STA130H1S TUT0109
W2: Data Wrangling

Jan 18, 2019

Email: sonia.chhay@mail.utoronto.ca
Website: soniachhay.github.io

Announcements

- Attendance sheet will be at the front of class
  - Must arrive to tutorial on time and sign in to receive attendance mark
- Next week will be a half tutorial (Jan 25)
  - Other half is for mentorship worth 3% of your grade

Writing tips!

- Provide some context (mention the variables/use the vocab learned)
  - E.g. Where was the data from, what variables are used and the units, what is being analyzed
- Should provide more detail in answers (makes it easier for readers to understand/visualize)
  - Transition words should also be used to help your words flow better
- Take note of important details in graph and its significance
  - Can describe the distribution (if you need reference, the vocab words are on Quercus)
  - E.g. If positive relation, what does that mean?
    - Always ask yourself “why?”

Want to improve your English language skills?

English language learning and writing centres (Check “Course Syllabus & Help” page) :)

Overview

- Vocabulary for this week
- Review Problem Set #2, Question (1)
- Group discussion on Question (3)
- Writing example
- Writing exercise

Vocabulary for this week:

- Average
- Variance
- Standard deviation
- Data frame
- Vector
- Types of variables
  - Numeric, character, logical, etc.
- Matrix
- Missing data

Question (1)

Used Galton dataset which contained heights of adult children and their parents in inches

Question (1)

- Used Galton dataset which contained heights of adult children and their parents in inches

```
## Observations: 898
## Variables: 6
## $ family <fct> 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5...
## $ father <dbl> 78.5, 78.5, 78.5, 78.5, 75.5, 75.5, 75.5, 75.5, 75.0, 7...
## $ mother <dbl> 67.0, 67.0, 67.0, 67.0, 66.5, 66.5, 66.5, 66.5, 64.0, 6...
## $ sex    <fct> M, F, F, F, M, M, F, F, M, F, M, M, F, F, F, M, M, M, F...
## $ height <dbl> 73.2, 69.2, 69.0, 69.0, 73.5, 72.5, 65.5, 65.5, 71.0, 6...
## $ nkids  <int> 4, 4, 4, 4, 4, 4, 4, 4, 2, 2, 5, 5, 5, 5, 5, 6, 6, 6, 6...
```

Question (1a)

Calculate average and variance of children's heights

Average, **variance**, **standard deviation**

- **Average**: Mean of a given set of data
- **Variance**: Average degree to which each point varies from the mean (i.e. Spread)
- **Standard deviation**: Square root of variance
  - Unlike variance, standard deviation is measured in the same units as the data

Question (1a)

- Calculated average and variance of children's heights in the first three families using the filter function

```
filter(Galton, family == 1 | family == 2 | family == 3)
```

```
##    family father mother sex height nkids
## 1       1   78.5   67.0   M   73.2     4
## 2       1   78.5   67.0   F   69.2     4
## 3       1   78.5   67.0   F   69.0     4
## 4       1   78.5   67.0   F   69.0     4
## 5       2   75.5   66.5   M   73.5     4
## 6       2   75.5   66.5   M   72.5     4
## 7       2   75.5   66.5   F   65.5     4
## 8       2   75.5   66.5   F   65.5     4
## 9       3   75.0   64.0   M   71.0     2
## 10      3   75.0   64.0   F   68.0     2
```

Question (1a)

```
summarise(filter(Galton, family == 1), n= n(), mean = mean(height), var = (sd(height))^2
)
```

```
##   n mean  var
## 1 4 70.1 4.28
```

```
summarise(filter(Galton, family == 2), n= n(), mean = mean(height), var = (sd(height))^2
)
```

```
##   n  mean      var
## 1 4 69.25 18.91667
```

```
summarise(filter(Galton, family == 3), n= n(), mean = mean(height), var = (sd(height))^2
)
```

```
##   n mean var
## 1 2 69.5 4.5
```

Question (1a)

(1) Which family had the largest variance?
(2) What does it mean in this context?

```
summarise(filter(Galton, family == 1), n= n(), mean = mean(height), var = (sd(height))^2
)
```

```
##   n mean  var
## 1 4 70.1 4.28
```

```
summarise(filter(Galton, family == 2), n= n(), mean = mean(height), var = (sd(height))^2
)
```

```
##   n  mean      var
## 1 4 69.25 18.91667
```

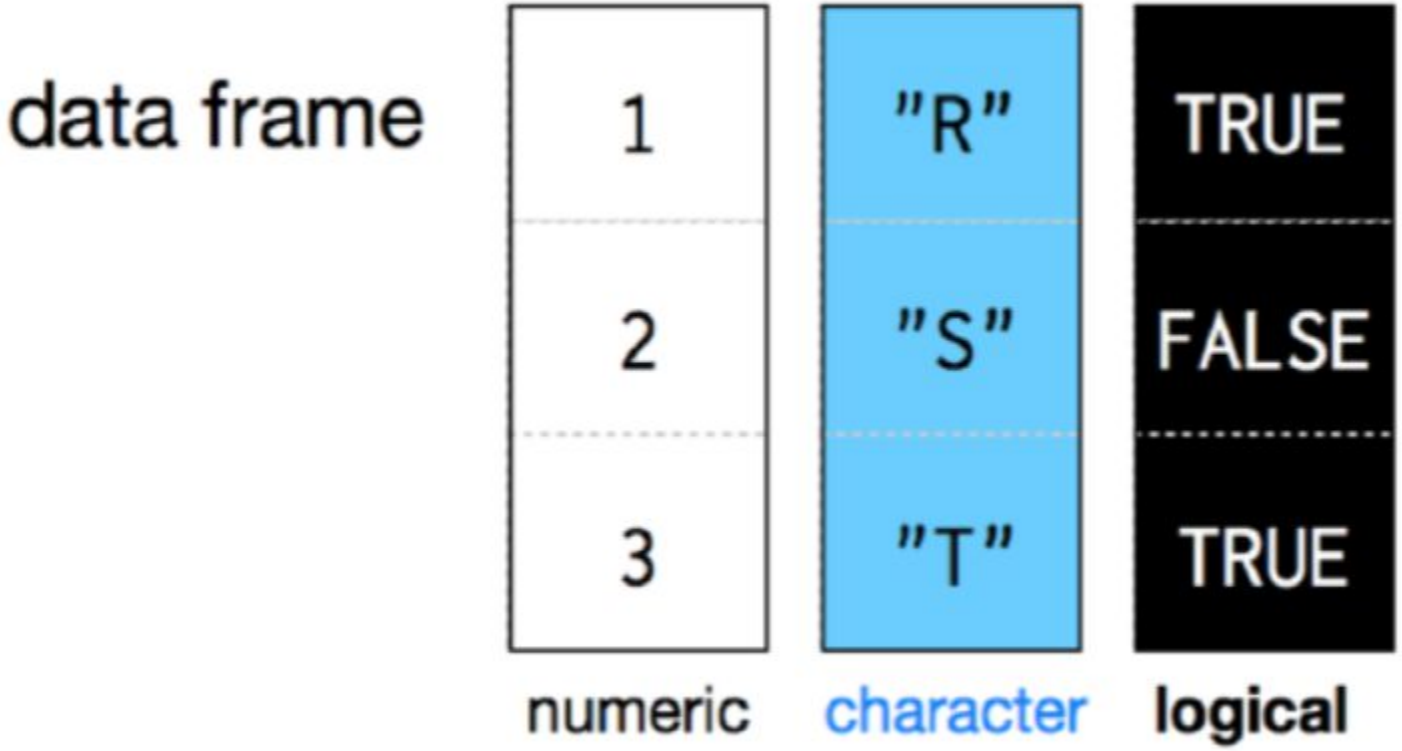
```
summarise(filter(Galton, family == 3), n= n(), mean = mean(height), var = (sd(height))^2
)
```

```
##   n mean var
## 1 2 69.5 4.5
```

Question (1b)

Create a data frame that contains the family ID# and the number of kids in each family

Data frame, vectors, and data type



Question 1(b)

- Created a data frame called data that contains the family id number and the numbers of kids in each family
- group_by() used together with summarise()
  - Group data by variable chosen
  - Output for summarise() will have one row for each group

```
library(tidyverse)
data <- summarise(group_by(Galton, family),
numkids = mean(nkids))
data<-data.frame(data)
```

Question 1(b)

- Created a data frame called data that contains the family id number and the numbers of kids in each family
- group_by() used together with summarise()
  - Group data by variable chosen
  - Output for summarise() will have one row for each group

```
library(tidyverse)
data <- summarise(group_by(Galton, family),
numkids = mean(nkids))
data<-data.frame(data)
```

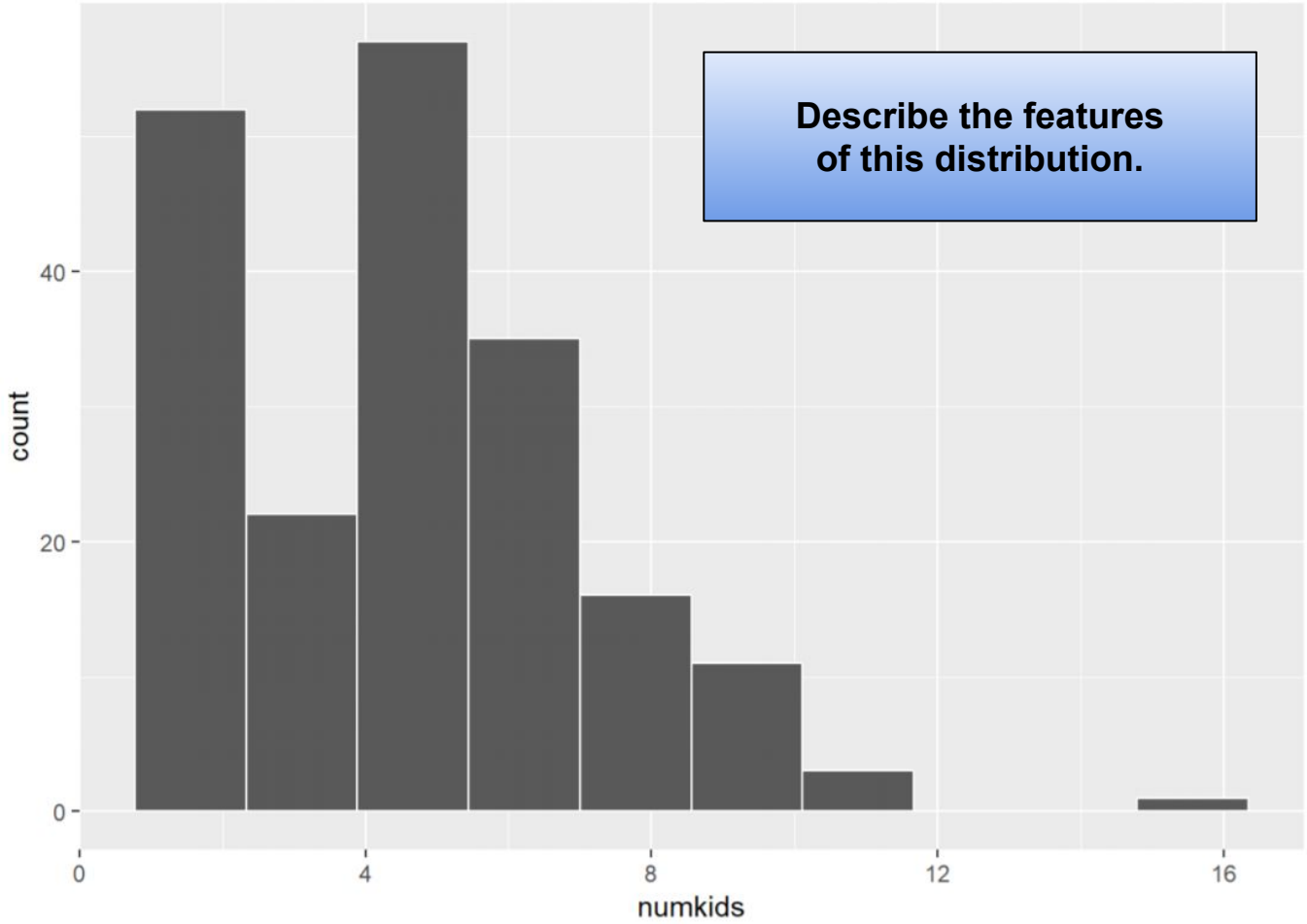
family <fctr>	numkids <dbl>
1	4
10	1
100	3
101	6
102	6
103	5
104	4
105	6
106	7
107	9

1-10 of 197 rows

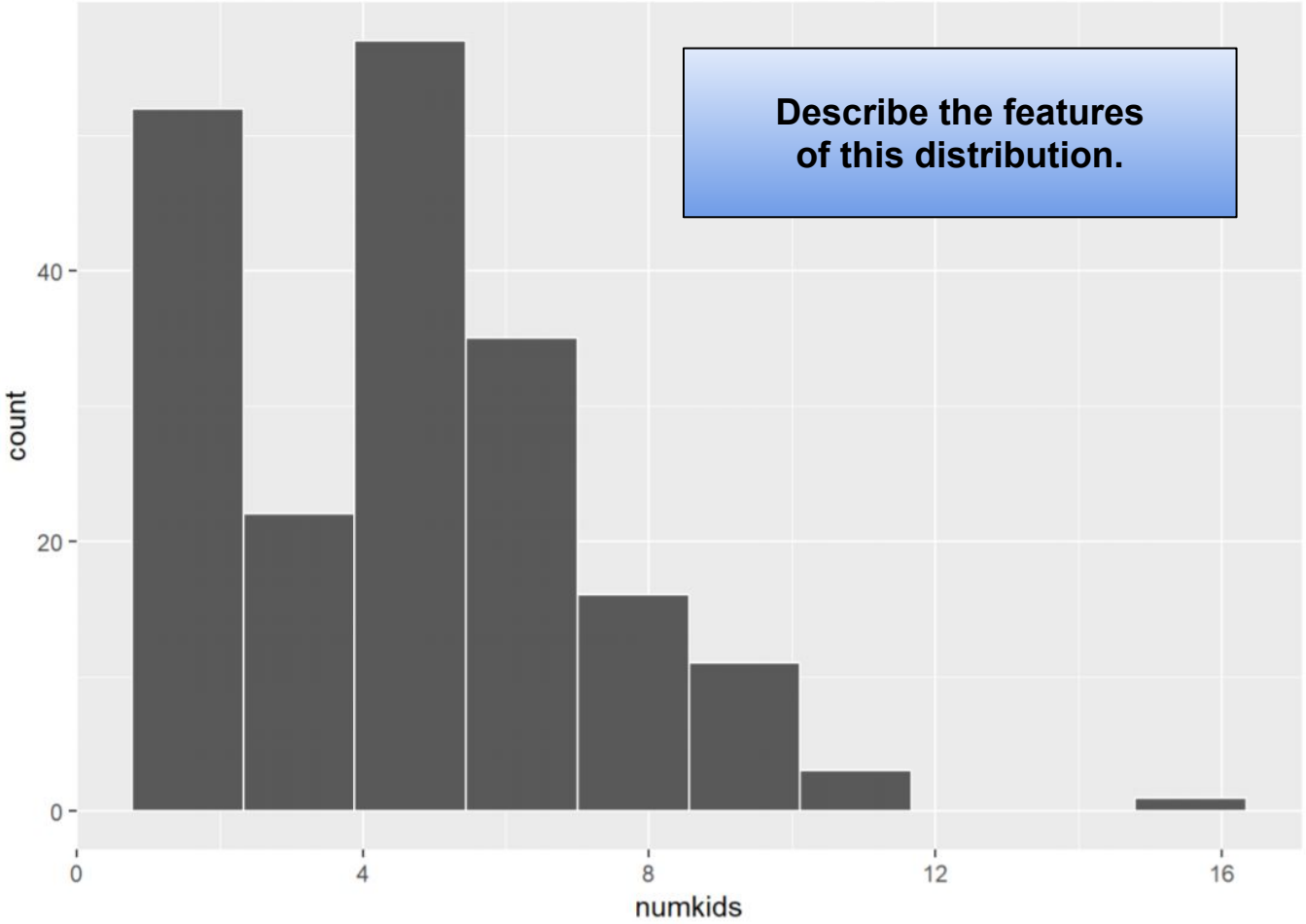
Question (1c)

Graph the distribution of the number of kids
in the Galton dataset families

```
ggplot(data) + aes(x = numkids) + geom_histogram(bins=10,color="white")
```



```
ggplot(data) + aes(x = numkids) + geom_histogram(bins=10,color="white")
```



Features:

- Positively skewed
- Range
- Mode
- Outlier

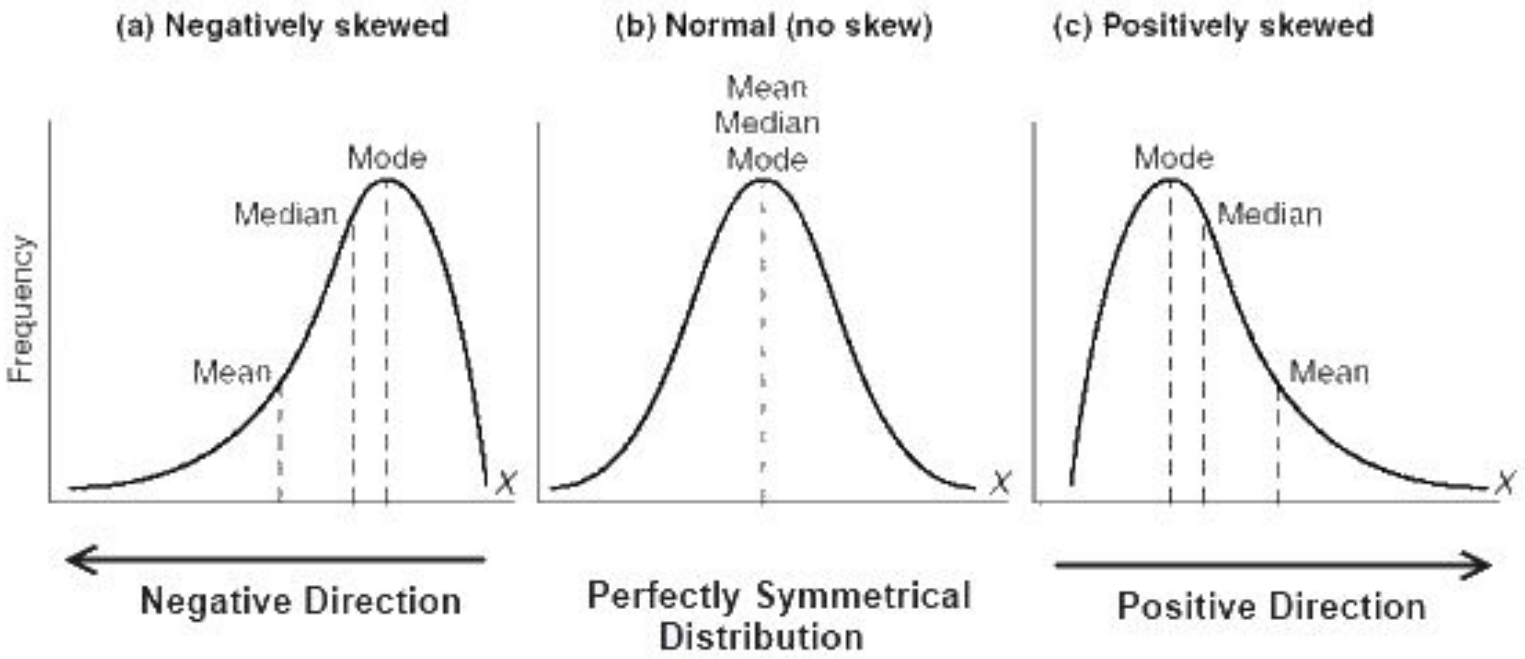
Question (1d)

Based on the graph from (1c),
how do you think mean and median would compare?

Some Measures of Central Tendency

- **Average**: Mean of a given set of data
- **Median**: A value such that 50% of the observations are less than the median and 50% of the observations are greater than it
  - I.e. The value in the middle

Measures of Central Tendency



Mean < Median < Mode

Mode < Median < Mean

Matrix

- Grid with more than one row and column
- Can only contain one type of data