Lecture 2 practice

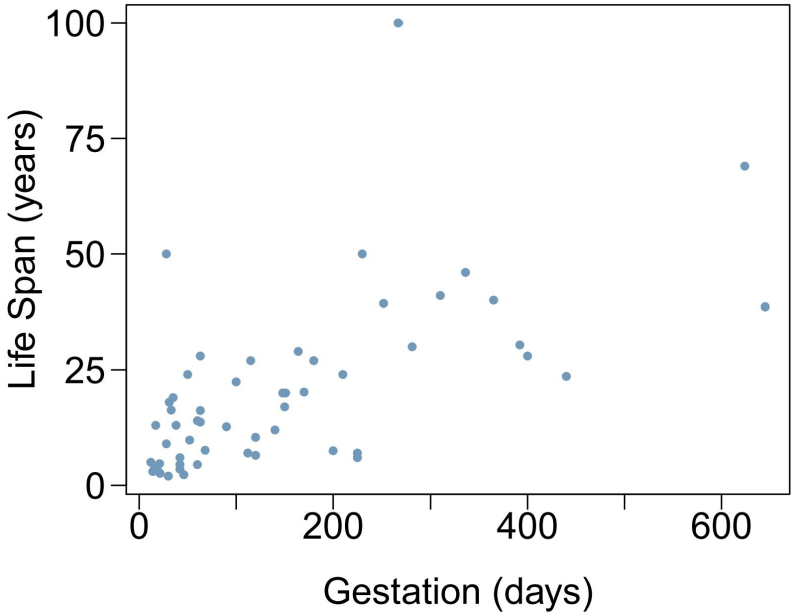
August 4, 2022

Mammal life spans

Data were collected on life spans (in years) and gestation lengths (in days) for 62 mammals. A scatterplot of life span versus length of gestation is shown below.

What type of an association is apparent between life span and length of gestation?

Positive association: mammals with longer gestation periods tend to live longer as well.

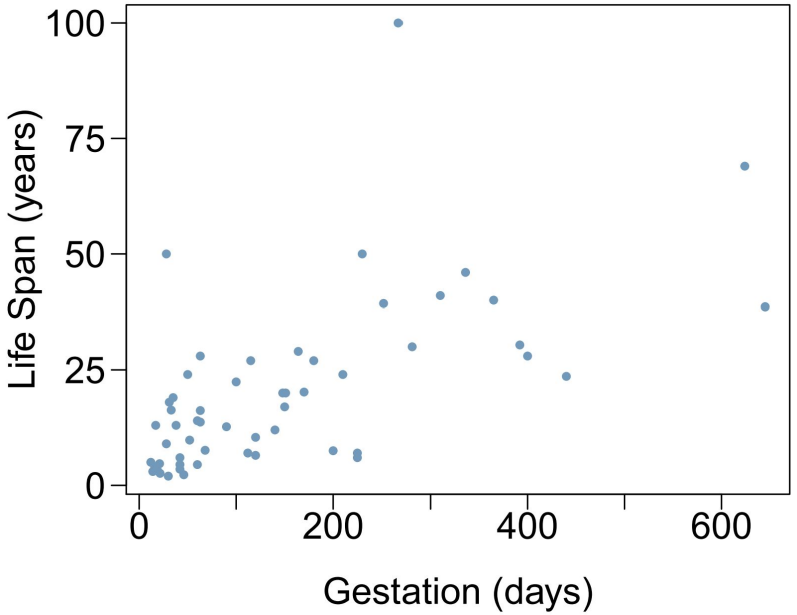


Mammal life spans

Data were collected on life spans (in years) and gestation lengths (in days) for 62 mammals. A scatterplot of life span versus length of gestation is shown below.

What type of an association would you expect to see if the axes of the plot were reversed, i.e. if we plotted length of gestation versus life span?

Association would still be positive

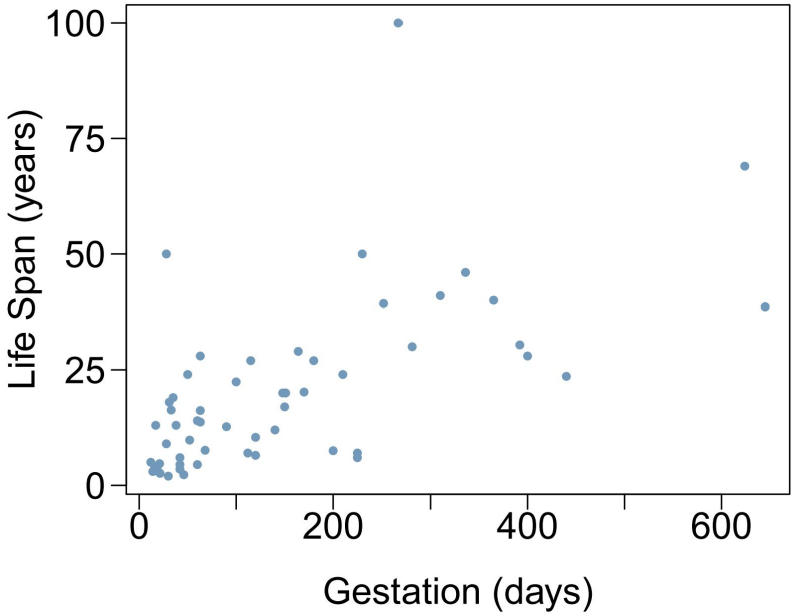


Mammal life spans

Data were collected on life spans (in years) and gestation lengths (in days) for 62 mammals. A scatterplot of life span versus length of gestation is shown below.

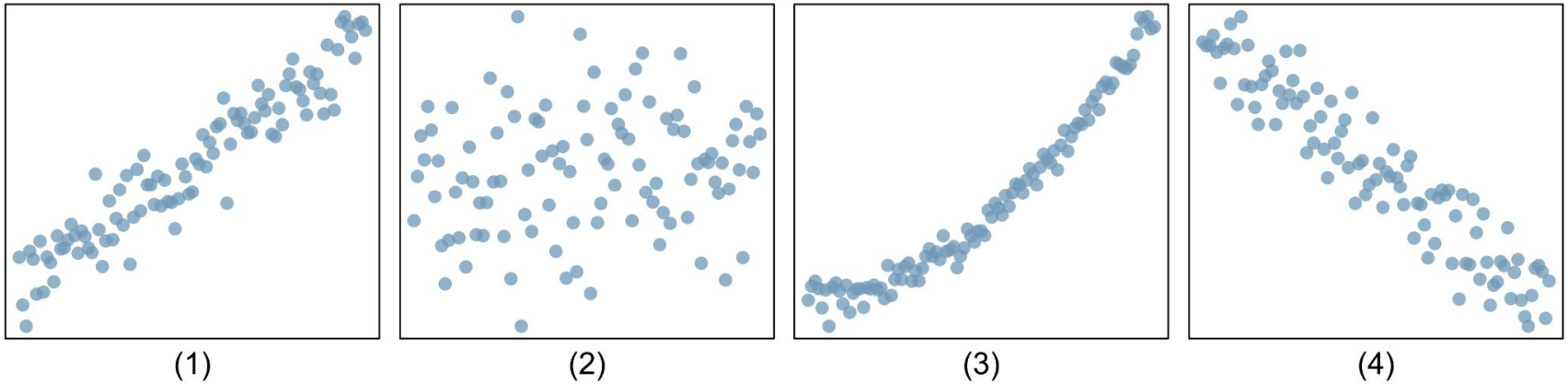
Are life span and length of gestation independent? Explain your reasoning.

No, they are not independent. Having an association means they are dependent variables.



Associations

Indicate which of the plots show (a) a positive association, (b) a negative association, or (c) no association. Also determine if the positive and negative associations are linear or nonlinear. Each part may refer to more than one plot.



Positive linear

No association

Positive nonlinear

Negative linear

Sample statistic versus population parameter

A recent article in a college newspaper stated that college students get an average of 5.5 hrs of sleep each night. A student who was skeptical about this value decided to conduct a survey by randomly sampling 25 students. On average, the sampled students slept 6.25 hours per night.

Identify which value represents the sample mean and which value represents the claimed population mean.

Sample mean: 6.25 hours per night, Population mean: 5.5 hours per night

Days off at a mining plant

Workers at a particular mining site receive an average of 35 days paid vacation, which is lower than the national average. The manager of this plant is under pressure from a local union to increase the amount of paid time off. However, he does not want to give more days off to the workers because that would be costly. Instead he decides he should fire 10 employees in such a way as to raise the average number of days off that are reported by his employees.

In order to achieve this goal, should he fire employees who have the most number of days off, least number of days off, or those who have about the average number of days off?

Any 10 employees whose average number of days off is between the minimum and the mean number of days off for the entire workforce at this plant.

Medians and IQRs

For each part, compare distributions (1) and (2) based on their medians and IQRs. You do not need to calculate these statistics; simply state how the medians and IQRs compare. Make sure to explain your reasoning.

(1) 3, 5, 6, 7, 9
(2) 3, 5, 6, 7, 20

median 1 = median 2; IQR 1 = IQR 2

(1) 3, 5, 6, 7, 9
(2) 3, 5, 7, 8, 9

median 1 < median 2; IQR 1 < IQR 2

(1) 1, 2, 3, 4, 5
(2) 6, 7, 8, 9, 10

median 1 < median 2; IQR 1 = IQR 2

Means and standard deviations (sd)

For each part, compare distributions (1) and (2) based on their means and standard deviations. You do not need to calculate these statistics; simply state how the means and the standard deviations compare. Make sure to explain your reasoning.

(1) 3, 5, 5, 5, 8, 11, 11, 11, 13
(2) 3, 5, 5, 5, 8, 11, 11, 11, 20

mean 1 < mean 2; sd 1 < sd 2

(1) -20, 0, 0, 0, 15, 25, 30, 30
(2) -40, 0, 0, 0, 15, 25, 30, 30

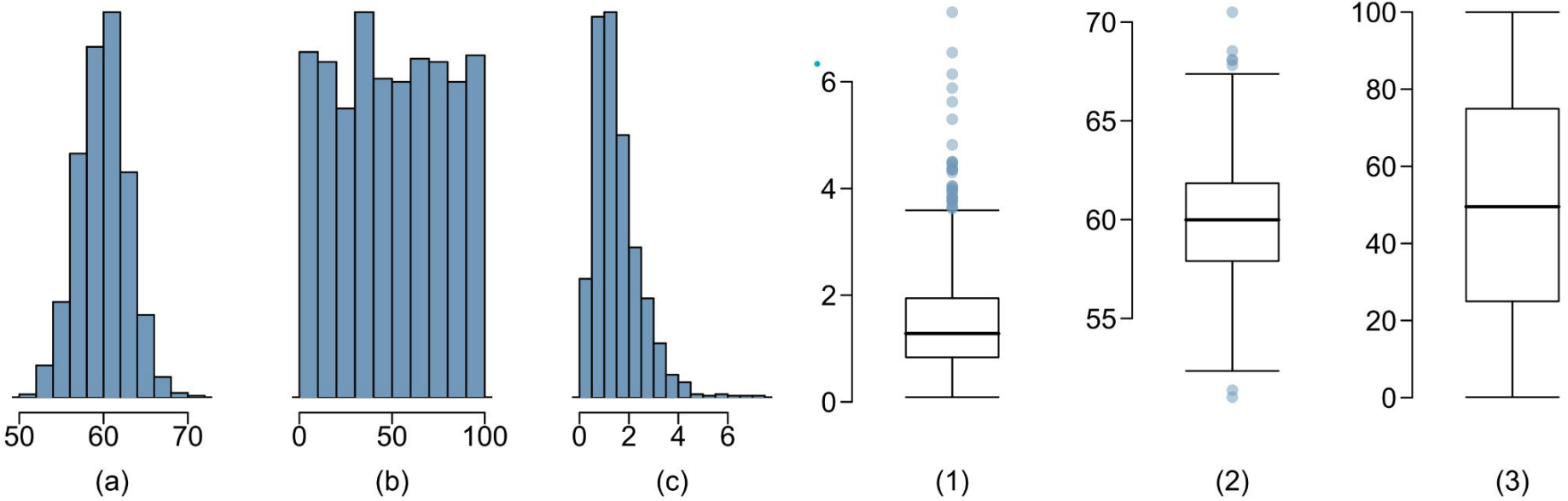
mean 1 > mean 2; sd 1 < sd 2

(1) 0, 2, 4, 6, 8, 10
(2) 20, 22, 24, 26, 28, 30

mean 1 < mean 2; sd 1 = sd 2

Visualizing a distribution

Describe the distribution in the histograms below and match them to the box plots



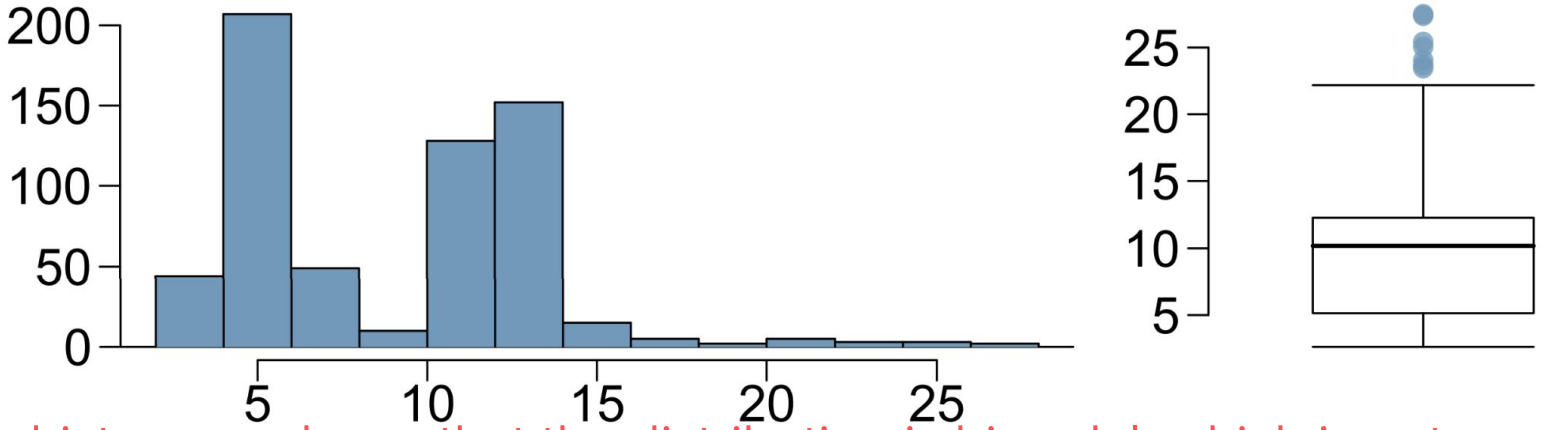
a-2 unimodal, symmetric

b-3 uniform

c-1 unimodal, right skewed

Histograms vs. box plots

Compare the two plots below. What characteristics of the distribution are apparent in the histogram and not in the box plot? What characteristics are apparent in the box plot but not in the histogram?



The histogram shows that the distribution is bimodal, which is not apparent in the box plot. The box plot makes it easy to identify more precise values of observations outside of the whiskers.

Distributions and appropriate statistics

For each of the following

- state whether you expect the distribution to be symmetric, right skewed, or left skewed
- specify whether the mean or median would best represent a typical observation in the data
- specify the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(a) Number of pets per household.

The distribution of number of pets per household is likely right skewed as there is a natural boundary at 0 and only a few people have many pets. Therefore the center would be best described by the median, and variability would be best described by the IQR.

Distributions and appropriate statistics

For each of the following

- state whether you expect the distribution to be symmetric, right skewed, or left skewed
- specify whether the mean or median would best represent a typical observation in the data
- specify the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

(b) Distance to work, i.e. number of miles between work and home.

The distribution of number of distance to work is likely right skewed as there is a natural boundary at 0 and only a few people live a very long distance from work. Therefore the center would be best described by the median, and variability would be best described by the IQR.

Distributions and appropriate statistics

For each of the following

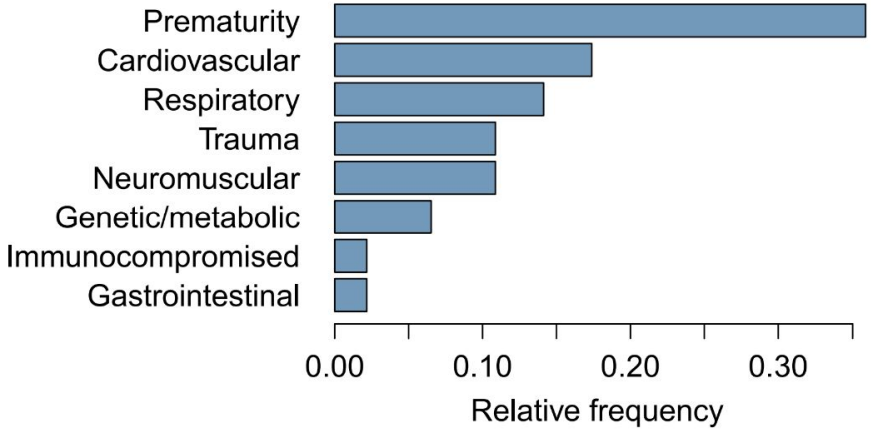
- state whether you expect the distribution to be symmetric, right skewed, or left skewed
- specify whether the mean or median would best represent a typical observation in the data
- specify the variability of observations would be best represented using the standard deviation or IQR. Explain your reasoning.

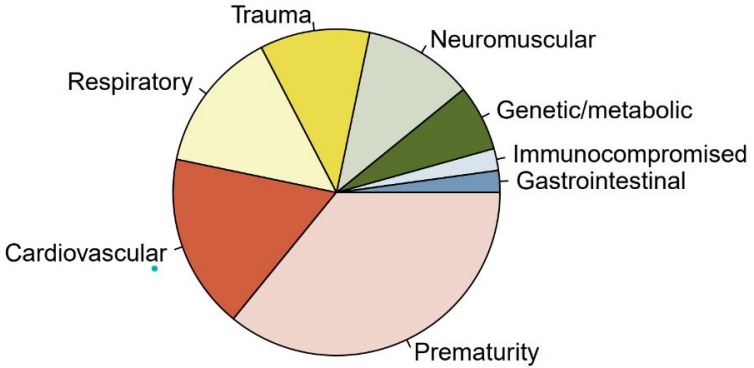
(c) Heights of adult males

The distribution of heights of males is likely symmetric. Therefore the center would be best described by the mean, and variability would be best described by the standard deviation.

Antibiotic use in children

The bar plot and the pie chart below show the distribution of pre-existing medical conditions of children involved in a study on the optimal duration of antibiotic use in treatment of tracheitis, which is an upper respiratory infection.



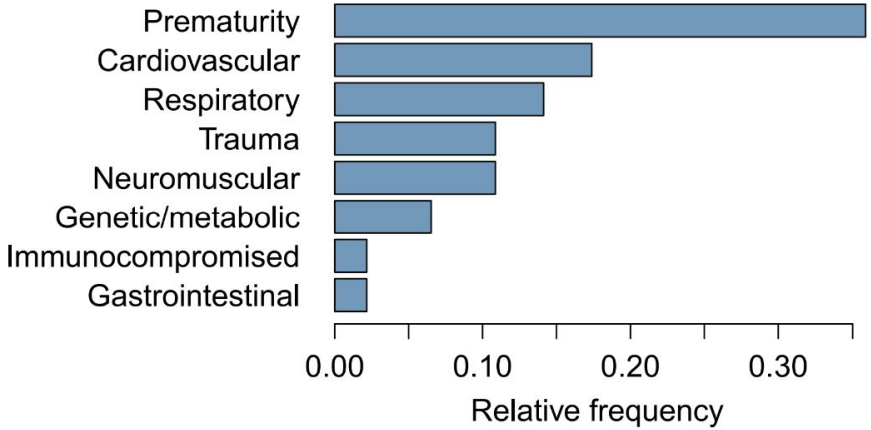


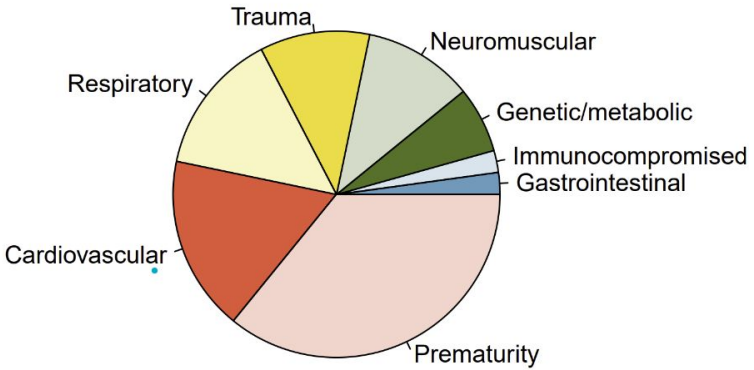
(a) What features are apparent in the bar plot but not in the pie chart?

We see the order of the categories and the relative frequencies in the bar plot.

Antibiotic use in children

The bar plot and the pie chart below show the distribution of pre-existing medical conditions of children involved in a study on the optimal duration of antibiotic use in treatment of tracheitis, which is an upper respiratory infection.



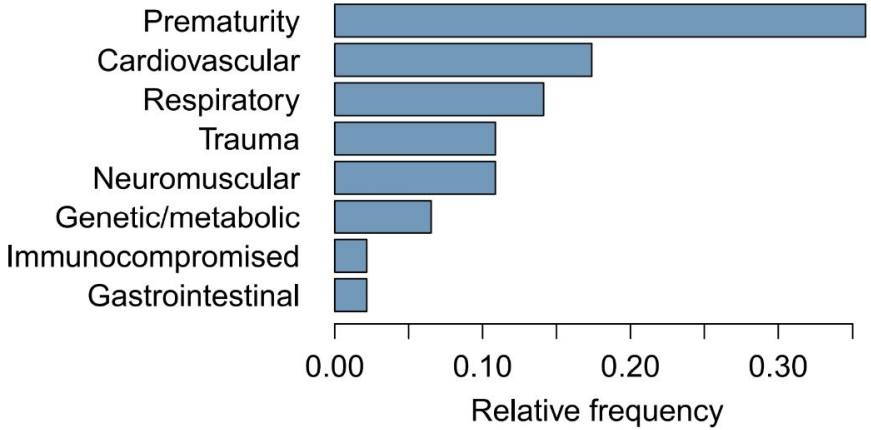


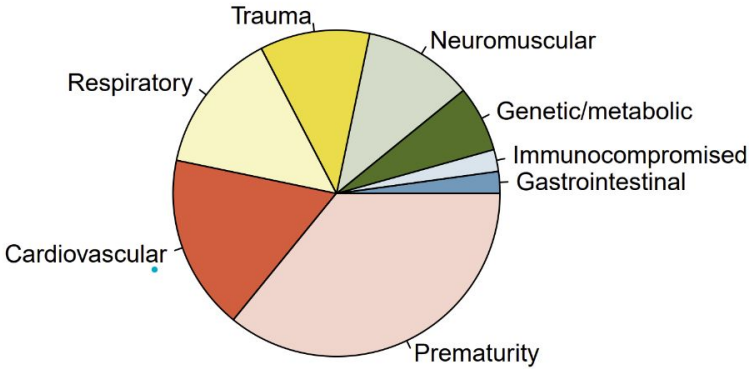
(b) What features are apparent in the pie chart but not in the bar plot?

There are no features that are apparent in the pie chart but not in the bar plot.

Antibiotic use in children

The bar plot and the pie chart below show the distribution of pre-existing medical conditions of children involved in a study on the optimal duration of antibiotic use in treatment of tracheitis, which is an upper respiratory infection.



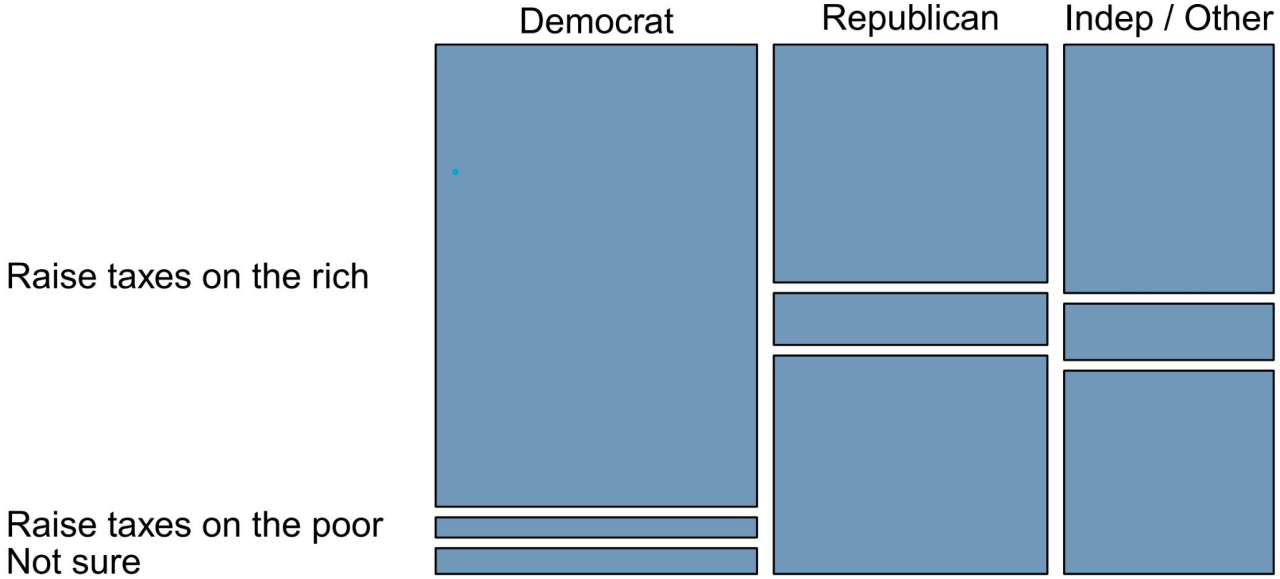


(c) Which graph would you prefer to use for displaying these categorical data?

We usually prefer to use a bar plot as we can also see the relative frequencies of the categories in this graph.

Raise

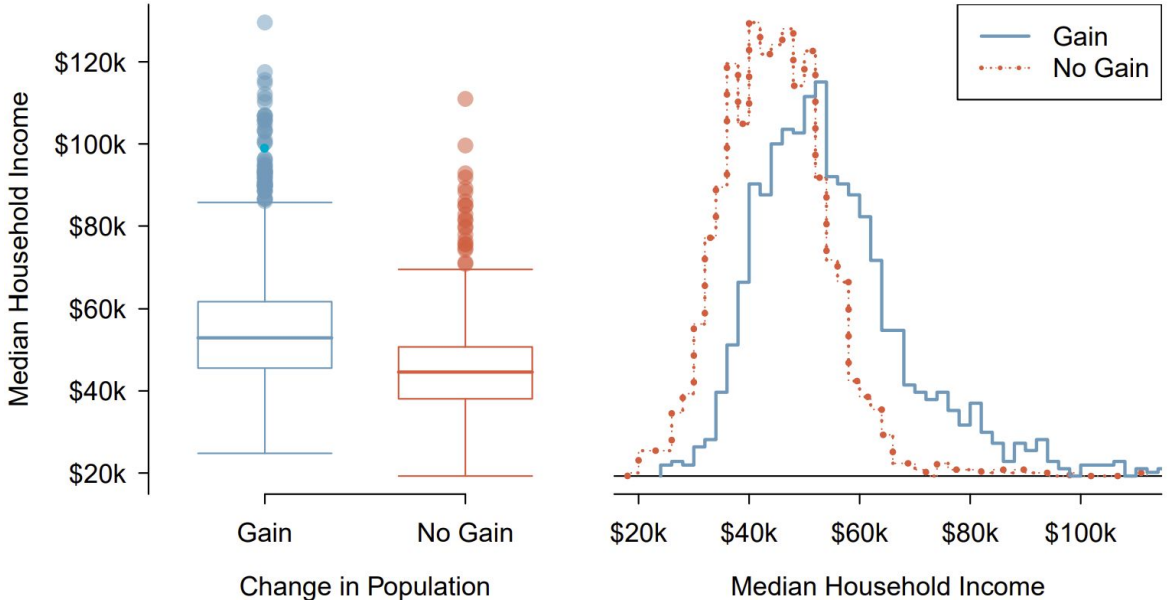
A random sample of registered voters nationally were asked whether they think it's better to raise taxes on the rich or raise taxes on the poor. Based on the mosaic plot shown below, do views on raising taxes and political affiliation appear to be independent? Explain your reasoning.



The vertical locations of the breaks differ, suggesting these variables are dependent.

Comparing distributions for counties

- What do you notice about the approximate center of each group?
- What do you notice about the variability between groups?
- Is the shape relatively consistent between groups?
- How many prominent modes are there for each group?



Counties with a gain have higher median and variability, which is visualized in the box plot.

Both are unimodal with right skew, seen from examining the histogram.

Questions?

Credits

Examples adapted from OpenIntro Statistics (4th edition) by David Diez, Mine Cetinkaya-Rundel, and Christopher D Barr https://www.openintro.org/book/os/ protected under the Creative Commons License