M358K - Applied Statistics (Fall 2017)

Professor: Ngoc M. Tran

UT Austin

Last updated: August 31, 2017

A few words about R

- R is THE software for statistical analysis. It is free and user-contributed
- R = calculator for statistics. R $\neq$ statistics

Cooking is easy. You put some stuff on the stove, burn it and then order pizza. My mom does it all the time.

- Install R: `https://cran.r-project.org/`
- Get familiar with R: `http://www.stat.berkeley.edu/share/rvideos/R_Videos/R_Videos.html`.
- More R resources, codes and lecture slides: Canvas
- Questions? Post on Piazza

How to analyze a dataset I: Descriptive Statistics

Descriptive statistics = tell me what you see

1. State the questions
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

How to analyze a dataset I: Descriptive Statistics

Descriptive statistics = tell me what you see

1. State the questions
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Example. Titanic

- Most famous shipwreck in history
- Only one-third survived
- Common perception: higher classes, women and children are better off. Is this true?

How to analyze a dataset I: Descriptive Statistics

Descriptive statistics = tell me what you see

1. State the questions
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Example. Titanic

- Most famous shipwreck in history
- Only one-third survived
- Common perception: higher classes, women and children are better off. Is this true?

The dataset. `data/titanic.csv` on Canvas.
2201 entries, one per person. 4 variables.

- `class`: crew, first, second, third
- age: adult, child
- `sex`: male, female
- `survived`: yes, no

Source: `https://ww2.amstat.org/publications/jse/v3n3/datasets.dawson.html`

How to analyze a dataset I: Descriptive Statistics

Descriptive statistics = tell me what you see

1. State the questions
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Example. Titanic

- Most famous shipwreck in history
- Only one-third survived
- Common perception: higher classes, women and children are better off. Is this true?

The dataset. `data/titanic.csv` on Canvas.
2201 entries, one per person. 4 variables.

- `class`: crew, first, second, third
- age: adult, child
- `sex`: male, female
- `survived`: yes, no

Source: `https://ww2.amstat.org/publications/jse/v3n3/datasets.dawson.html`

State the questions?

How to analyze a dataset I: Descriptive Statistics

Descriptive statistics = tell me what you see

1. State the questions
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Example. Titanic

- Most famous shipwreck in history
- Only one-third survived
- Common perception: higher classes, women and children are better off. Is this true?

The dataset. `data/titanic.csv` on Canvas.
2201 entries, one per person. 4 variables.

- `class`: crew, first, second, third
- `age`: adult, child
- `sex`: male, female
- `survived`: yes, no

Source: `https://ww2.amstat.org/publications/jse/v3n3/datasets.dawson.html`

1. What is the difference in survival rates between different classes?

How to analyze a dataset I: Descriptive Statistics

Descriptive statistics = tell me what you see

1. State the questions
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Example. Titanic

- Most famous shipwreck in history
- Only one-third survived
- Common perception: higher classes, women and children are better off. Is this true?

The dataset. `data/titanic.csv` on Canvas.
2201 entries, one per person. 4 variables.

- `class`: crew, first, second, third
- `age`: adult, child
- `sex`: male, female
- `survived`: yes, no

Source: `https://ww2.amstat.org/publications/jse/v3n3/datasets.dawson.html`

1. What is the difference in survival rates between different classes?
   Summarize the data?

How to analyze a dataset I: Descriptive Statistics

Descriptive statistics = tell me what you see

1. State the questions
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Example. Titanic

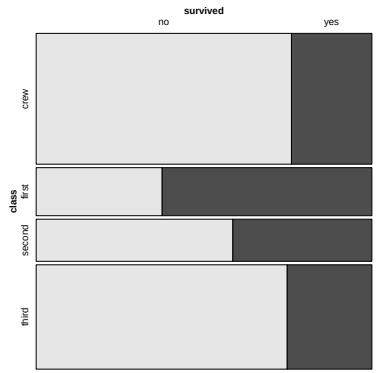
- Most famous shipwreck in history
- Only one-third survived
- Common perception: higher classes, women and children are better off. Is this true?

The dataset. `data/titanic.csv` on Canvas.
2201 entries, one per person. 4 variables.

- `class`: crew, first, second, third
- age: adult, child
- `sex`: male, female
- `survived`: yes, no

Source: `https://ww2.amstat.org/publications/jse/v3n3/datasets.dawson.html`

1. What is the difference in survival rates between different classes?



2.

3.

	crew	first	second	third
no	673	122	167	528
yes	212	203	118	178
survival %	24	62	41	25

How to analyze a dataset I: Descriptive Statistics

Descriptive statistics = tell me what you see

1. State the questions
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Example. Titanic

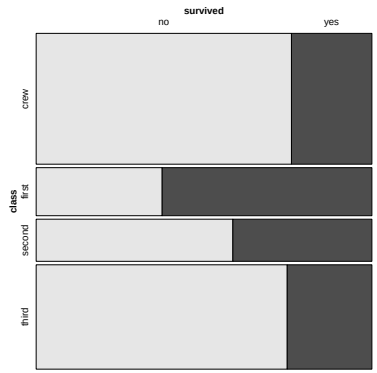
- Most famous shipwreck in history
- Only one-third survived
- Common perception: higher classes, women and children are better off. Is this true?

The dataset. `data/titanic.csv` on Canvas.
2201 entries, one per person. 4 variables.

- `class`: crew, first, second, third
- age: adult, child
- `sex`: male, female
- `survived`: yes, no

Source: `https://ww2.amstat.org/publications/jse/v3n3/datasets.dawson.html`

1. What is the difference in survival rates between different classes?



2.

3.

	crew	first	second	third
no	673	122	167	528
yes	212	203	118	178
survival %	24	62	41	25

Report findings?

How to analyze a dataset I: Descriptive Statistics

Descriptive statistics = tell me what you see

1. State the questions
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Example. Titanic

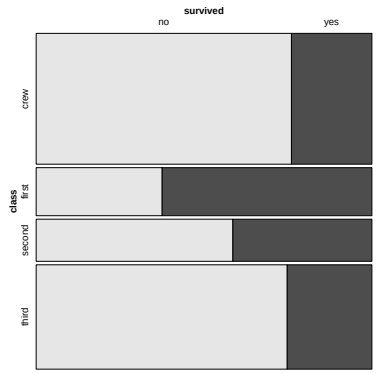
- Most famous shipwreck in history
- Only one-third survived
- Common perception: higher classes, women and children are better off. Is this true?

The dataset. `data/titanic.csv` on Canvas.
2201 entries, one per person. 4 variables.

- `class`: crew, first, second, third
- age: adult, child
- `sex`: male, female
- `survived`: yes, no

Source: `https://ww2.amstat.org/publications/jse/v3n3/datasets.dawson.html`

1. What is the difference in survival rates between different classes?



2.

3.

	crew	first	second	third
no	673	122	167	528
yes	212	203	118	178
survival %	24	62	41	25

4. Report: First class's survival rate is 62%, 1.5 times higher than second class, and 2.5 times higher than third class and crew.

How to analyze a dataset II: Inferential Statistics

Descriptive statistics = Tell me what you see
eg: What is the difference in survival rates between classes?

Inferential statistics = Is what you see significant?
eg: Is the difference in survival rates between classes unlikely due to chance alone?

How to analyze a dataset II: Inferential Statistics

Descriptive statistics = Tell me what you see
eg: What is the difference in survival rates between classes?

Inferential statistics = Is what you see significant?
eg: Is the difference in survival rates between classes (unlikely due to chance alone) statistically significant?

How to analyze a dataset II: Inferential Statistics

Descriptive statistics = Tell me what you see
eg: What is the difference in survival rates between classes?

Inferential statistics = Is what you see significant?
eg: Is the difference in survival rates between classes (unlikely due to chance alone) statistically significant?

Is 'significant' = 'large enough'?

How to analyze a dataset II: Inferential Statistics

Descriptive statistics = Tell me what you see
eg: What is the difference in survival rates between classes?

Report: First class's survival rate is **62%**, **1.5 times higher** than second class, and **2.5 times higher** than third class and crew.

Are these significant?

Inferential statistics = Is what you see significant?
eg: Is the difference in survival rates between classes (unlikely due to chance alone) statistically significant?

Is 'significant' = 'large enough'? How large is enough?

How to analyze a dataset II: Inferential Statistics

Descriptive statistics = Tell me what you see
eg: What is the difference in survival rates between classes?

1. State the question: what is ...?
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Inferential statistics = Is what you see significant?
eg: Is the difference in survival rates between classes (unlikely due to chance alone) statistically significant?

1. State the question: is ... significant or not?
2. Choose an appropriate statistical test / model
3. Do the test / fit the model
4. Report findings
5. Criticize the data AND the methods used

How to analyze a dataset II: Inferential Statistics

Descriptive statistics = Tell me what you see
eg: What is the difference in survival rates between classes?

1. State the question: what is ...?
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Inferential statistics = Is what you see significant?
eg: Is the difference in survival rates between classes (unlikely due to chance alone) statistically significant?

1. State the question: is ... significant or not?
2. Choose an appropriate statistical test / model
3. Do the test / fit the model
4. Report findings
5. Criticize the data AND the methods used

What is the hardest step in inference?

How to analyze a dataset II: Inferential Statistics

Descriptive statistics = Tell me what you see
eg: What is the difference in survival rates between classes?

1. State the question: what is ...?
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Inferential statistics = Is what you see significant?
eg: Is the difference in survival rates between classes (unlikely due to chance alone) statistically significant?

1. State the question: is ... significant or not?
2. Choose an appropriate statistical test / model
3. Do the test / fit the model
4. Report findings
5. Criticize the data AND the methods used

How to analyze a dataset II: Inferential Statistics

Descriptive statistics = Tell me what you see
eg: What is the difference in survival rates between classes?

1. State the question: what is ...?
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

Inferential statistics = Is what you see significant?
eg: Is the difference in survival rates between classes (unlikely due to chance alone) statistically significant?

1. State the question: is ... significant or not?
2. Choose an appropriate statistical test / model
3. Do the test / fit the model
4. Report findings
5. Criticize the data AND the methods used

Describe vs infer: which is more important? Harder?

How to analyze a dataset II: Inferential Statistics

Descriptive statistics = Tell me what you see
eg: What is the difference in survival rates between classes?

1. State the question: what is ...?
2. Summarize the data in pictures
3. Summarize the data in numbers
4. Report findings

- **No inference without descriptions!**
- **Both are hard to do well**
- **Both are widely used AND abused**

Inferential statistics = Is what you see significant?
eg: Is the difference in survival rates between classes (unlikely due to chance alone) statistically significant?

1. State the question: is ... significant or not?
2. Choose an appropriate statistical test / model
3. Do the test / fit the model
4. Report findings
5. Criticize the data AND the methods used