Unit 1: Introduction to data
Lecture 2: Exploratory data analysis

Statistics 104

Mine Çetinkaya-Rundel

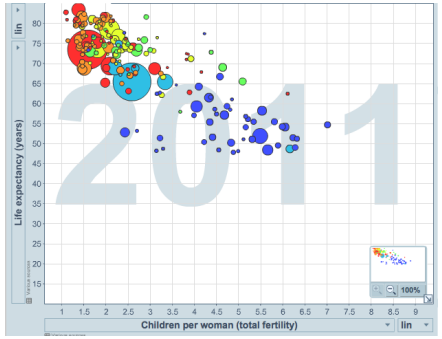
May 17, 2013

Scatterplot

Scatterplots are useful for visualizing the relationship between two numerical variables.

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

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

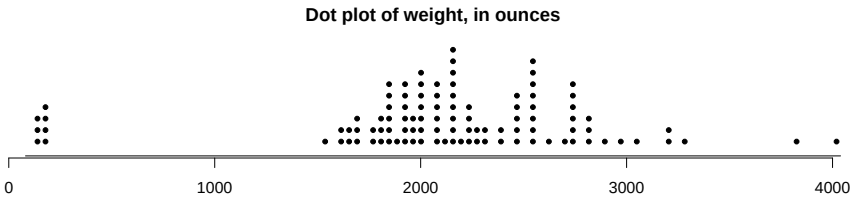


http://www.gapminder.org/world

Visualizing numerical variables

- *Dot plot*: Useful when individual values are of interest.
- *Histogram*: Provides a view of the *data density*, and are especially convenient for describing the shape of the data distribution.
- *Box plot*: Especially useful for displaying the median, quartiles, unusual observations, as well as the IQR.
- *Intensity map*: Useful for displaying the spatial distribution.

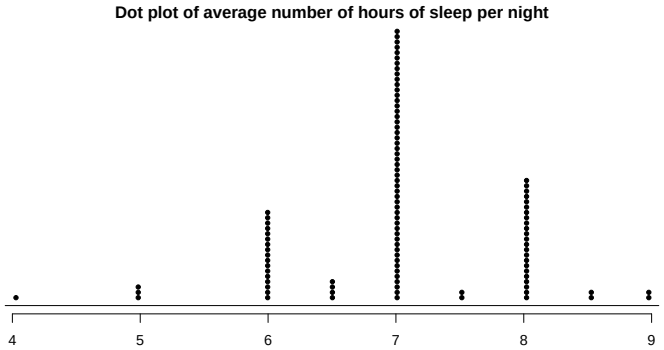
Why visualize?



Do you see anything out of the ordinary?

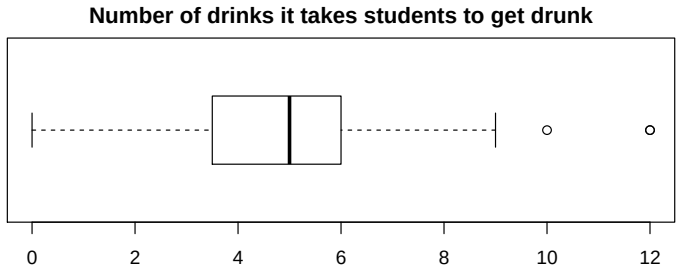
Why visualize?

What type of variable is average number of hours of sleep per night? Is this reflected in the dot plot below? If not, what might be the reason?



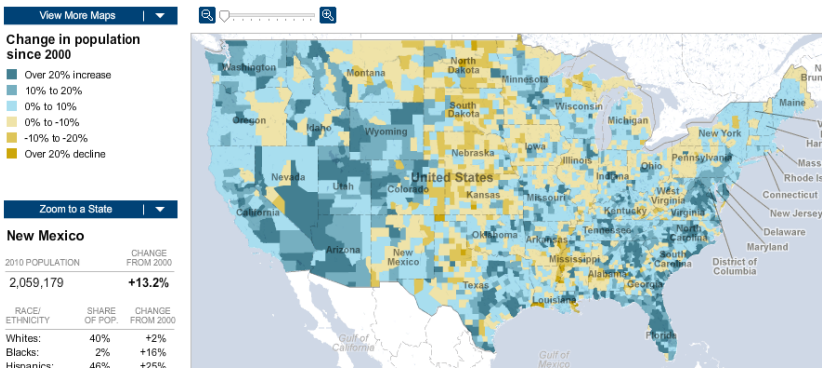
Why visualize?

What does a response of 0 mean in this distribution?



Why visualize?

What patterns are apparent in the change in population between 2000 and 2010?



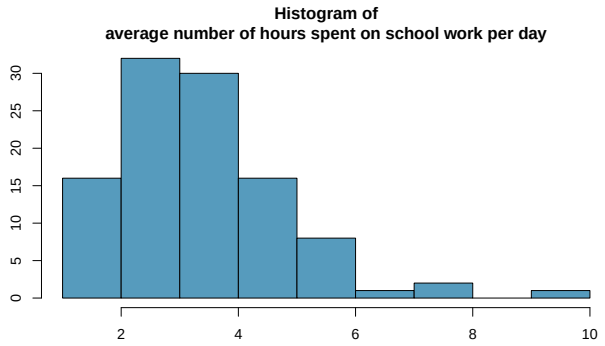
Describing distributions of numerical variables

When describing distributions of numerical variables always mention

- *Shape*: skewness, modality
- *Center*: an estimate of a *typical* observation in the distribution (mean, median, mode, etc.)
- *Spread*: measure of variability in the distribution (SD, IQR, range, etc.)
- *Unusual observations*: observations that stand out from the rest of the data that may be suspected outliers

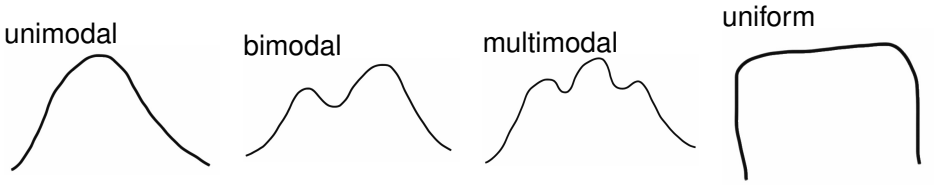
Shape

How would you describe the shape of this distribution?

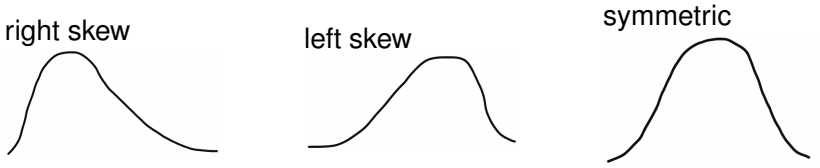


Commonly observed shapes of distributions

- modality



- skewness



Poll

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)

Application exercise: Shapes of distributions

Determine and sketch the expected distributions of the following variables.

- number of piercings
- scores on an exam
- IQ scores

Come up with a concise way (1-2 sentences) to teach someone how to determine the expected distribution of any variable.

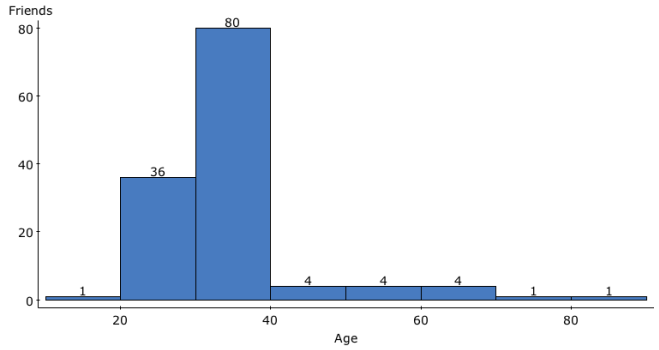
Measures of center

- Mean: arithmetic average
  - *Sample mean, $\bar{x}$*: Arithmetic average of values in sample.
  - *Population mean, μ*: Computed the same way but it is often not possible to calculate μ since population data is rarely available.
- Median: midpoint of the distribution, 50th percentile
- Mode: most frequent observation

The *sample statistics* are *point estimates* of the *population parameters*. These estimates may not be perfect, but if the sample is good (representative of the population)they are usually good guesses.

Ages of my FB friends

Can you guess my age based on data on the ages of my Facebook friends?



http://www.statcrunch.com/frienddata

Are you typical?

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

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

Variance

Variance, s^2

Roughly the average squared deviation from the mean

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

Given that the average number of hours students sleep per night is 7.029, the variance of amount of sleep students get per night can be calculated as:

$$s^2 = \frac{(7.5 - 7.029)^2 + (7 - 7.029)^2 + \cdots + (8 - 7.029)^2}{106 - 1} = 0.72$$

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.

Standard deviation

Standard deviation, s

Square root of the variance, has the same units as the data

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

The variance of amount of sleep students get per night can be calculated as:

$$s = \sqrt{0.72} = 0.85 \text{ hours}$$

Student on average sleep 7.029 hours, give or take 0.85 hours.

Range and IQR

Range

Range of the entire data.

$$range = max - min$$

IQR

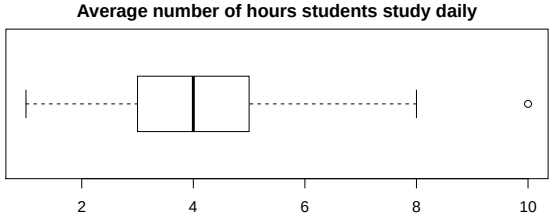
Range of the middle 50% of the data.

$$IQR = Q3 - Q1$$

Is the range or the IQR more robust to outliers?

Poll

Which of the following is *false* about the distribution of average number of hours students study daily?



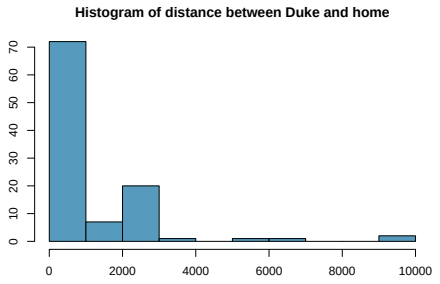

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   3.000   4.000   3.821   5.000  10.000
```

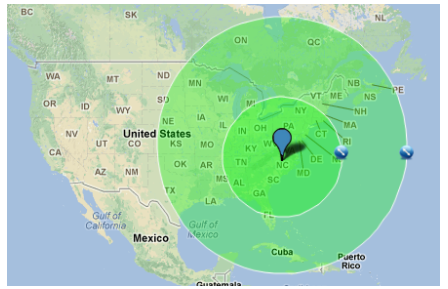
(a) There are no students who don't study at all.
(b) 75% of the students study more than 5 hours daily, on average.
(c) 25% of the students study less than 3 hours, on average.
(d) IQR is 2 hours.

Typical observation

How far is the *typical* student's home from Duke?

mean = 1250 miles median = 600 miles





http://www.freemaptools.com/radius-around-point.htm

Robust statistics

Since the median and IQR are more robust to skewness and outliers than mean and SD:

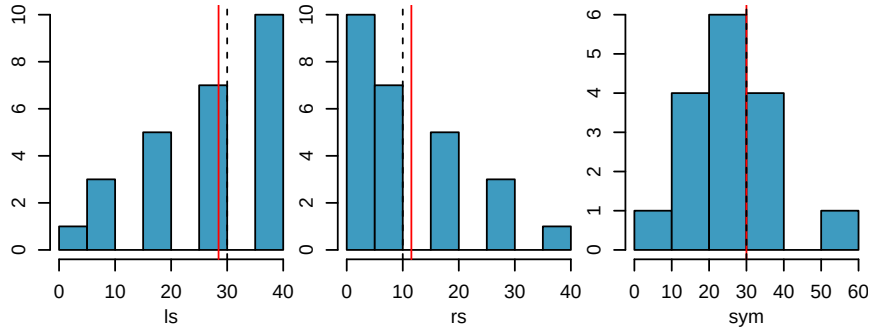
- skewed → median and IQR
- symmetric → mean and SD

If you were searching for a car, and you are price conscious, would you be more interested in the mean or median vehicle price when considering a car?

Mean vs. median

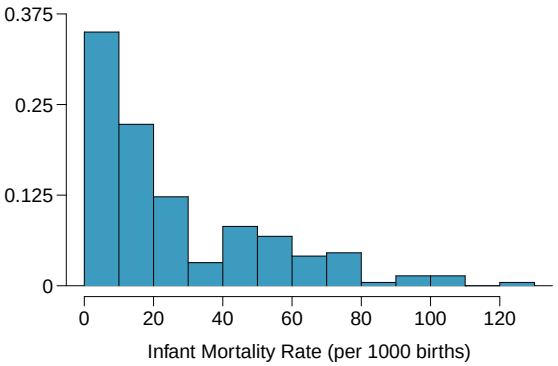
- If the distribution is symmetric, center is the mean
  - Symmetric: mean is roughly equal to the median
- If the distribution is skewed or has outliers center is the median
  - Right-skewed: mean is likely greater than the median
  - Left-skewed: mean is likely less than the median

red solid - mean, black dashed - median



Poll

The infant mortality rate is defined as the number of infant deaths per 1,000 live births. The relative frequency histogram below shows the distribution of estimated infant death rates in 2012 for 222 countries. Estimate Q3.



(a) between 0 and 10
(b) between 10 and 20
(c) between 20 and 30
(d) between 30 and 40
(e) between 40 and 50

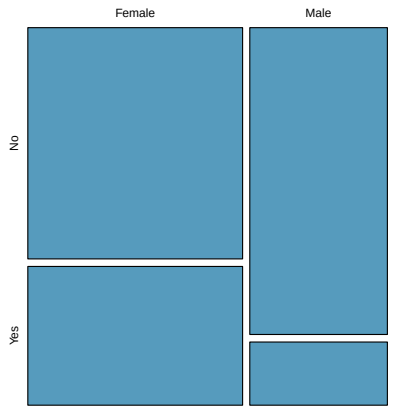
Contingency table and mosaic plot

Is there a relationship between gender and whether the student is looking for a spouse in college?

	No	Yes
Female	40	24
Male	34	7

% Females looking for a spouse:
24 / (40 + 24) = 0.375

% Males looking for a spouse:
7 / (34 + 7) = 0.17



Who survived the Titanic?

- On April 14, 1912, four days after it set sail from the port of Southampton England, the Titanic struck an iceberg in the North Atlantic.
- The ship sank and 1490 of the 2201 passengers perished.
- We'll focus on the third class passengers and the crew:

Who had a better chance of surviving, third class passengers or the crew?

	Survived	Died	Total	Survival rate
Third class	178	528	706	
Crew	212	673	885	
Total	390	1201	1591	

Who survived the Titanic? (cont.)

Or did they? Let's break the data down by gender:

Males	Survived	Died	Total	Survival rate
Third class	88	422	510	17%
Crew	192	670	862	*22%*
Total	280	1092	1372	

Females	Survived	Died	Total	Survival rate
Third class	90	106	196	46%
Crew	20	3	23	*87%*
Total	110	109	219	

Contrary to the previous slide, the crew had a better chance of surviving than third class passengers for both women and men. How can this be?

Who survived the Titanic? (cont.)

- Tthere are more men in the crew group than in the third class group and the reverse is true for women (more women in the third class group than in the crew).
- Since men were less likely to survive than women ("women and children first") and the crew was mostly males that gave the impression that the crew had a worse survival rate.
- This phenomenon is called *Simpson's Paradox*: An association, or a comparison, that holds when we compare two groups can disappear or even be reversed when the original groups are broken down into smaller groups according to some other feature (a confounding/lurking variable).

Side-by-side box plot

How does the number of the average number of times students go out per week vary by involvement? Do the two variables appear to be associated or independent?