Chapter 2: Summarizing data

- Examining numerical data
- Considering categorical data

Examining Numerical Data

- Visualizing Numerical Data
- Measures of Center
- Measures of Spread
- Outliers
- Transforming

visualizing numerical data

- scatterplots for paired data
- other visualizations for describing distributions of numerical variables

data

country	income per person ($,2011)	life expectancy (year, 2011)
Afghanistan	1660.74	60.4
Albania	10207.76	77.7
Algeria	12990.35	76.3
…	…	…
Zimbabwe	1667.138	52.4

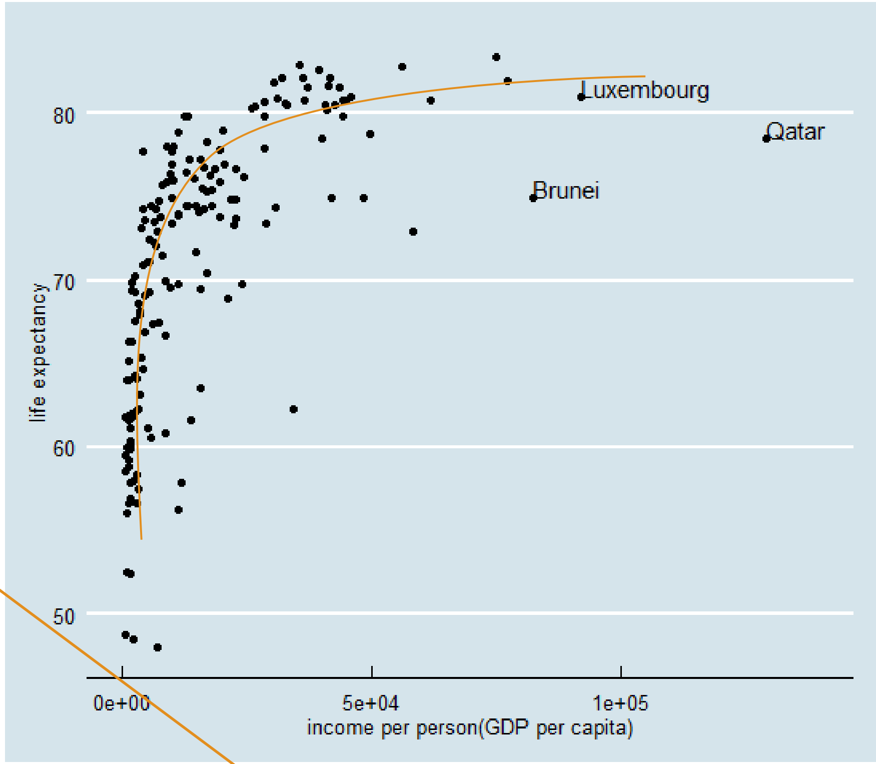
Source: ourworldindata.org, gapminder.org

scatterplots

A common tool for visualizing the *relationship* between two numerical variables is a scatter plot.

To identify the *explanatory* variable in a pair of variables, we identify which of the two is suspected of effecting the other and plan an appropriate analysis.

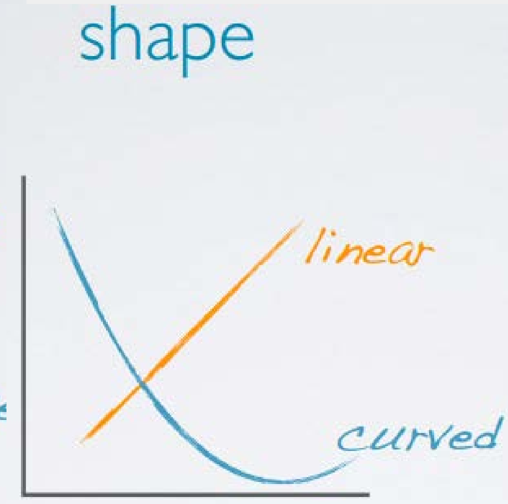
response

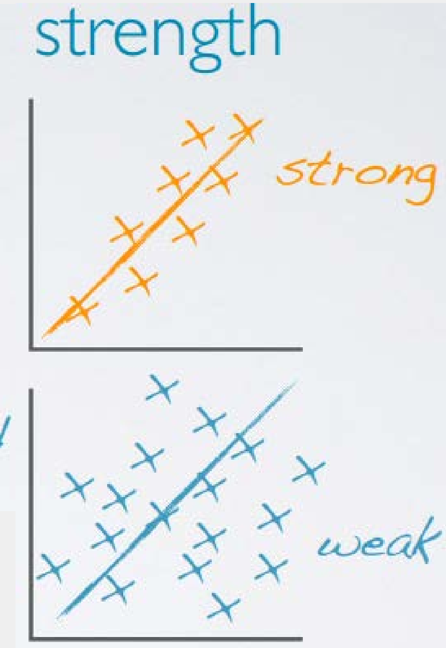


correlation
not causation

explanatory

evaluating the relationship





Positive: Is it increasing?

Negative: Or decreasing?

Is it **linear**?
Or **non-linear**?

Any potential **outliers**?

Strong indicated by little scatter?

Or **weak**, indicated by lots of scatter?

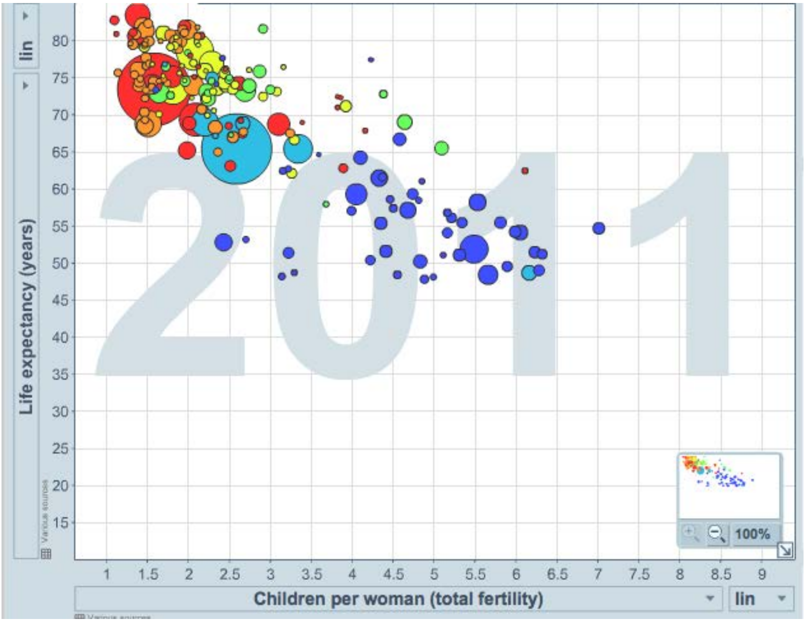
Practice

Do life expectancy and total fertility appear to be *associated* or *independent*?

They appear to be linearly and negatively associated: as fertility increases, life expectancy decreases.

Was the relationship the same throughout the years, or did it change?

The relationship changed over the years.

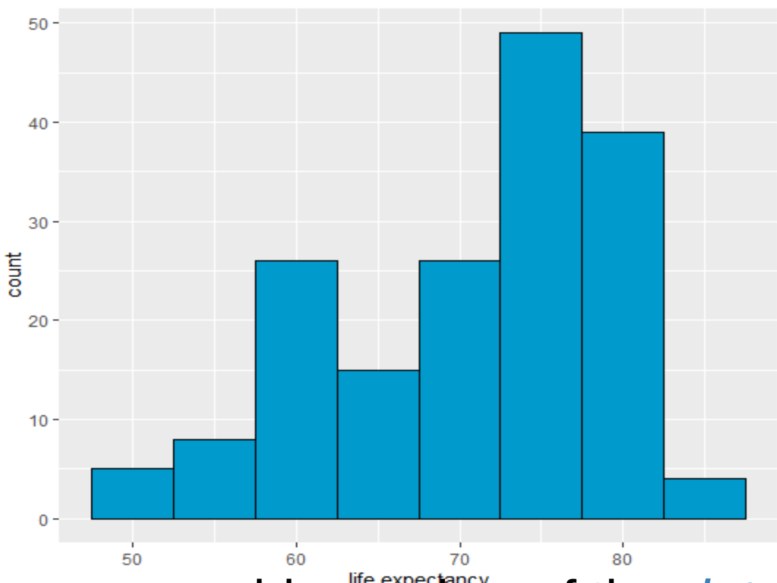


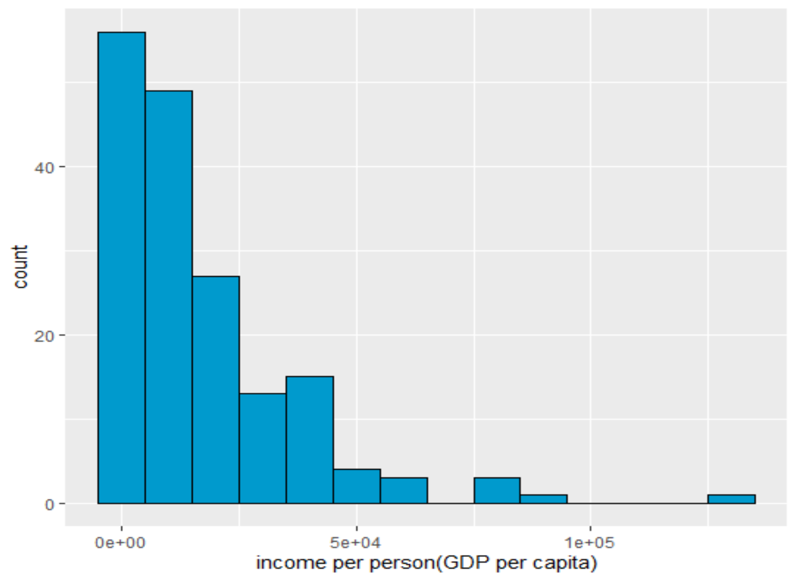
http://www.gapminder.org/world

Histogram

one good way of visualizing the distribution of a numerical variable

- data are binned into intervals
- heights of the bars represent the number of cases that fall into each interval.





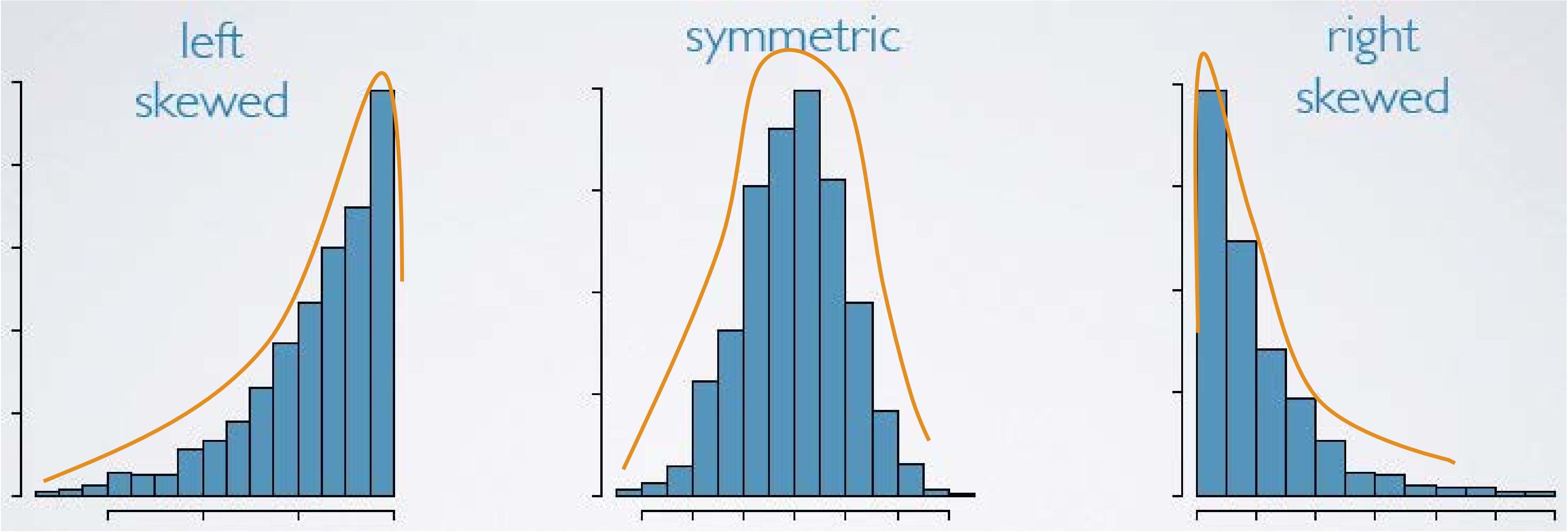
- provides a view of the *data density*.

  - higher bars represent where the data are relatively more common

- especially useful for describing the *shape* of the distribution.
- The chosen *bin width* can alter the story the histogram is telling.

Shape of a Distribution: Skewness

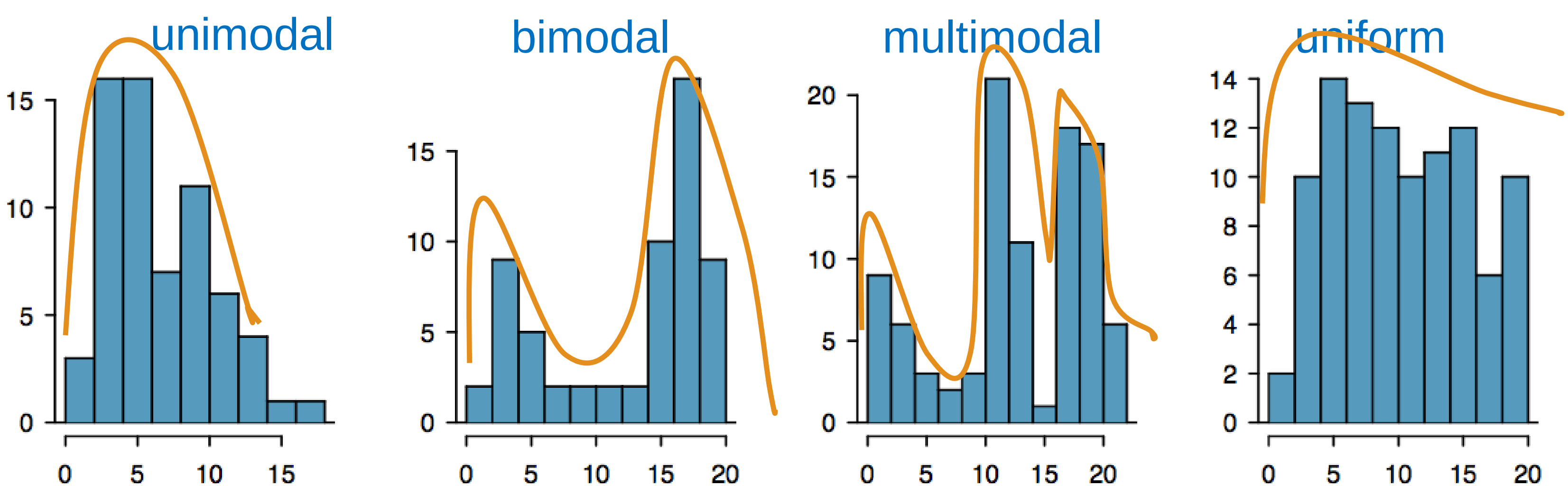
- Distribution *are skewed to the side of the long tail.*
- Is the histogram *right skewed*, *left skewed*, or *symmetric*?



Shape of a Distribution: Modality

Prominent peak determine modality

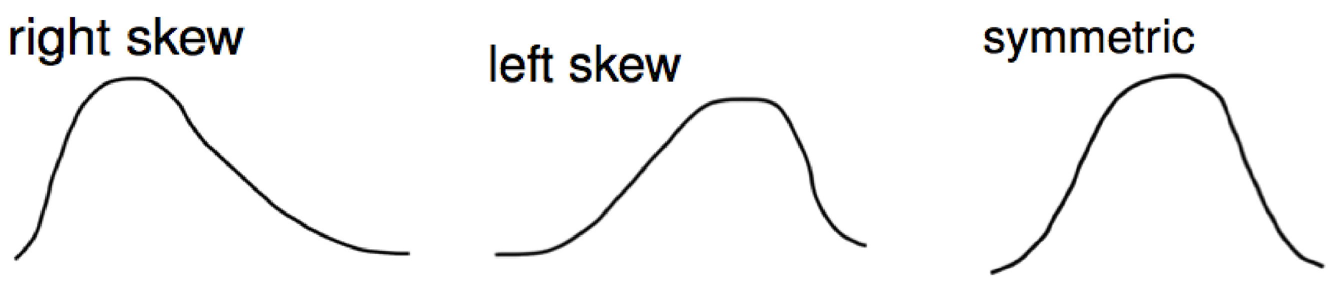
- A distribution might be *unimodal with one prominent peak*
- *Bimodal* with two prominent peaks
- *Uniform* with no prominent peaks
- *Multimodal*is what we call a distribution when it has more than two prominent peaks



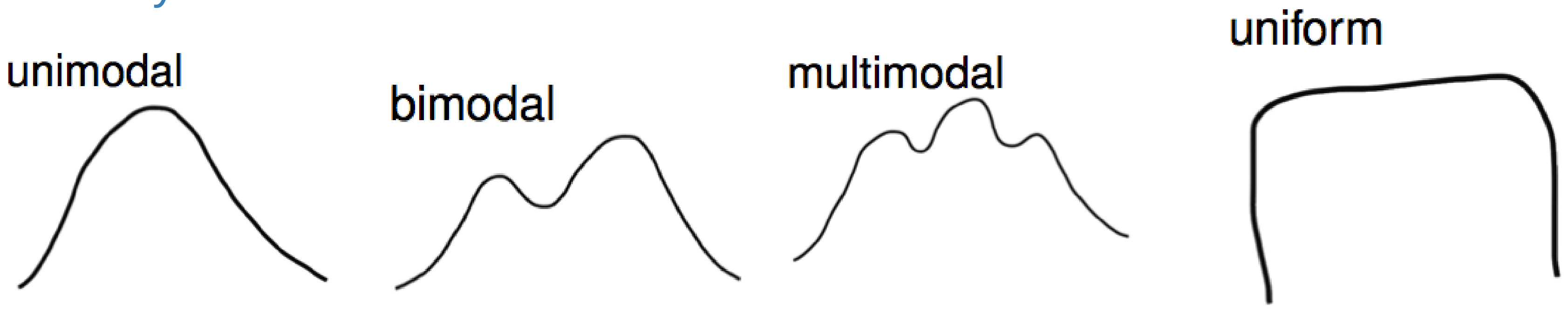
Note: In order to determine modality, step back and imagine a smooth curve over the histogram -- imagine that the bars are wooden blocks and you drop a limp spaghetti over them, the shape the spaghetti would take could be viewed as a smooth curve.

Commonly observed shapes of distributions

Skewness



Modality



Practice

Which of these variables do you expect to be uniformly distributed?

(a) weights of adult females

(b) salaries of a random sample of people from North Carolina

(c) house prices

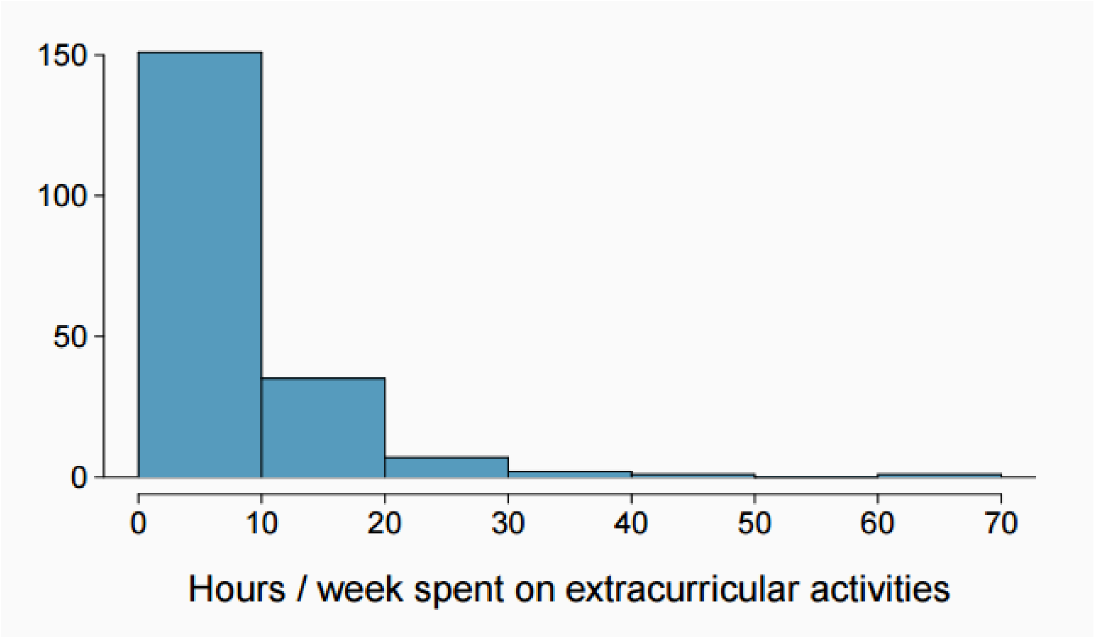
(d) birthdays of classmates (day of the month)

Answer : (d)

People are equally likely to be born at the beginning, middle, or the end of the month; hence we would expect the distribution of the birthdays to be uniform (no trend).

Extracurricular activities

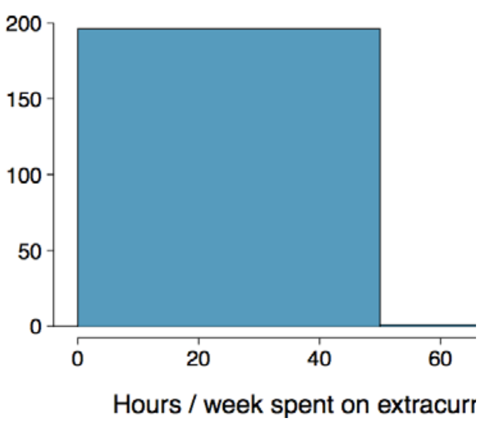
How would you describe the shape of the distribution of hours per week students spend on extracurricular activities?

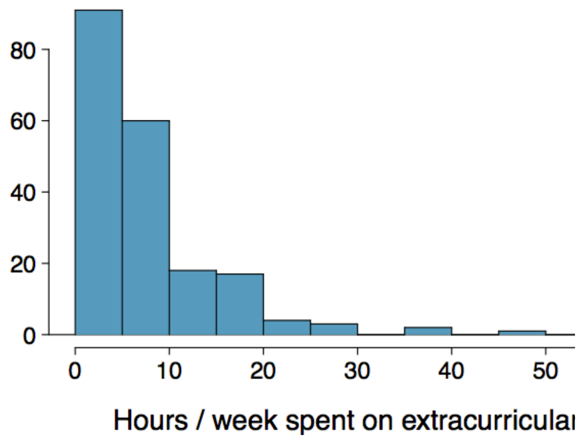


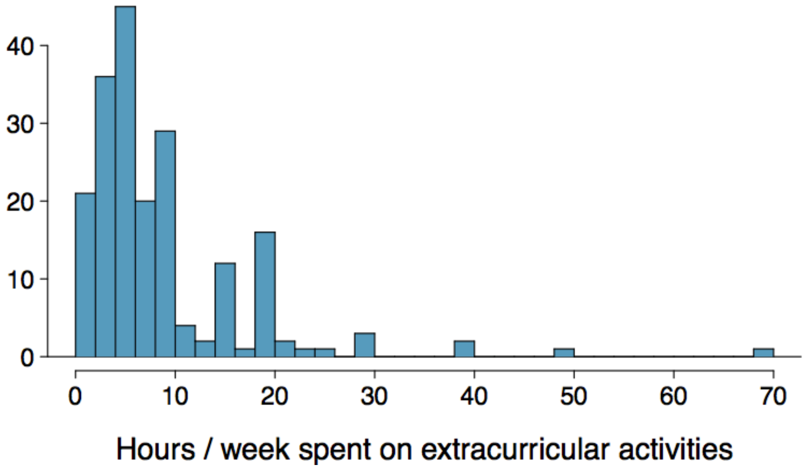
Unimodal and right skewed

Bin Width

Which one(s) of these histograms are useful? Which reveal too much about the data? Which hide too much?







The chosen bin width of a histogram can alter the story the histogram is telling

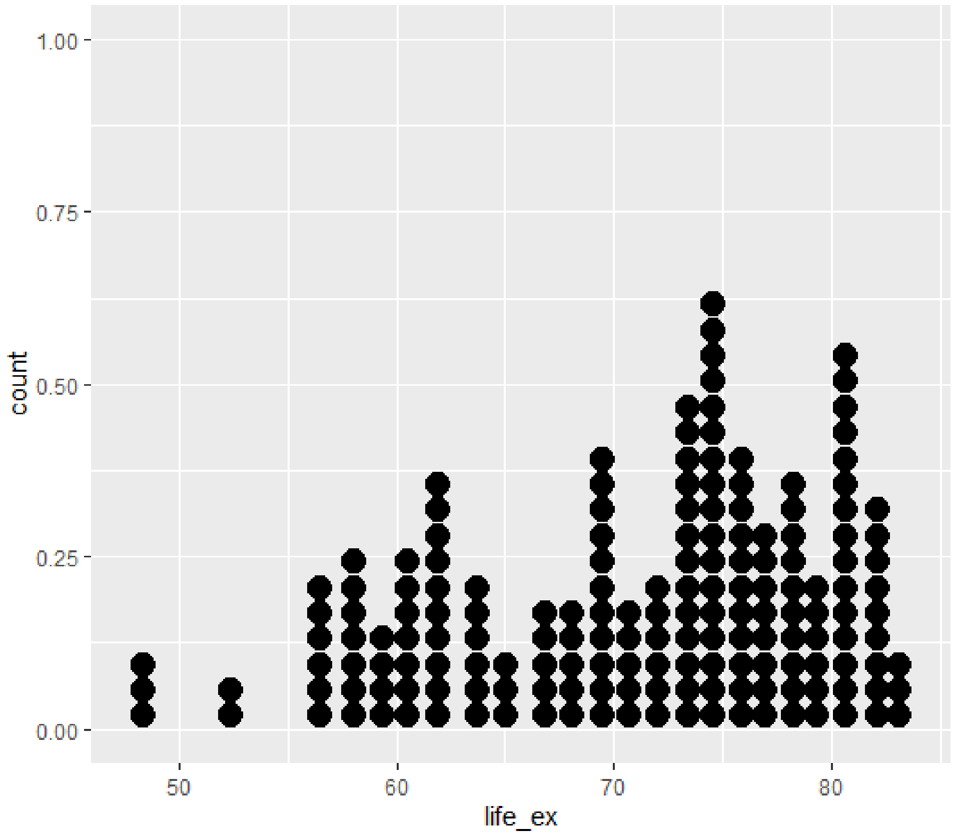
- too wide $\longrightarrow$ lose interesting details
- too narrow $\longrightarrow$ difficult to get an overall picture of the distribution.
- the ideal bin width depends on the data you're working with

You should try playing with it until you're satisfied with the visualization.

Dot Plots

Useful for especially when individual values are of interest

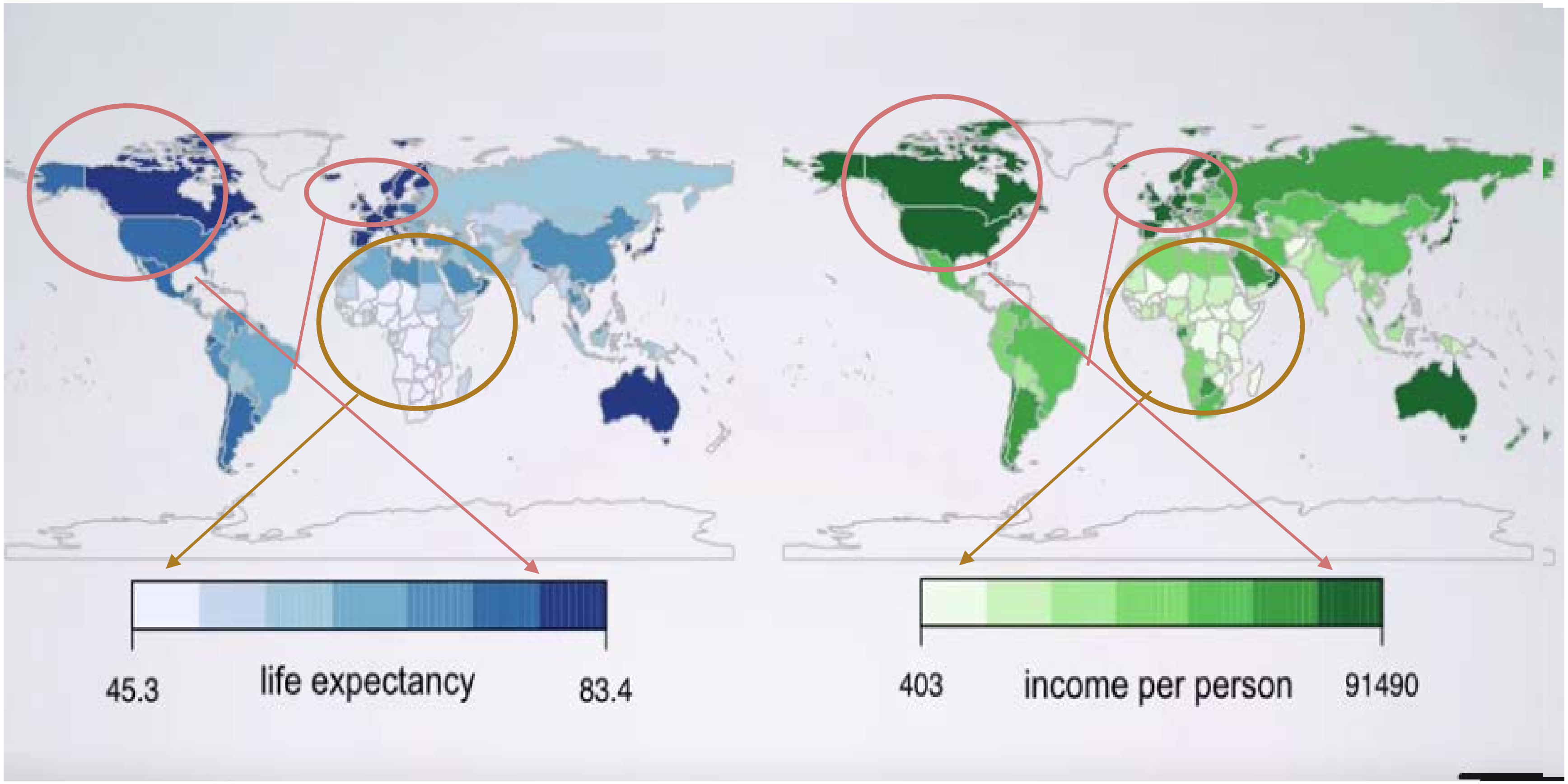
As the sample size increases, the dot plot may get too busy



Box plot *introduced later*

Intensity Maps

- Useful for highlighting the spatial distribution



measures of center

mean

arithmetic average

$\bar{x}$: sample mean

μ : population mean

median

midpoint of the distribution (50percentile)

observations: $x_1, x_2, \ldots, x_n$

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

sample statistic

point estimate

population parameter

mode

most frequent observation

Intuitively speaking, a numerical measure of center describes a “typical value” of the distribution.

Are you typical?

http://www.youtube.com/watch?v=4B2xOvKFFz4

How useful are centers alone for conveying the true characteristics of a distribution?

Example 1: Odd number of observations

9 student’s exam scores:

75, 69, 88, 93, 95, 54, 87, 88, 24

- Mean: (75 + 69 +88 + 93 + 95 + 54 + 87 + 88 + 24)/9 = 74.78
- Mode: 88 (The most frequent observed value, in this case it is 88)
- Median: 87

  - firstly sort the data in increasing order

  24, 64, 69, 75, 87, 88, 88, 93, 95

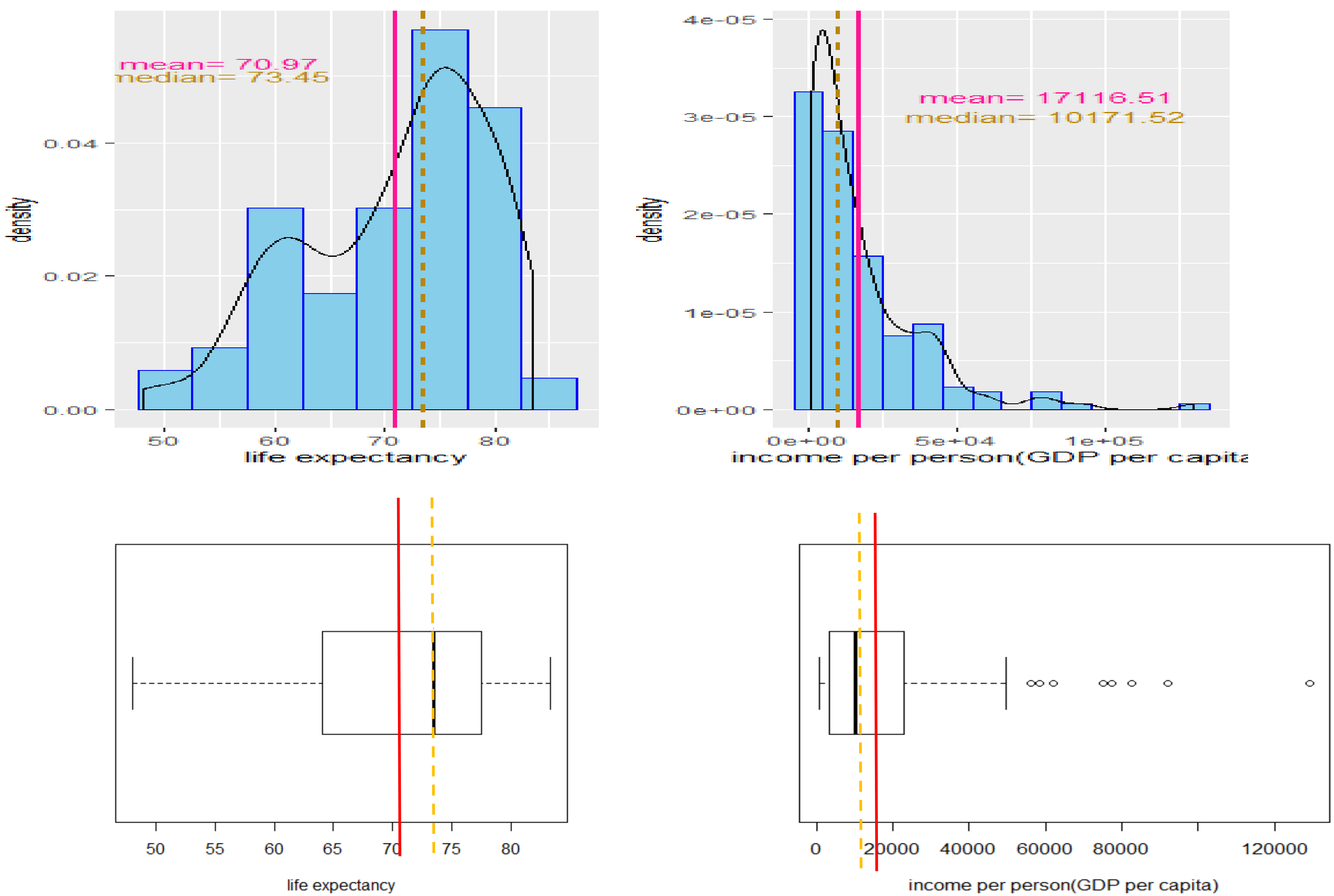
  - then we find the mid-point of the ordered data

Example 2: Even number of observations

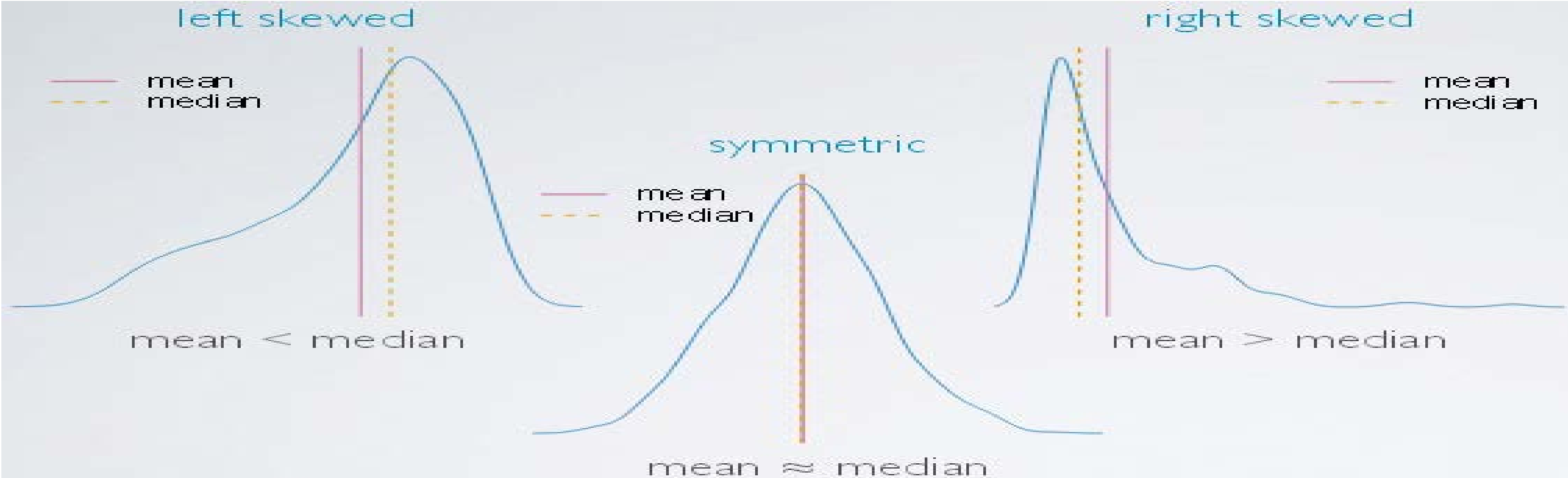
10 student’s exam scores:

24, 54, 69, 75, 87, 88, 88, 93, 95, 100

Median: (87+88)/2=87.5


mean= 70.97
median= 73.45
density
0.00
0.02
0.04
50
60
70
80
life expectancy
mean= 17116.51
median= 10171.52
density
0e+00
1e-05
2e-05
3e-05
4e-05
0e+00
5e+04
1e+05
income per person(GDP per capita
50
55
60
65
70
75
80
life expectancy
0
20000
40000
60000
80000
120000
income per person(GDP per capita)

Skewness vs. Measures of Center



left skewed

- the lower valued observations pull the mean to themselves
- the *mean is generally lower than the median*.

symmetric

- *the mean and the median are roughly equal* to each other in the center of the distribution

right skewed

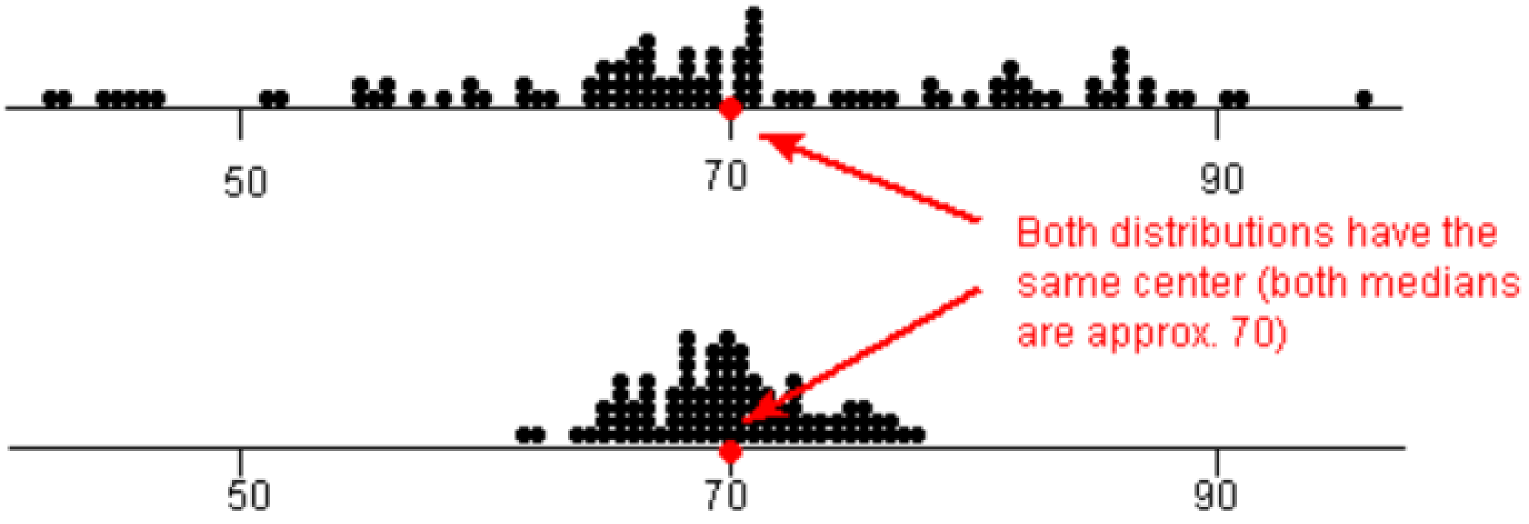
- the high valued observations pull the mean to themselves
- the *mean is generally larger than the median*.

measures of spread

Consider the following two distributions of exam scores.

Both distributions are centered at 70 but the distributions are quite different.

The first distribution has a *much larger variability* in scores compared to the second one.



measures of spread

three most commonly used measures of spread:

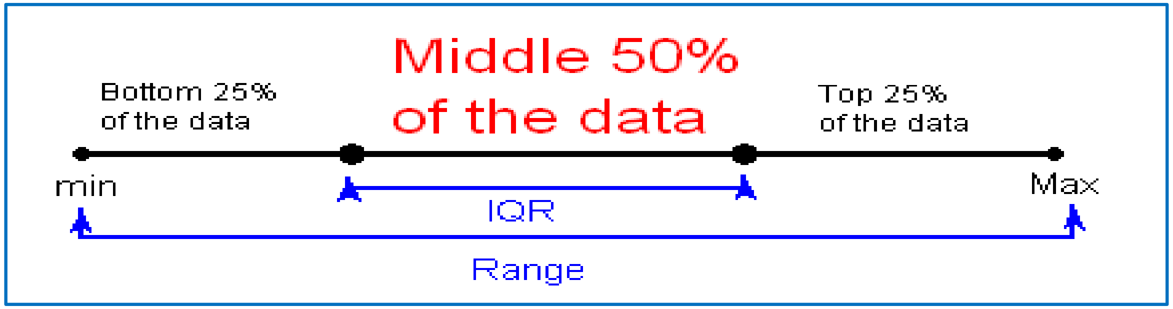
- Range : max- min
- Inter-quartile range (IQR)
- Standard Deviation

The **range** covered by the data is the most intuitive measure of variability. The range is exactly the distance between the smallest data point (min) and the largest one (max).

measures of spread : IQR

While the range quantifies the variability by looking at the range covered by ALL the data,

the *Inter-Quartile Range* or **IQR** measures the variability of a distribution by giving us the range covered by the MIDDLE 50% of the data.



IQR = Q3 – Q1

Q3 = 3rd Quartile = 75th Percentile : three quarters (75%) of the data points fall below it,

Q1 = 1st Quartile = 25th Percentile: one quarter (25%) of the data points fall below it

*The **IQR** is generally used as a measure of spread of a distribution when the **median** is used as a measure of center*

measures of spread : standard deviation

- roughly *the average deviation around the mean*

$$s = \sqrt{s^2}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- *standard deviation* = the square root of the variance, has the same units as the data.
- *Variance* = roughly the *average squared deviation* from the mean.

sample variance	population variance
s^2	σ^2
sample standard deviation	population standard deviation
s	σ^2

many notations for the standard deviation: SD, s, Sd, StDev.

It is appropriate to use the ***standard deviation*** *as a measure of spread with the* ***mean*** *as the measure of center.*

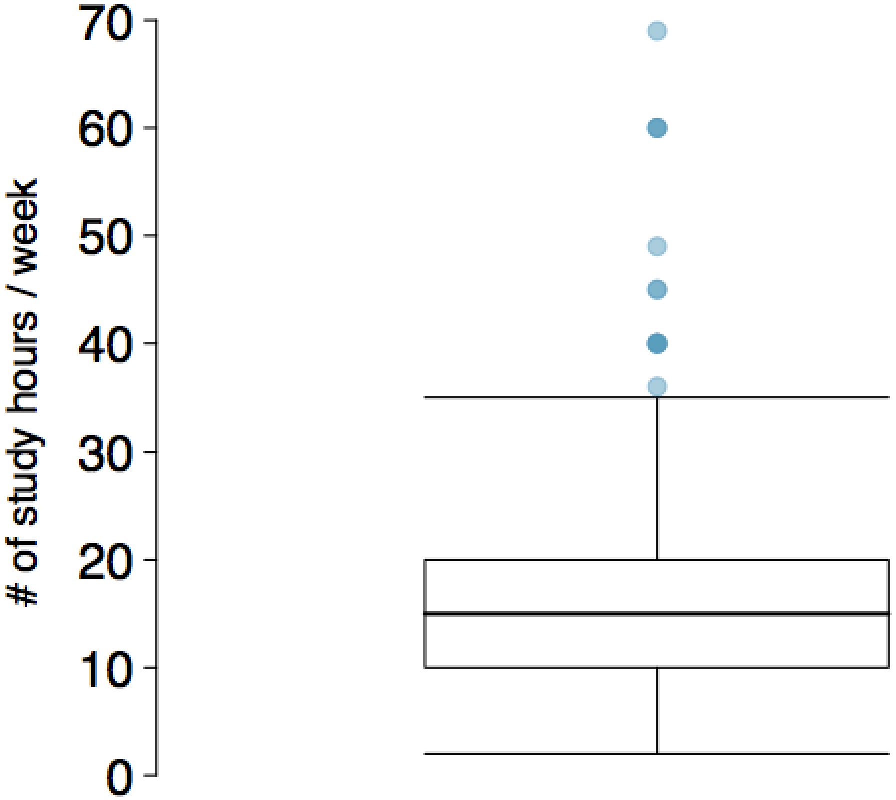
variance

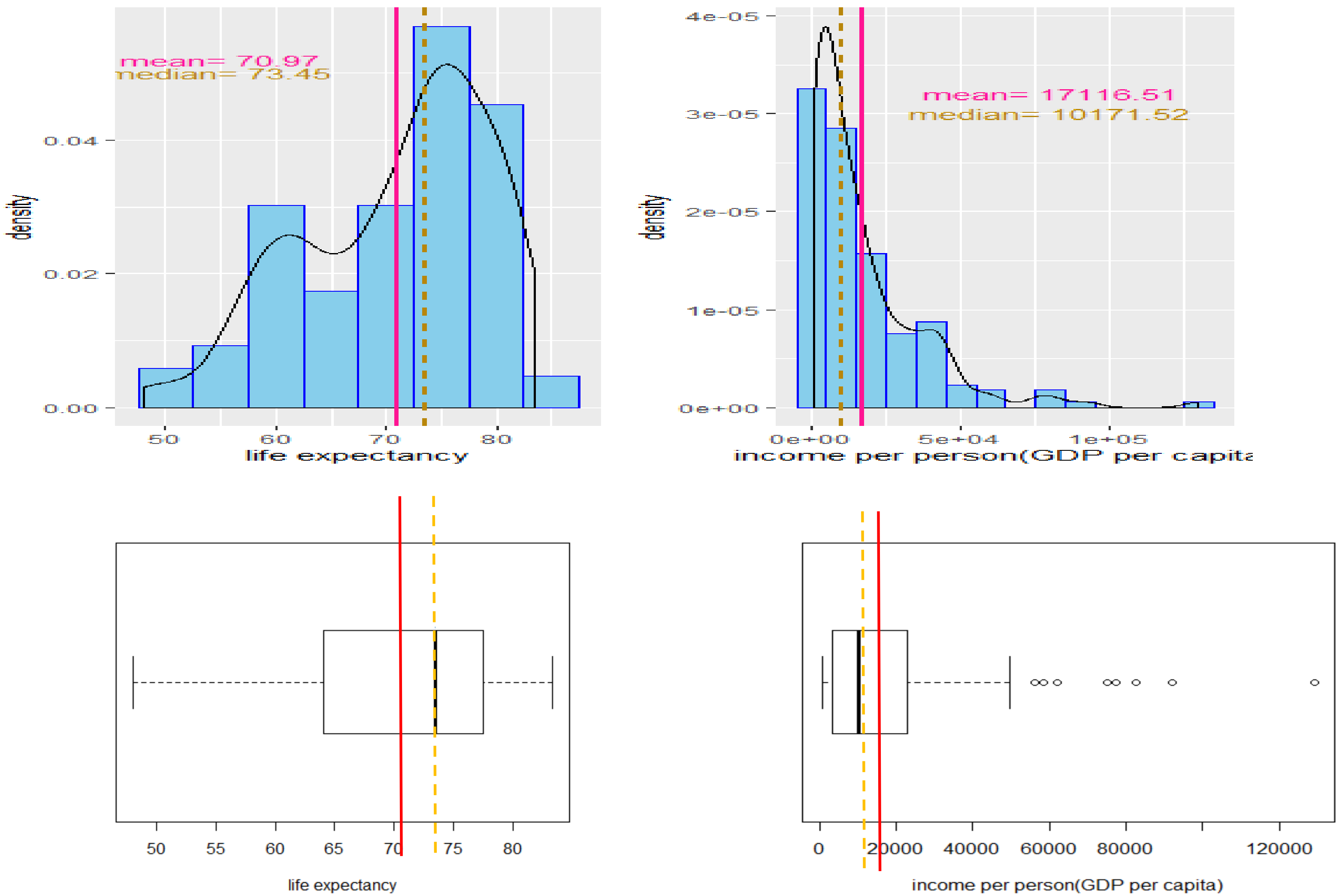
Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

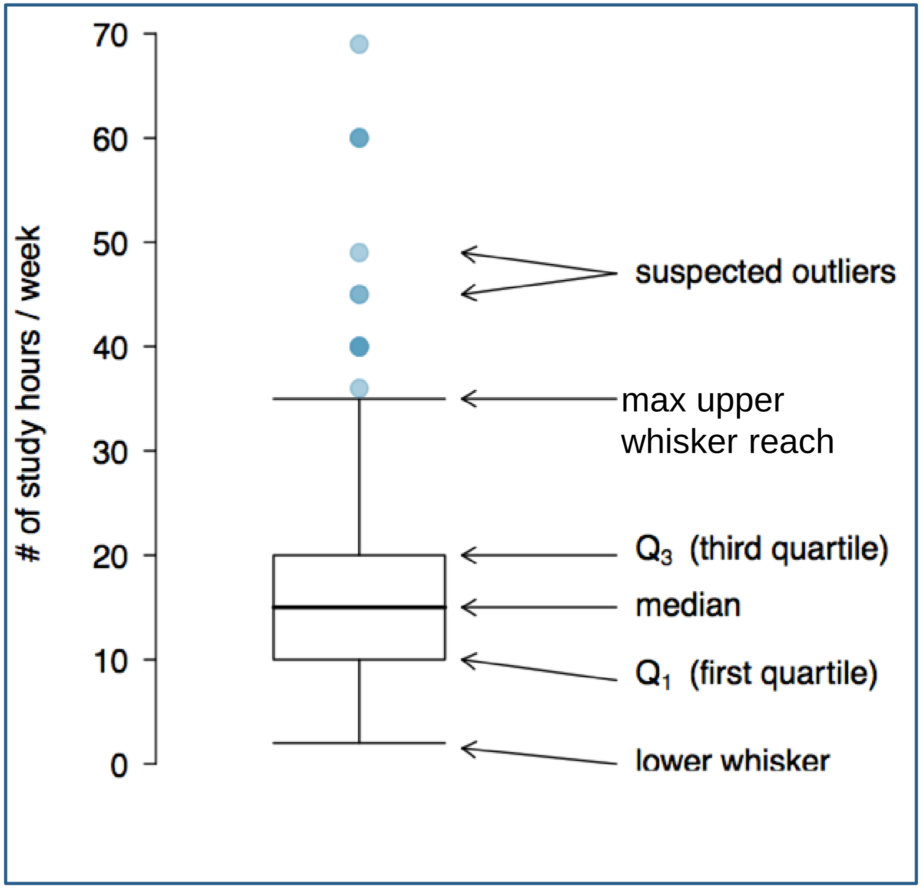
Box Plot

The box in a *box plot* represents the middle 50% of the data, and the thick line in the box is the median.




mean= 70.97
median= 73.45
density
0.00
0.02
0.04
50
60
70
80
life expectancy
mean= 17116.51
median= 10171.52
density
0e+00
1e-05
2e-05
3e-05
4e-05
0e+00
5e+04
1e+05
income per person(GDP per capita
50
55
60
65
70
75
80
life expectancy
0
20000
40000
60000
80000
120000
income per person(GDP per capita)

Anatomy of a Box Plot



Whiskers of a box plot can extend

up to 1.5 x IQR away from the quartiles.

- max upper whisker reach = Q3 + 1.5 x IQR
- max lower whisker reach = Q1 - 1.5 x IQR

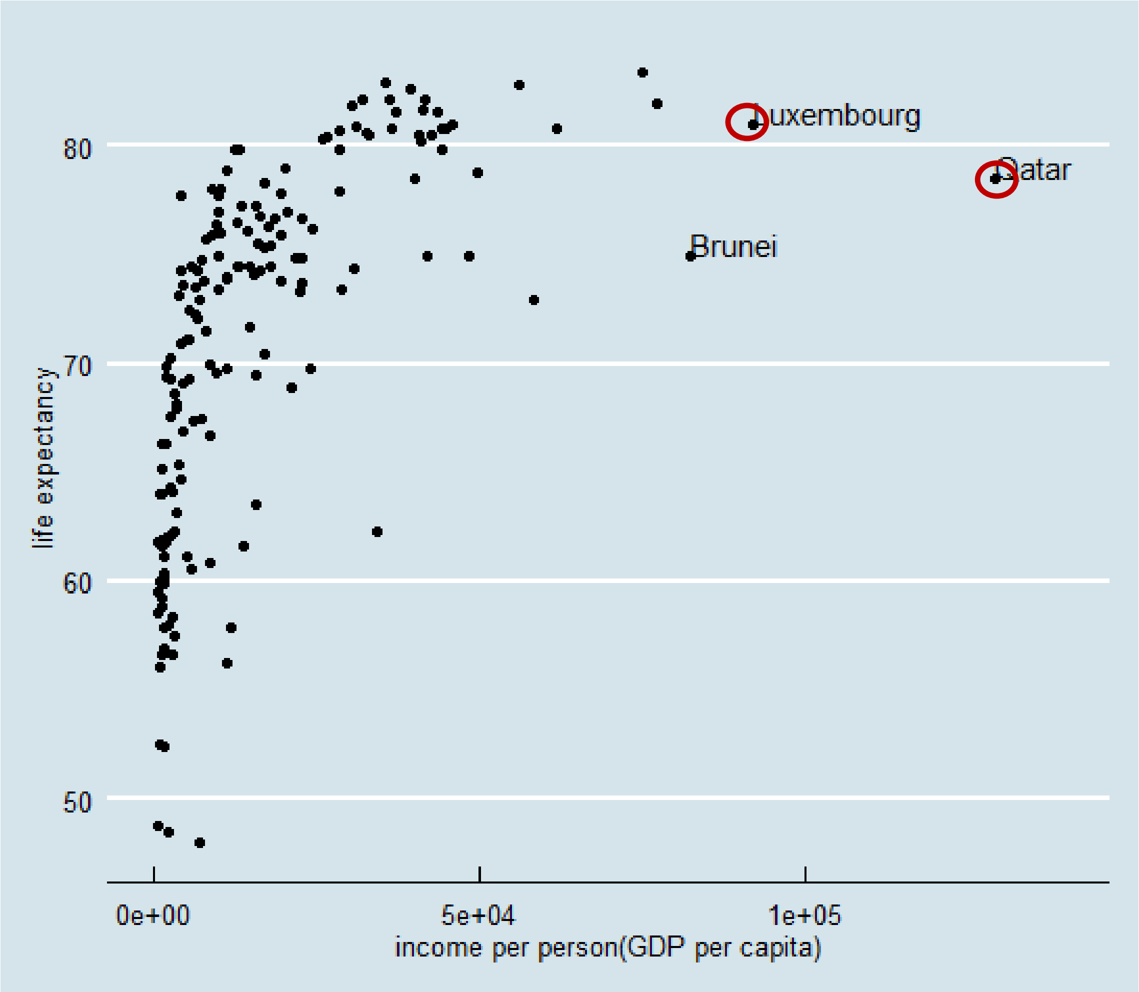
IQR: 20 - 10 = 10

max upper whisker reach = 20 + 1.5 x 10 = 35

max lower whisker reach = 10 - 1.5 x 10 = -5

1.5(IQR) criterion for outliers :A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

outliers



Luxembourg
Qatar
Brunei
80
70
60
50
life expectancy
0e+00
5e+04
1e+05
income per person(GDP per capita)

Outliers

Why is it important to look for outliers?

- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

Understanding outliers

Why is it important to identify possible outliers, and how should they be dealt with?

The answers to these questions depend on the reasons for the outlying values.

- by *essentially the same process* as the rest of the data, expected to *eventually occur again*

  such an outlier indicates something important and interesting about the process you’re investigating, and it *should be kept* in the data.

- under fundamentally *different* conditions from the rest of the data

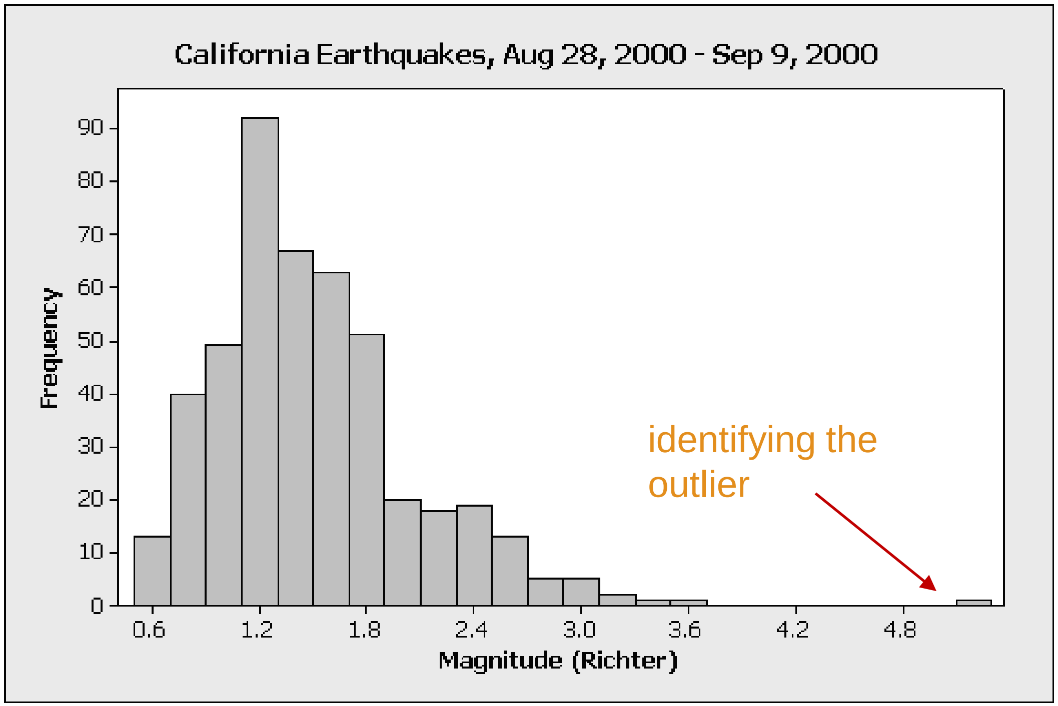
  such an outlier *can be removed* from the data if your goal is to investigate only the process that produced the rest of the data.

- by a *mistake* in the data (like a typo, or a measuring error

  should be corrected if possible or else removed

Example of types of outliers

The following histogram displays the magnitude of 460 earthquakes in California, occurring in the year 2000, between August 28 and September 9

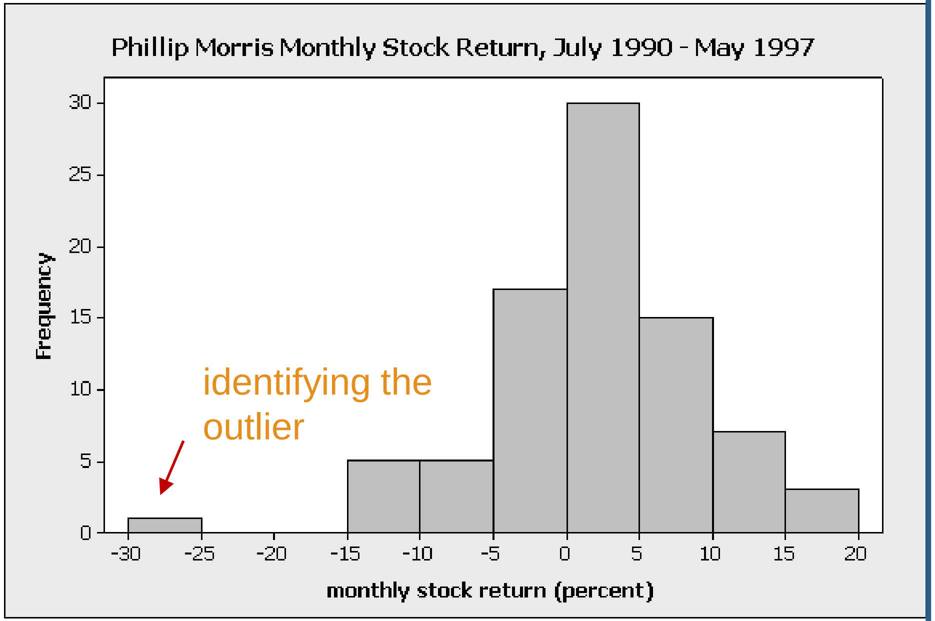


Understanding the outlier**:** In this case, the outlier represents a much stronger earthquake, which is relatively rarer than the smaller quakes that happen more frequently in California.

How to handle the outlier**:** For many purposes, the relatively severe quakes represented by the outlier might be the most important (because, for instance, that sort of quake has the potential to do more damage to people and infrastructure). So, it could *be important to keep this outlier in the data.*

Example of types of outliers

The following histogram displays the monthly percent return on the stock of Phillip Morris (a large tobacco company) from July 1990 to May 1997:



Understanding the outlier: In the early 1990s, there were highly-publicized federal hearings being conducted regarding the addictiveness of smoking, and there was growing public sentiment against the tobacco companies. The unusually low monthly value in the Phillip Morris dataset was due to public pressure against smoking, which negatively affected the company's stock for that particular month.

How to handle the outlier: the outlier was due to unusual conditions during one particular month that *aren't expected to be repeated*, and that were *fundamentally different from the conditions* that produced the values in all the other months. So in this case, it would be *reasonable to remove the outlier*, if we wanted to characterize the "typical" monthly return on Phillip Morris stock.

Extreme observations

Here are two datasets:

Data set A $\rightarrow$ 64 65 66 68 70 71 73

Data set B $\rightarrow$ 64 65 66 68 70 71 730

- dataset A: mean = 68.1, median = 68.
- dataset B: mean = 162, median = 68

notice that all of the observations except the last one are close together.

The observation 730 is very large, and is certainly an outlier.

Comparing the Mean, SD and the Median, IQR

mean, SD		median, IQR
the *actual values* of the data points play an important role. *very sensitive to outliers*	the most common **measures of center and spread**	the *order* of the data is the key. *resistant (or robust) to outliers*

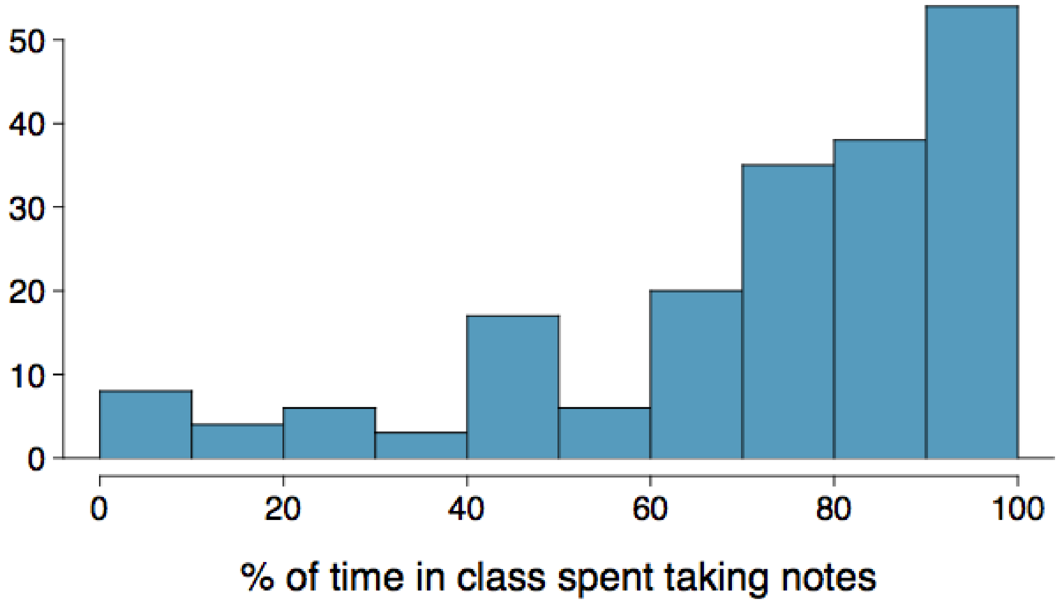
Median and IQR are more robust to skewness and outliers than mean and SD.

Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

Practice

Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



Answer: (c)

median: 80%

mean: 76%

(a) mean > median

(b) mean ~ median

(c) mean < median

(d) impossible to tell

transforming the data

- define transformations
- review when it might be useful/necessary to transform data

transformations

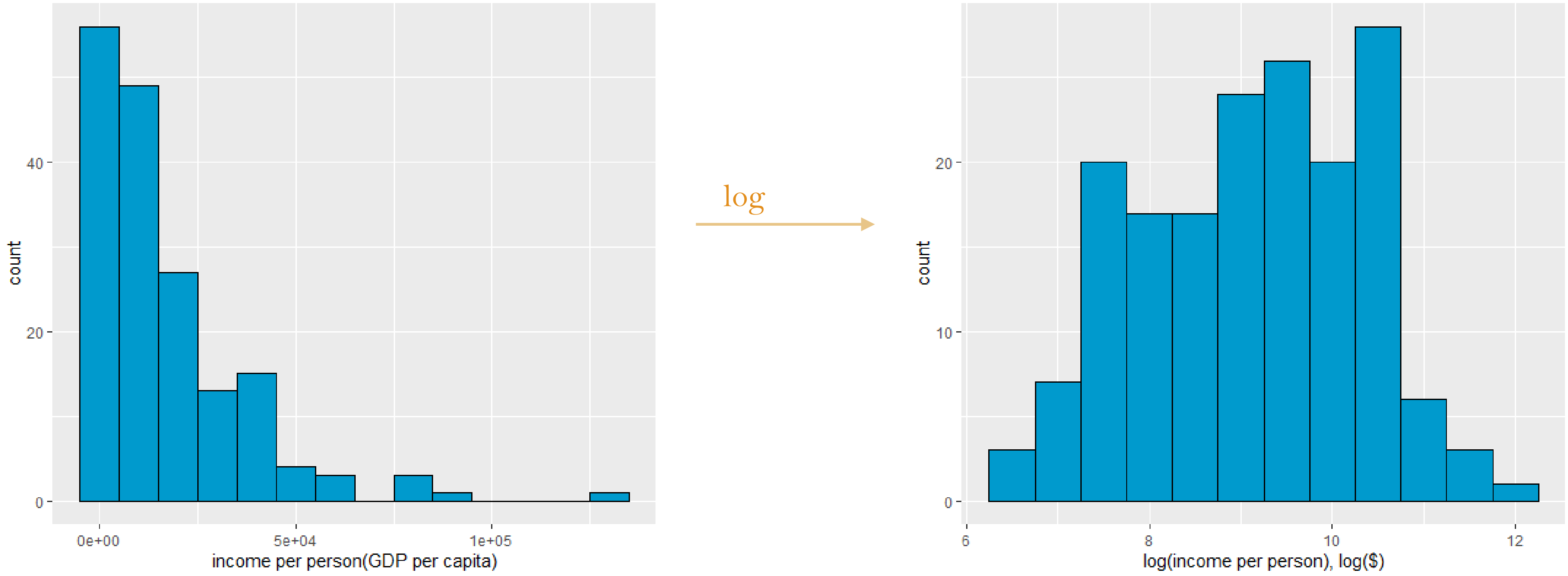
- a transformation is a rescaling of the data using a function
- when data are very strongly skewed, we sometimes transform them so they are easier to model

(natural) log transformations

Natural log transformation is often applied when:

- Much of the data cluster near zero

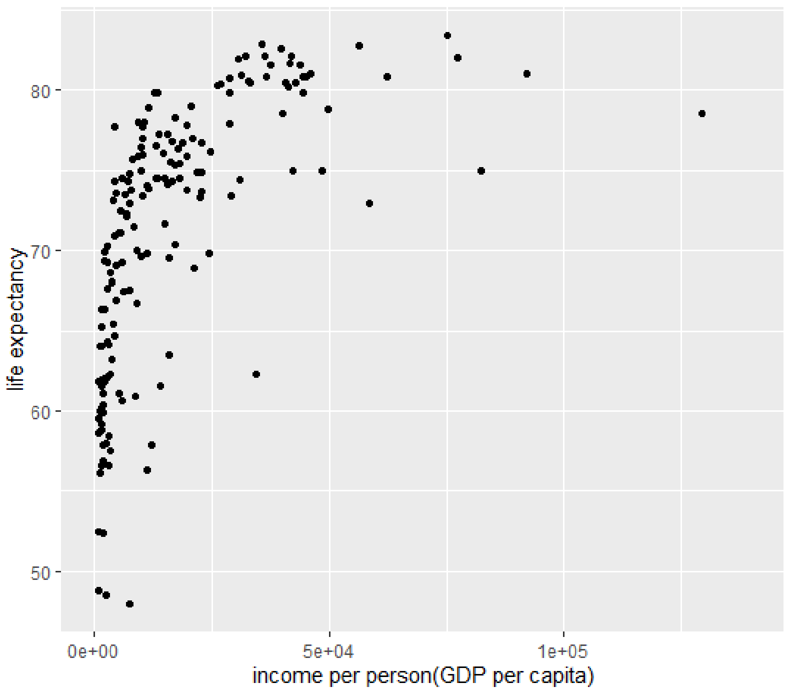
  (relative to larger values in the data set).
- and, all observations are positive.



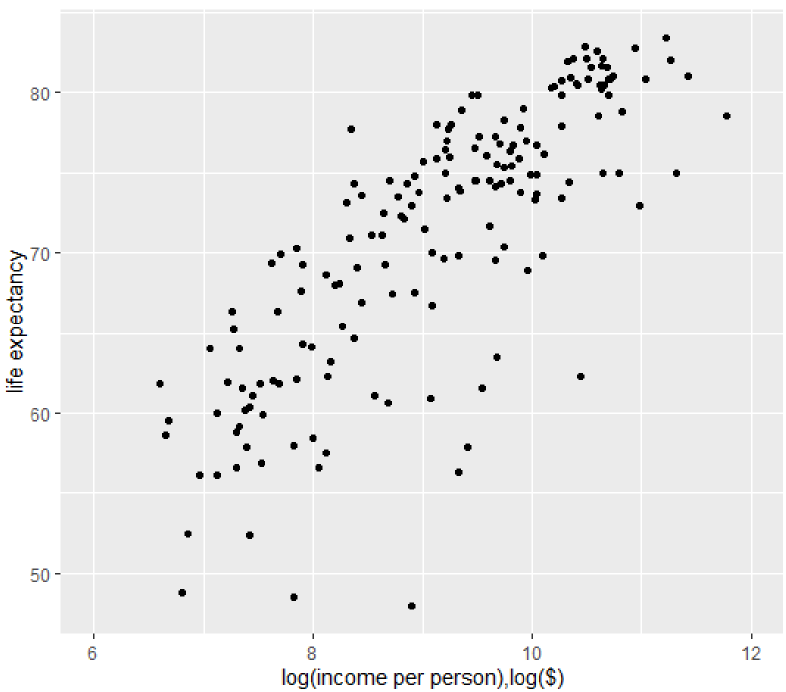
log transformation

to make the relationship between the variables more linear

and hence easier to model with simple methods



log



other transformations

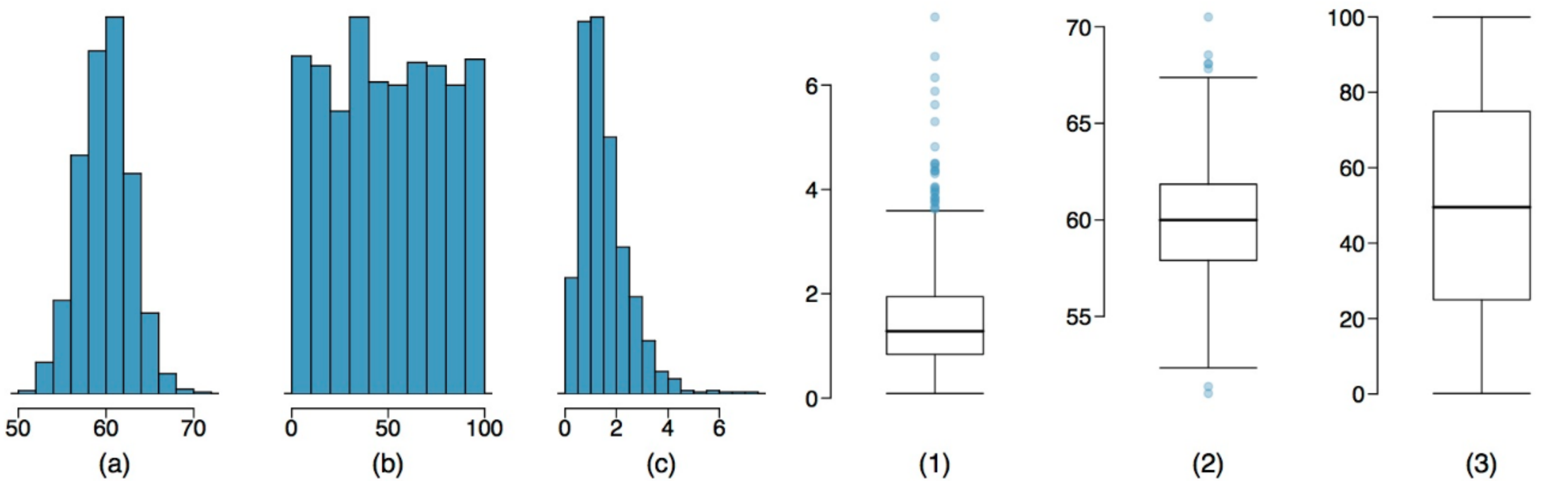
goals of transformations

- to see the data structure differently
- to reduce skew assist in modeling
- to straighten a nonlinear relationship in a scatterplot

Homework Week 3

Problem 1: Find the recent news article that refer to some type of measure of center and spread of the data. Explain which measures were referred.

Problem 2: Determine which histogram matches which box plot



Homework

Problem 3: For each of the following situations, state whether you expect the distribution to be symmetric, left-skewed, or right-skewed. Explain with a concise way (1-2 sentences) to teach someone how to determine the expected distribution of any variable.

a. Heights of a sample of 100 women

b. Family income in the United States

c. Speeds of cars on a road where a visible patrol car is using radar to detect speeders