of the population** will be independent.

Let's calculate some 95% confidence intervals for `time`

```
# Find Mean of time
MeanTime <- mean(Run10Samp$time)

# Find Standard Error of the mean of time
SETime <- std.error(Run10Samp$time)

# Lower CI
LowerTime <- MeanTime - 1.96 * SETime

# Upper CI
UpperTime <- MeanTime + 1.96 * SETime
```

```
# Show Confidence Intervals
LowerTime
```

```
## [1] 93.5
```

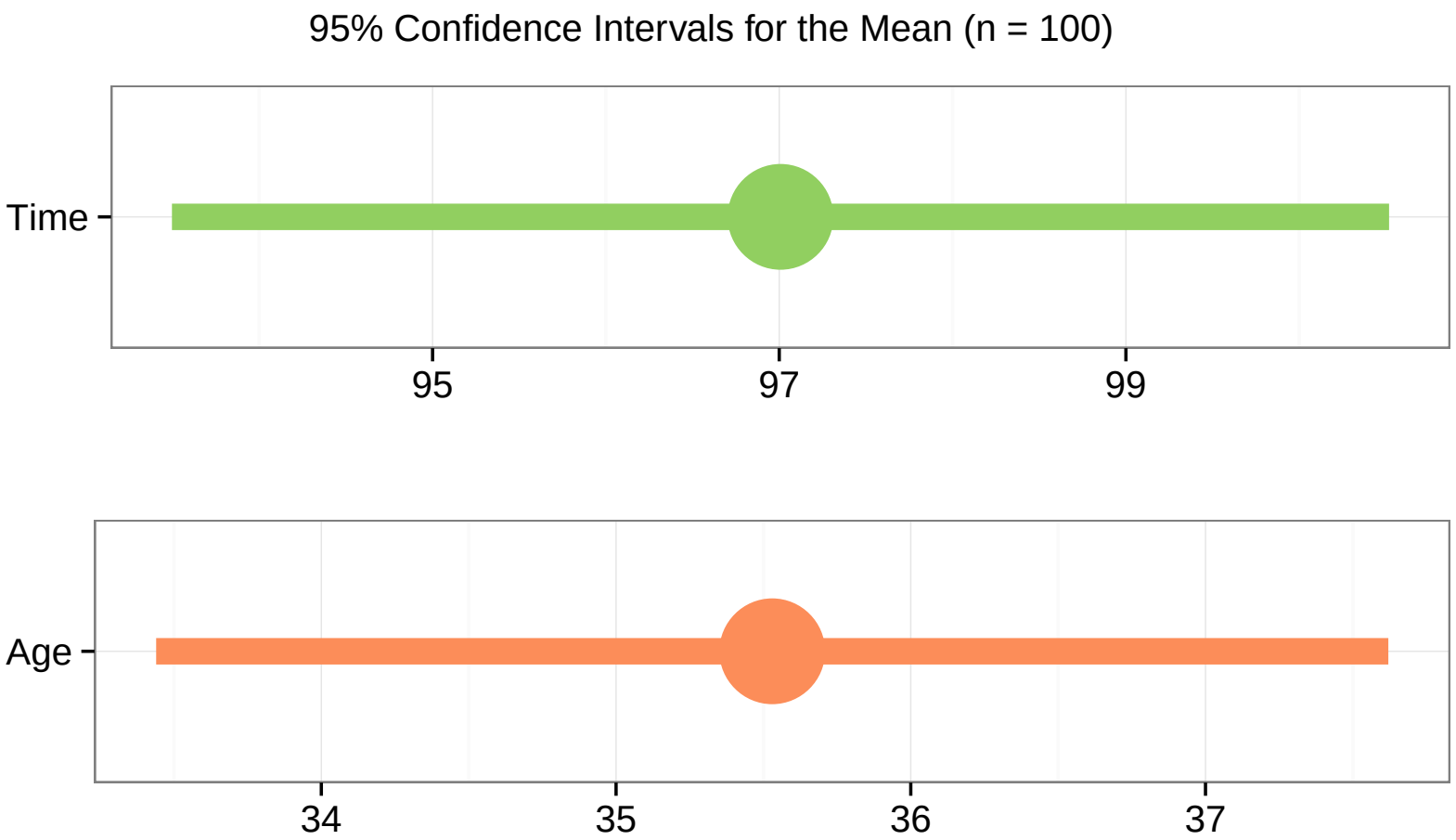
```
UpperTime
```

```
## [1] 100.5
```

Summary of our Parameter Estimates

Variable	Lower CI	Mean	Upper CI	Population Mean
`time`	93.5	97	100.5	94.5
`age`	33.4	35.5	37.6	35.5

Alternative



References I

Diaz, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel. 2011. OpenIntro Statistics. 1st ed. http://www.openintro.org/stat/downloads.php.