AUA CS 108, Statistics, Fall 2019

Lecture 01

Michael Poghosyan
YSU, AUA
michael@ysu.am, mpoghosyan@aua.am

26 Aug 2019

Welcome to the AUA Statistics Course

Welcome to the AUA Statistics Course

And Happy New ~~Year~~ Semester ! ◡̈

Contents

- Syllabus highlights
- Intro to the Course
- Intro to the Descriptive Statistics

Syllabus Highlights

- Course name: **CS 108, Statistics, Section B**

Syllabus Highlights

- Course name: **CS 108, Statistics, Section B**
- No. of Credits: **3**

Syllabus Highlights

- Course name: **CS 108, Statistics, Section B**
- No. of Credits: **3**
- Instructor: **MP**

Syllabus Highlights

- Course name: **CS 108, Statistics, Section B**
- No. of Credits: **3**
- Instructor: **MP**
- Instructor's Office: **#336W, PAB**

Syllabus Highlights

- Course name: **CS 108, Statistics, Section B**
- No. of Credits: **3**
- Instructor: **MP**
- Instructor's Office: **#336W, PAB**
- Instructor's OH: Mon? Wed?

Syllabus Highlights

- Course name: **CS 108, Statistics, Section B**
- No. of Credits: **3**
- Instructor: **MP**
- Instructor's Office: **#336W, PAB**
- Instructor's OH: Mon? Wed?
- Teaching Associate: **Mane Margaryan**

Syllabus Highlights

- Course name: **CS 108, Statistics, Section B**
- No. of Credits: **3**
- Instructor: **MP**
- Instructor's Office: **#336W, PAB**
- Instructor's OH: Mon? Wed?
- Teaching Associate: **Mane Margaryan**
- PSS day/time: **TBD**

Syllabus Highlights

- Course name: **CS 108, Statistics, Section B**
- No. of Credits: **3**
- Instructor: **MP**
- Instructor's Office: **#336W, PAB**
- Instructor's OH: Mon? Wed?
- Teaching Associate: **Mane Margaryan**
- PSS day/time: **TBD**
- TA's OH: **TBD**

Supplementary Info

- Section A Instructor: **Ruben Gevorgyan**

Supplementary Info

- Section A Instructor: **Ruben Gevorgyan**
- Section A TA: **Lusine Zilfimyan**

Course Materials

- Moodle Page: shared with the Section A

Course Materials

- Moodle Page: shared with the Section A
- Moodle Enrollment Key: **TraLaLa**

Course Materials

- Moodle Page: shared with the Section A
- Moodle Enrollment Key: **TraLaLa**
- Syllabus: uploaded to our Moodle page

Course Materials

- Moodle Page: shared with the Section A
- Moodle Enrollment Key: **TraLaLa**
- Syllabus: uploaded to our Moodle page
- Textbooks: uploaded to our Moodle page

Course Materials

- Moodle Page: shared with the Section A
- Moodle Enrollment Key: **TraLaLa**
- Syllabus: uploaded to our Moodle page
- Textbooks: uploaded to our Moodle page
- Software: **R** and **R Studio** (freeware)

Course Materials

- Moodle Page: shared with the Section A
- Moodle Enrollment Key: **TraLaLa**
- Syllabus: uploaded to our Moodle page
- Textbooks: uploaded to our Moodle page
- Software: **R** and **R Studio** (freeware)
- R Textbooks: uploaded to our Moodle Page

Course Materials

- Moodle Page: shared with the Section A
- Moodle Enrollment Key: **TraLaLa**
- Syllabus: uploaded to our Moodle page
- Textbooks: uploaded to our Moodle page
- Software: **R** and **R Studio** (freeware)
- R Textbooks: uploaded to our Moodle Page
- Other: I have prepared some R Intro Slides

Course Materials

- Moodle Page: shared with the Section A
- Moodle Enrollment Key: **TraLaLa**
- Syllabus: uploaded to our Moodle page
- Textbooks: uploaded to our Moodle page
- Software: **R** and **R Studio** (freeware)
- R Textbooks: uploaded to our Moodle Page
- Other: I have prepared some R Intro Slides
- Question: Do we need some supplementary R Labs? YESSS/No

Syllabus Highlights, Cont'd

- Exams: **2 Midterms** and a **Final Exam**

Syllabus Highlights, Cont'd

- Exams: **2 Midterms** and a **Final Exam**
- Homework: **(almost) weekly**, due on Fridays

Syllabus Highlights, Cont'd

- Exams: **2 Midterms** and a **Final Exam**
- Homework: **(almost) weekly**, due on Fridays
- Quizzes: Yeah, we will have them!

Syllabus Highlights, Cont'd

- Exams: **2 Midterms** and a **Final Exam**
- Homework: **(almost) weekly**, due on Fridays
- Quizzes: Yeah, we will have them!
- Final Grade Formula:

$$Total = 0.1 \cdot (HW + Q) + 0.2 \cdot (M1 + M2) + 0.4 \cdot F$$

Syllabus Highlights, Cont'd

- Exams: **2 Midterms** and a **Final Exam**
- Homework: **(almost) weekly**, due on Fridays
- Quizzes: Yeah, we will have them!
- Final Grade Formula:

$$Total = 0.1 \cdot (HW + Q) + 0.2 \cdot (M1 + M2) + 0.4 \cdot F$$

- No Makeups for Quizzes! Sorry!

Syllabus Highlights, Cont'd

- Exams: **2 Midterms** and a **Final Exam**
- Homework: **(almost) weekly**, due on Fridays
- Quizzes: Yeah, we will have them!
- Final Grade Formula:

$$Total = 0.1 \cdot (HW + Q) + 0.2 \cdot (M1 + M2) + 0.4 \cdot F$$

- No Makeups for Quizzes! Sorry!
- No late HWs (except some veeery special cases)

Syllabus Highlights, Cont'd

- Exams: **2 Midterms** and a **Final Exam**
- Homework: **(almost) weekly**, due on Fridays
- Quizzes: Yeah, we will have them!
- Final Grade Formula:

$$Total = 0.1 \cdot (HW + Q) + 0.2 \cdot (M1 + M2) + 0.4 \cdot F$$

- No Makeups for Quizzes! Sorry!
- No late HWs (except some veeery special cases)
- No Grades Curving. ☹

Syllabus Highlights, Cont'd

- Exams: **2 Midterms** and a **Final Exam**
- Homework: **(almost) weekly**, due on Fridays
- Quizzes: Yeah, we will have them!
- Final Grade Formula:

$$Total = 0.1 \cdot (HW + Q) + 0.2 \cdot (M1 + M2) + 0.4 \cdot F$$

- No Makeups for Quizzes! Sorry!
- No late HWs (except some veeery special cases)
- No Grades Curving. ☹
- Advice: Always ask your questions, attend OHs, solve HWs by yourself!

Syllabus Highlights, Cont'd

- Exams: **2 Midterms** and a **Final Exam**
- Homework: **(almost) weekly**, due on Fridays
- Quizzes: Yeah, we will have them!
- Final Grade Formula:

$$Total = 0.1 \cdot (HW + Q) + 0.2 \cdot (M1 + M2) + 0.4 \cdot F$$

- No Makeups for Quizzes! Sorry!
- No late HWs (except some veeery special cases)
- No Grades Curving. ☹
- Advice: Always ask your questions, attend OHs, solve HWs by yourself!
- Advice: Run over the Probability Topics, especially, about RVs and Distributions

QA Session

Questions?

About Statistics

- What is Statistics?

About Statistics

- What is Statistics?

Statistics is an Art and Science of Learning from Data.

About Statistics

- What is Statistics?

Statistics is an Art and Science of Learning from Data.

- What is the difference between Statistical and Mathematical thinking?

About Statistics

- What is Statistics?

Statistics is an Art and Science of Learning from Data.

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms).

About Statistics

- What is Statistics?

Statistics is an Art and Science of Learning from Data.

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms). Say, our lovely MVT is correct in all cases!

About Statistics

- What is Statistics?

Statistics is an Art and Science of Learning from Data.

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms). Say, our lovely MVT is correct in all cases! The real roots of $x^2 = 4$ are

About Statistics

- What is Statistics?

Statistics is an Art and Science of Learning from Data.

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms). Say, our lovely MVT is correct in all cases! The real roots of $x^2 = 4$ are $x = \pm 2$.

About Statistics

- What is Statistics?

Statistics is an Art and Science of Learning from Data.

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms). Say, our lovely MVT is correct in all cases! The real roots of $x^2 = 4$ are $x = \pm 2$.

In Statistics (Inferential Statistics), we use data to get an insight about the unknown process behind the generation of the data. Usually, our data is finite, and this is not giving a chance to get a complete and 100 percent information about that unknown process.

About Statistics

- What is Statistics?

Statistics is an Art and Science of Learning from Data.

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms). Say, our lovely MVT is correct in all cases! The real roots of $x^2 = 4$ are $x = \pm 2$.

In Statistics (Inferential Statistics), we use data to get an insight about the unknown process behind the generation of the data. Usually, our data is finite, and this is not giving a chance to get a complete and 100 percent information about that unknown process. So, in Statsitsics, we are "inferring", "guessing", ***estimating*** the unknown process, its parameters, and sometimes give the level of our confidence.

About Statistics

- What is Statistics?

Statistics is an Art and Science of Learning from Data.

- What is the difference between Statistical and Mathematical thinking?

In Mathematics, we prove facts (using other, already proven facts, and using some small number of Axioms). Say, our lovely MVT is correct in all cases! The real roots of $x^2 = 4$ are $x = \pm 2$.

In Statistics (Inferential Statistics), we use data to get an insight about the unknown process behind the generation of the data. Usually, our data is finite, and this is not giving a chance to get a complete and 100 percent information about that unknown process. So, in Statsitsics, we are "inferring", "guessing", ***estimating*** the unknown process, its parameters, and sometimes give the level of our confidence. Say, if I am tossing a coin 10 times and get 8 Hs and 2 Ts, is it a sign that the coin is not fair?

About Statistics

- What is the difference between Data Scienece and Statistics?

About Statistics

- What is the difference between Data Scienece and Statistics?

Well, maybe we will talk a little bit about this at the end of the course.

- Why I need to learn Statistics?

About Statistics

- What is the difference between Data Scienece and Statistics?

Well, maybe we will talk a little bit about this at the end of the course.

- Why I need to learn Statistics?

Simple - to pass this course ¨‿

About Statistics

- What is the difference between Data Scienece and Statistics?

Well, maybe we will talk a little bit about this at the end of the course.

- Why I need to learn Statistics?

Simple - to pass this course ⌣̈ Or, it is very important if you want to become a Statistician (one of the fastest growing and Best Jobs in USA), learn DS/ML, become a Biostatistician, learn Econometrics, make Investments and play in Financial Markets, and many-many more, see, for example, the Wiki page.

About Statistics

- What is the difference between Data Scienece and Statistics?

Well, maybe we will talk a little bit about this at the end of the course.

- Why I need to learn Statistics?

Simple - to pass this course ⌣̈ Or, it is very important if you want to become a Statistician (one of the fastest growing and Best Jobs in USA), learn DS/ML, become a Biostatistician, learn Econometrics, make Investments and play in Financial Markets, and many-many more, see, for example, the Wiki page.

And finally, to understand the everyday usage of Statistical language, graphs and estimates, say, about polls and salaries ⌣̈

Course Structure: Topics at a glance

The structure of our course will be the following:

- Intro + Descriptive Statistics

Course Structure: Topics at a glance

The structure of our course will be the following:

- Intro + Descriptive Statistics
- Quick reminder on RVs, Convergence Types for RVs, and our good old LLN and CLT:

The rest will use these topics intensively.

Course Structure: Topics at a glance

The structure of our course will be the following:

- Intro + Descriptive Statistics
- Quick reminder on RVs, Convergence Types for RVs, and our good old LLN and CLT:

The rest will use these topics intensively.

- Models, Statistical Inference and Learning:

Here we will talk about Parametric and Non-Parametric Statistics, and the main problems of the Parametric Statistics: Parameter Estimation, Confidence Intervals and Hypothesis Testing

Course Structure: Topics at a glance

Then we will run over these three problems:

- Parameter Point Estimates

Course Structure: Topics at a glance

Then we will run over these three problems:

- Parameter Point Estimates
- Confidence Intervals

Course Structure: Topics at a glance

Then we will run over these three problems:

- Parameter Point Estimates
- Confidence Intervals
- Hypothesis Testing

Course Structure: Topics at a glance

Then we will run over these three problems:

- Parameter Point Estimates
- Confidence Intervals
- Hypothesis Testing

Then we will focus on the simplest Statistical Model for the relationship between different Variables: we will learn

- Linear Regression

Course Structure: Topics at a glance

Then we will run over these three problems:

- Parameter Point Estimates
- Confidence Intervals
- Hypothesis Testing

Then we will focus on the simplest Statistical Model for the relationship between different Variables: we will learn

- Linear Regression

And at the end of the course we will return back to Testing and cover:

- Goodness of fit tests

Descriptive Statistics

Descriptive Statistics

Descriptive Statistics is to get the first, basic information about the Data, either in the Visual or Numerical form.

Descriptive Statistics

Descriptive Statistics is to get the first, basic information about the Data, either in the Visual or Numerical form.

Consider, for example, the dataset `mpg` from the `ggplot2` package.

Descriptive Statistics

Descriptive Statistics is to get the first, basic information about the Data, either in the Visual or Numerical form.

Consider, for example, the dataset `mpg` from the `ggplot2` package. From the official description of the data,

> *This dataset contains a subset of the fuel economy data that the EPA makes available on http://fueleconomy.gov. It contains only models which had a new release every year between 1999 and 2008.*

Descriptive Statistics

Descriptive Statistics is to get the first, basic information about the Data, either in the Visual or Numerical form.

Consider, for example, the dataset `mpg` from the `ggplot2` package. From the official description of the data,

> *This dataset contains a subset of the fuel economy data that the EPA makes available on http://fueleconomy.gov. It contains only models which had a new release every year between 1999 and 2008.*

Lets look at the first 3 rows of our dataset:

```
head(ggplot2::mpg, 3)
```

```
## # A tibble: 3 x 11
##   manufacturer model displ  year   cyl trans      drv      cty
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18
## 2 audi         a4      1.8  1999     4 manual(m5) f        21
## 3 audi         a4      2    2008     4 manual(m6) f        20
```

Descriptive Statistics

Descriptive Statistics

The variable `cty` is the *city miles per gallon*, and the variable `cyl` is the *number of cylinders*. Let's separate that Variables:

```
cty <- ggplot2::mpg$cty
cyl <- ggplot2::mpg$cyl
```

Descriptive Statistics

Let's see the results:

```
cyl
```

```
##   [1] 4 4 4 4 6 6 6 4 4 4 4 6 6 6 6 6 8 8 8 8 8 8 8 8 8 8
##  [38] 4 6 6 6 6 6 6 6 6 6 6 6 6 6 6 8 8 8 8 8 6 8 8 8 8 8 8 8
##  [75] 8 8 8 6 6 6 6 8 8 6 6 8 8 8 8 8 6 6 6 6 8 8 8 8 8 4 4 4
## [112] 4 6 6 6 4 4 4 4 6 6 6 6 6 6 8 8 8 8 8 8 8 8 8 8 8 6 6
## [149] 6 6 6 6 6 8 6 6 6 6 8 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6
## [186] 6 4 4 4 4 6 6 6 4 4 4 4 4 8 8 4 4 4 6 6 6 6 4 4 4 4 6 4
## [223] 4 4 4 5 5 4 4 4 4 6 6 6
```

Descriptive Statistics

Let's see the results:

```
cyl
```

```
##   [1] 4 4 4 4 6 6 6 4 4 4 4 6 6 6 6 6 8 8 8 8 8 8 8 8 8 8
##  [38] 4 6 6 6 6 6 6 6 6 6 6 6 6 6 6 8 8 8 8 8 6 8 8 8 8 8 8 8
##  [75] 8 8 8 6 6 6 6 8 8 6 6 8 8 8 8 8 6 6 6 6 8 8 8 8 8 4 4 4
## [112] 4 6 6 6 4 4 4 4 6 6 6 6 6 6 8 8 8 8 8 8 8 8 8 8 8 6 6
## [149] 6 6 6 6 6 8 6 6 6 6 8 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6
## [186] 6 4 4 4 4 6 6 6 4 4 4 4 4 8 8 4 4 4 6 6 6 6 4 4 4 4 6 4
## [223] 4 4 4 5 5 4 4 4 4 6 6 6
```

Can you describe this data? What can be said about the No. of Cylinders of these cars?

Descriptive Statistics

Let's see the results for `cty`:

```
cty
```

```
##   [1] 18 21 20 21 16 18 18 18 16 20 19 15 17 17 15 15 17 16 1
##  [26] 16 15 15 14 11 11 14 19 22 18 18 17 18 17 16 16 17 17 1
##  [51] 13 14 14 14  9 11 11 13 13  9 13 11 13 11 12  9 13 13 1
##  [76] 11 12 14 15 14 13 13 13 14 14 13 13 13 11 13 18 18 17 1
## [101] 24 25 23 24 26 25 24 21 18 18 21 21 18 18 19 19 19 20 2
## [126] 14  9 14 13 11 11 12 12 11 11 11 12 14 13 13 13 21 19 2
## [151] 14 15 14 12 18 16 17 18 16 18 18 20 19 20 18 21 19 19 1
## [176] 15 15 16 14 21 21 21 21 18 18 19 21 21 21 22 18 18 18 2
## [201] 15 16 17 15 15 15 16 21 19 21 22 17 33 21 19 22 21 21 2
## [226] 20 20 21 18 19 21 16 18 17
```

Descriptive Statistics

Let's see the results for `cty`:

```
cty
```

```
##   [1] 18 21 20 21 16 18 18 18 16 20 19 15 17 17 15 15 17 16 1
##  [26] 16 15 15 14 11 11 14 19 22 18 18 17 18 17 16 16 17 17 1
##  [51] 13 14 14 14  9 11 11 13 13  9 13 11 13 11 12  9 13 13 1
##  [76] 11 12 14 15 14 13 13 13 14 14 13 13 13 11 13 18 18 17 1
## [101] 24 25 23 24 26 25 24 21 18 18 21 21 18 18 19 19 19 20 2
## [126] 14  9 14 13 11 11 12 12 11 11 11 12 14 13 13 13 21 19 2
## [151] 14 15 14 12 18 16 17 18 16 18 18 20 19 20 18 21 19 19 1
## [176] 15 15 16 14 21 21 21 21 18 18 19 21 21 21 22 18 18 18 2
## [201] 15 16 17 15 15 15 16 21 19 21 22 17 33 21 19 22 21 21 2
## [226] 20 20 21 18 19 21 16 18 17
```

Again, can you describe this data? What can be said about the City Miles per Gallon values of these cars?

Descriptive Statistics

Descriptive Statistics gives us tools to Describe the Data, get some basic, general information about the dataset.

Descriptive Statistics

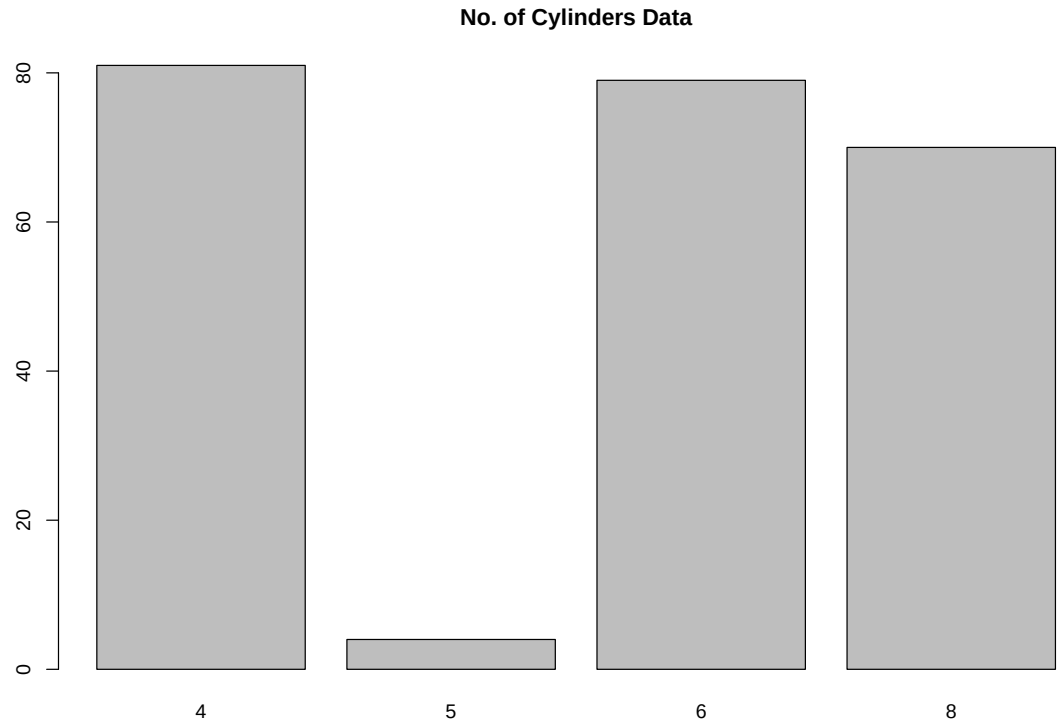
Descriptive Statistics gives us tools to Describe the Data, get some basic, general information about the dataset. We will describe data Graphically and/or by giving some numerical summaries.

Descriptive Statistics

Descriptive Statistics gives us tools to Describe the Data, get some basic, general information about the dataset. We will describe data Graphically and/or by giving some numerical summaries.

For example, let us draw the BarPlot for the frequencies of the `cyl` variable:

```
barplot(table(cyl), main = "No. of Cylinders Data")
```



Descriptive Statistics

Now, let us give some numerical summaries for `cty`: calculate the average Miles per Gallon for a City, and its max and min.

```
cat("mean = ", mean(cty))
```

```
## mean =  16.85897
```

```
cat("Max = ", max(cty))
```

```
## Max =  35
```

```
cat("Min = ", min(cty))
```

```
## Min =  9
```

Descriptive Statistics

Now, let us give some numerical summaries for `cty`: calculate the average Miles per Gallon for a City, and its max and min.

```
cat("mean = ", mean(cty))
```

```
## mean =  16.85897
```

```
cat("Max = ", max(cty))
```

```
## Max =  35
```

```
cat("Min = ", min(cty))
```

```
## Min =  9
```

And we can use the `summary` command to get some numerical info:

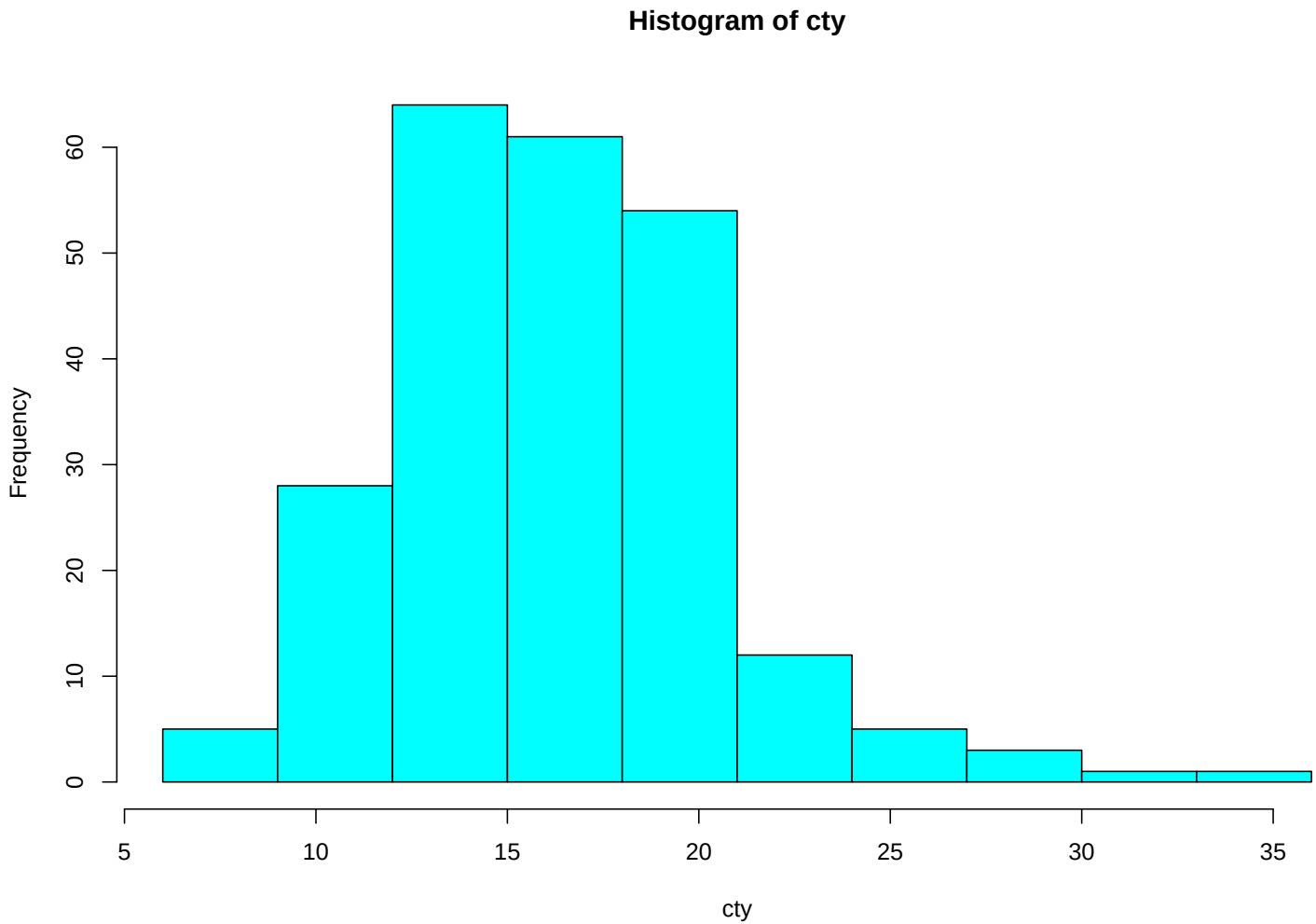
```
summary(cty)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.00   14.00   17.00   16.86   19.00   35.00
```

Descriptive Statistics

To get some visual information about the Variable `cty`, its distribution, we can draw the Histogram:

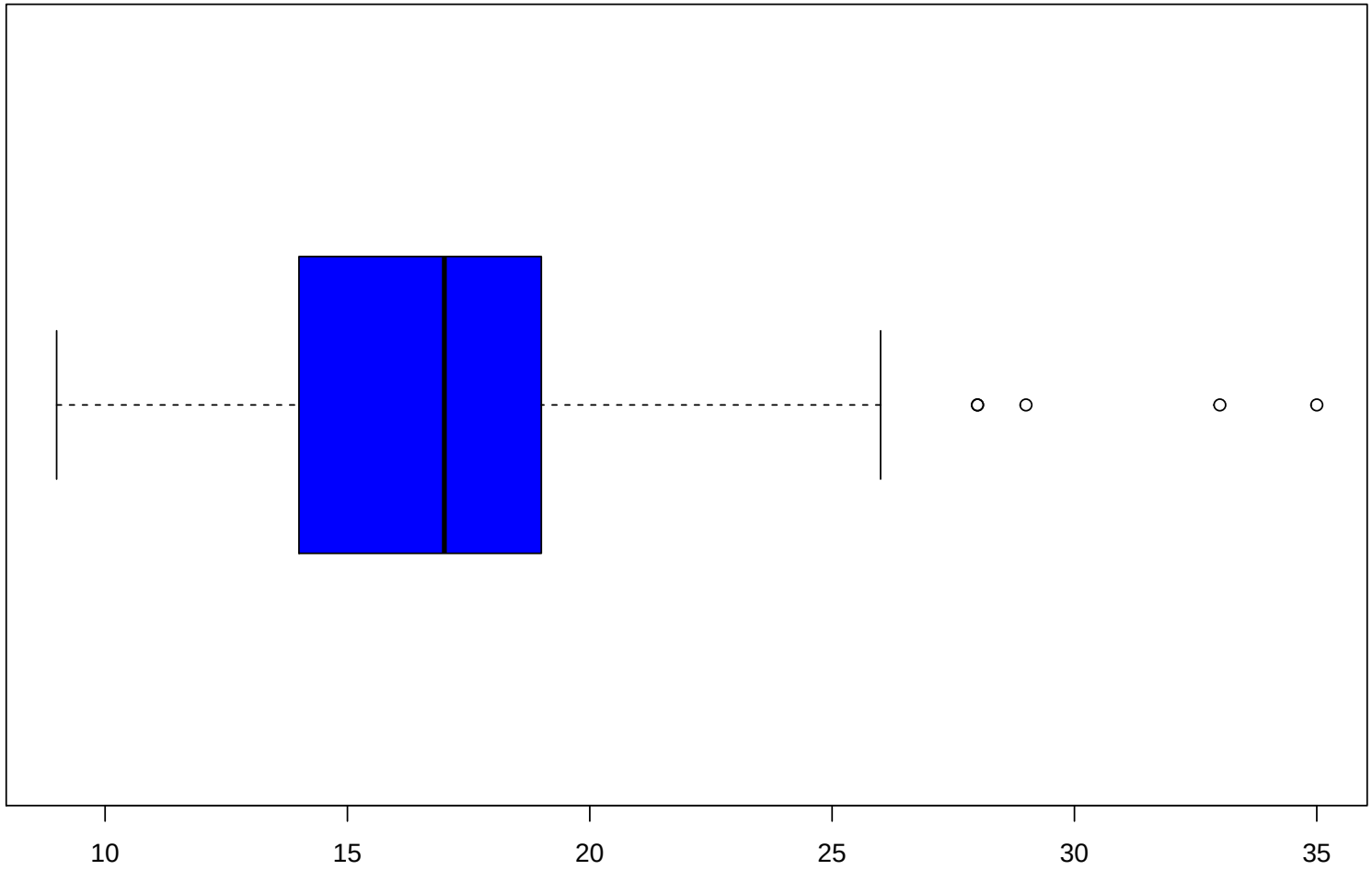
```
hist(cty, breaks=seq(6,36, 3),col="cyan")
```



Descriptive Statistics

Now, we can draw the BoxPlot of the `cty` data:

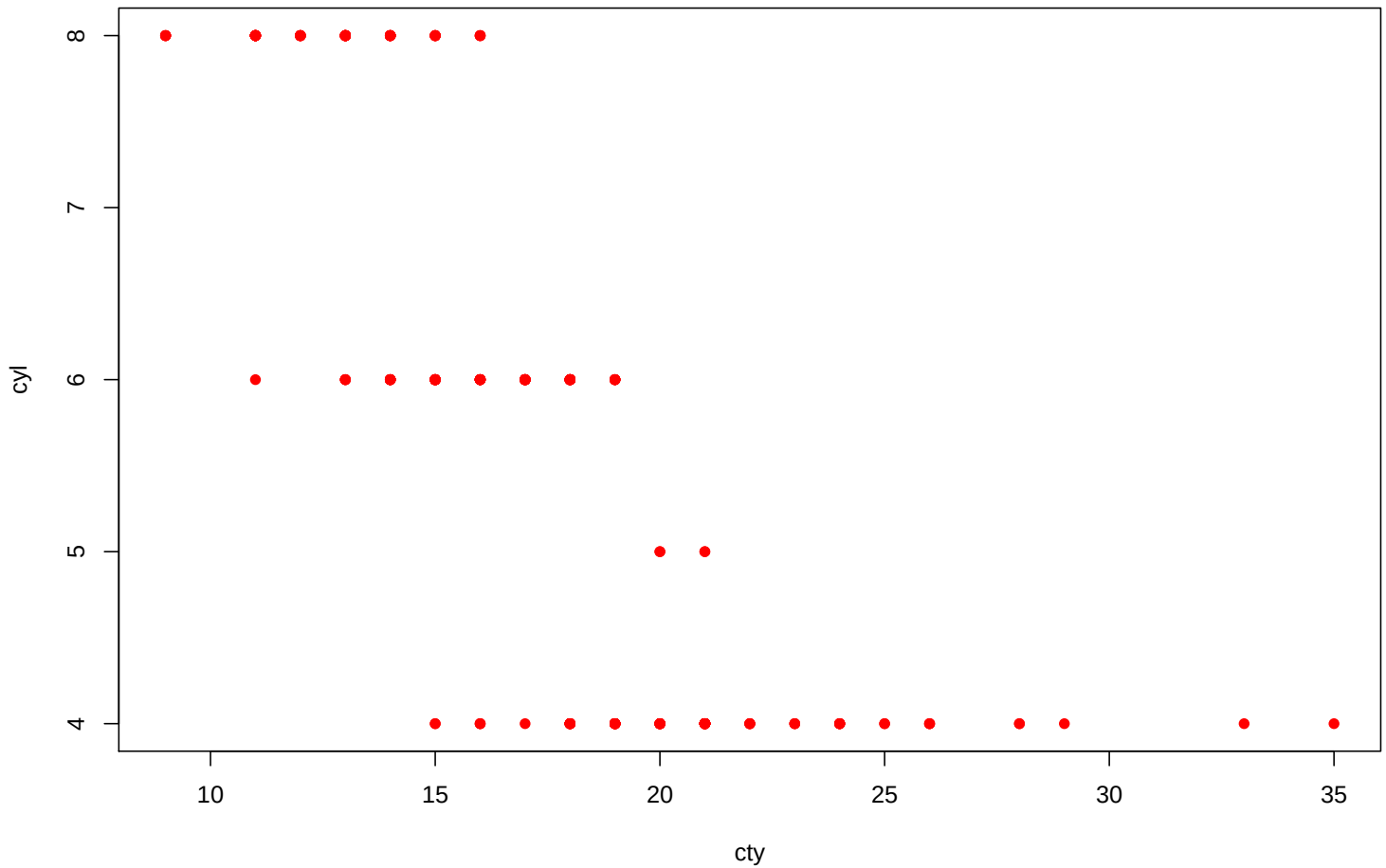
```
boxplot(cty, horizontal = T, col = "blue")
```



Descriptive Statistics

Now, instead of just getting information about `cyl` and `cty` separately, let us give visually the relationship between them:

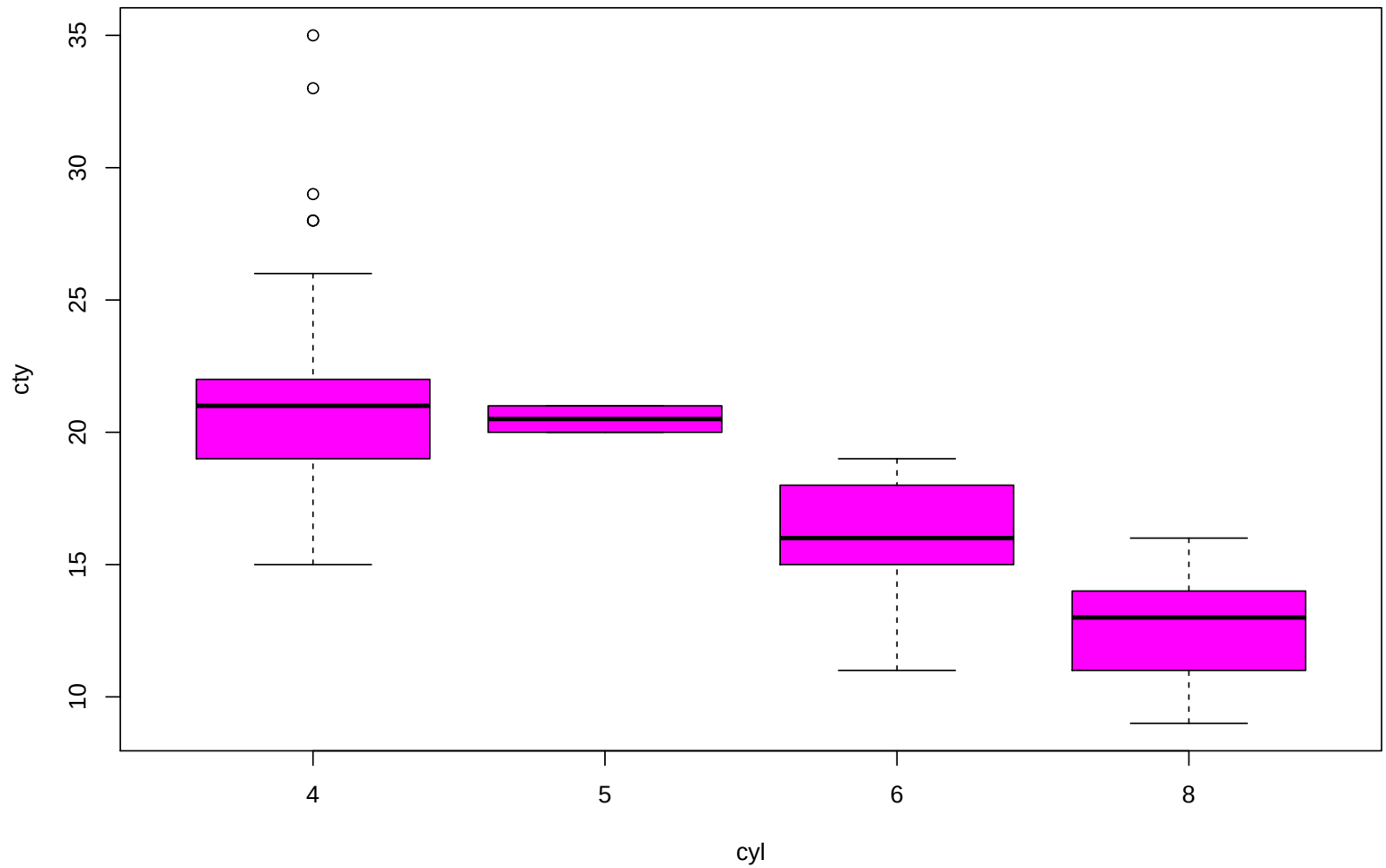
```
plot(cty, cyl, pch=16, col = "red")
```



Descriptive Statistics

. . . or draw a BoxPlot of `cty` for each type of the cylinder:

```
boxplot(cty~cyl, col="magenta")
```



Descriptive Statistics

Moral: our brain cannot get an insight from the list of numbers, but Descriptive Statistics can help ⌣̈