Descriptive statistics

Last updated: August 31, 2017

What we learned from Titanic

A dataset consists of **observations** and **variables**.
Often represented as a big matrix.
Each row = an observation, or a case. (eg: a person)
Each column = a variable (characteristic). (eg: `crew`)
When we summarize data, we summarize the variables and their relationships.

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. `data/titanic.csv` on Canvas.
2201 observations, 4 variables.

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

Numbers: what worked: tables.
Pictures: what worked

	class	age	sex	survived	class vs survived	class, sex vs survived
bar	✓	✓	✓	✓	✓	✓
mosaic					✓	✓

What we learned from Titanic

A dataset consists of **observations** and **variables**.
Often represented as a big matrix.
Each row = an observation, or a case. (eg: a person)
Each column = a variable (characteristic). (eg: `crew`)
When we summarize data, we summarize the variables and their relationships.

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. `data/titanic.csv` on Canvas.
2201 observations, 4 variables.

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

Numbers: what worked: tables.
Pictures: what worked

	class	age	sex	survived	class vs survived	class, sex vs survived
bar	✓	✓	✓	✓	✓	✓
mosaic					✓	✓

Any thing else we could use?

What we learned from Titanic

A dataset consists of **observations** and **variables**.
Often represented as a big matrix.
Each row = an observation, or a case. (eg: a person)
Each column = a variable (characteristic). (eg: `crew`)
When we summarize data, we summarize the variables and their relationships.

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. `data/titanic.csv` on Canvas.
2201 observations, 4 variables.

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

Numbers: what worked: tables.
Pictures: what worked

	class	age	sex	survived	class vs survived	class, sex vs survived
bar	✓	✓	✓	✓	✓	✓
mosaic					✓	✓
dot plot						
histogram						
frequency polygon						
density						
violin						
box plot						
scatter plot						

What we learned from Titanic

A dataset consists of **observations** and **variables**.
Often represented as a big matrix.
Each row = an observation, or a case. (eg: a person)
Each column = a variable (characteristic). (eg: `crew`)
When we summarize data, we summarize the variables and their relationships.

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. `data/titanic.csv` on Canvas.
2201 observations, 4 variables.

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

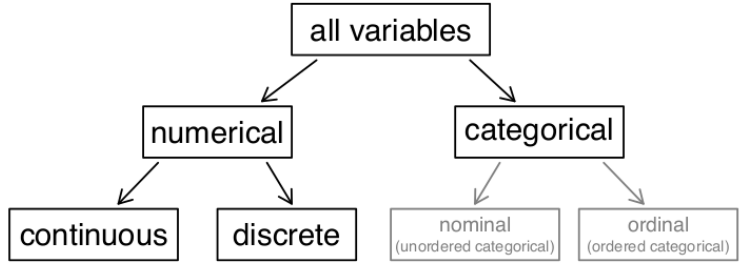
Numbers: what worked: tables.
Pictures: what worked

	class	age	sex	survived	class vs survived	class, sex vs survived
bar	✓	✓	✓	✓	✓	✓
mosaic					✓	✓
dot plot						
histogram						
frequency polygon						
density						
violin						
box plot						
scatter plot						

When to use what?

When to use what? Key: a variable's type

Each variable has a **type**, defined by what sort of values this variable can take.



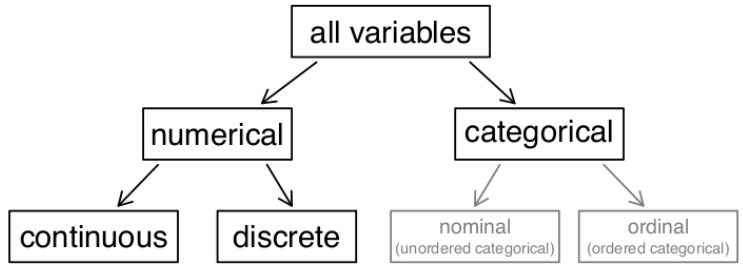
Picture guide	num.	cat.	num. vs num.	cat. vs cat.	num. vs cat.
bar		✓		✓	
mosaic				✓	
dot plot	✓				✓
histogram	✓				✓
frequency polygon	✓				✓
density	✓				✓
violin	✓				✓
box plot	✓				✓
scatter plot			✓		

Number guide	num.	cat.	num. vs num.	cat. vs cat.	num. vs cat.
frequency table		✓			
contingency table				✓	
min, max, median, quantiles	✓				✓
mean, standard deviation	✓				✓
correlation, R^2			✓		

The tables above are GUIDES, not rules. The only rule is: display the most informative pictures and numbers

When to use what? Key: a variable's type

Each variable has a **type**, defined by what sort of values this variable can take.



Picture guide	num.	cat.	num. vs num.	cat. vs cat.	num. vs cat.
bar		✓		✓	
mosaic				✓	
dot plot	✓				✓
histogram	✓				✓
frequency polygon	✓				✓
density	✓				✓
violin	✓				✓
box plot	✓				✓
scatter plot			✓		

Number guide	num.	cat.	num. vs num.	cat. vs cat.	num. vs cat.
frequency table		✓			
contingency table				✓	
min, max, median, quantiles	✓				✓
mean, standard deviation	✓				✓
correlation, R^2			✓		

The tables above are GUIDES, not rules. The only rule is: display the most informative pictures and numbers
R assigns default types, but these are often wrong. You must tell R what the correct types are, so commands work as intended.

Example 2: emails

Problem: How can a machine decide if an incoming email is spam or not?

The Emails dataset. `data/emails.csv` on Canvas.
3921 lines, one per email, 21 variables. Variable description: `data/emails-descrip.txt`

Variables summary

- `spam`: yes/no
- `to_multiple`, `from`, `winner`, `format`, `re_subj`, `exclaim_subj`, `urgent_subj`: yes/no
- `number`: none, small, big
- `line_breaks`, `cc`, `image`, `attach`, `dollar`, `inherit`, `viagra`, `password`, `exclaim_mess`: integer
- `num_char`: integer/1000 (unit: thousands)
- `time`: time

Question: Are any of these variables useful markers for spam?

Example 2: emails

Problem: How can a machine decide if an incoming email is spam or not?

The Emails dataset. `data/emails.csv` on Canvas.
3921 lines, one per email, 21 variables. Variable description: `data/emails-descrip.txt`

Variables summary

- `spam`: yes/no
- `to_multiple, from, winner, format, re_subj, exclaim_subj, urgent_subj`: yes/no
- `number`: none, small, big
- `line_breaks, cc, image, attach, dollar, inherit, viagra, password, exclaim_mess`: integer
- `num_char`: integer/1000 (unit: thousands)
- `time`: time

Question: Are any of these variables useful markers for spam?
What types are these?

Example 2: emails

Problem: How can a machine decide if an incoming email is spam or not?

The Emails dataset. `data/emails.csv` on Canvas.
3921 lines, one per email, 21 variables. Variable description: `data/emails-descrip.txt`

Variables summary

- `spam`: yes/no **categorical**
- `to_multiple, from, winner, format, re_subj, exclaim_subj, urgent_subj`: yes/no **categorical**
- `number`: none, small, big **ordinal**
- `line_breaks, cc, image, attach, dollar, inherit, viagra, password, exclaim_mess`: integer **numeric (integer)**
- `num_char`: integer/1000 (unit: thousands) **numeric (real)**
- `time`: time **ordinal**

Question: Are any of these variables useful markers for spam?

What is a ...

bar plot?

- y-axis: count (frequency), x-axis: variable values

mosaic?

- represent each cell in a table by rectangles, whose areas are proportional to the frequency by that group.

scatter plot?

- y vs x for two numerical variables x,y

dot plot?

- A one-variable scatter plot
- y-axis: one line per observation. x-axis: variable values

histogram?

- A numeric version of barplot
- y-axis: count (frequency), x-axis: variable values, grouped in bins

frequency polygon?

- An interpolated version of histogram
- y-axis: counts (frequency), x-axis: variable values. Starting from a histogram, adjacent y-values are interpolated to create a polygon for all x-values

density plot?

- A scaled and smoothed version of frequency polygon
- y-axis: counts normalized, x-axis: variable values. Starting from a histogram, adjacent y-values are locally averaged to create a smooth curve for all x-values. Then the whole curve is scaled so that the total area sums to 1.

violin plot?

- Density plot tilted sideway and mirrored. Great for comparison between multiple densities.

box plot?

- Middle line: median, box: 25th and 75th percentile (interquartile ragne), whisker lengths: at most 1.5 times the IQR, dots: points outside.
- Great for detecting extreme values, comparing distributions