Announcements

- HW1 due next Monday night
- Lab 1 write-up due Monday night as well
  - Make sure it includes:
    - Answers/requested plots to Exercise questions
    - Answers/plots to “On Your Own” questions
  - You can always resubmit it if you left something out and have submitted already
- Monday is another In-class Lab!
  - Working with datasets and plotting
- Read first half of Appendix A: Basic Probability for next Wednesday

Warm Up

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the least effective?

A) Simple random sampling

B) Cluster sampling

C) Stratified sampling

Warm Up

A city council has requested a household survey be conducted in a suburban area of their city. The area is broken into many distinct and unique neighborhoods, some including large homes, some with only apartments. Which approach would likely be the least effective?

A) Simple random sampling

B) Cluster sampling

C) Stratified sampling

Experiments

Principles of experimental design

1. ***Control:*** Compare treatment of interest to a control group.
2. ***Randomize:*** Randomly assign subjects to treatments, and randomly sample from the population whenever possible.
3. ***Replicate:*** Within a study, replicate by collecting a sufficiently large sample. Or replicate the entire study.
4. ***Block:*** If there are variables that are known or suspected to affect the response variable, first group subjects into ***blocks*** based on these variables, and then randomize cases within each block to treatment groups.

More on blocking

- We would like to design an experiment to investigate if energy gels makes you run faster:

More on blocking

- We would like to design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel

More on blocking

- We would like to design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel
- It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:

More on blocking

- We would like to design an experiment to investigate if energy gels makes you run faster:
  - Treatment: energy gel
  - Control: no energy gel
- It is suspected that energy gels might affect pro and amateur athletes differently, therefore we block for pro status:
  - Divide the sample to pro and amateur
  - Randomly assign pro athletes to treatment and control groups
  - Randomly assign amateur athletes to treatment and control groups
  - Pro/amateur status is equally represented in the resulting treatment and control groups

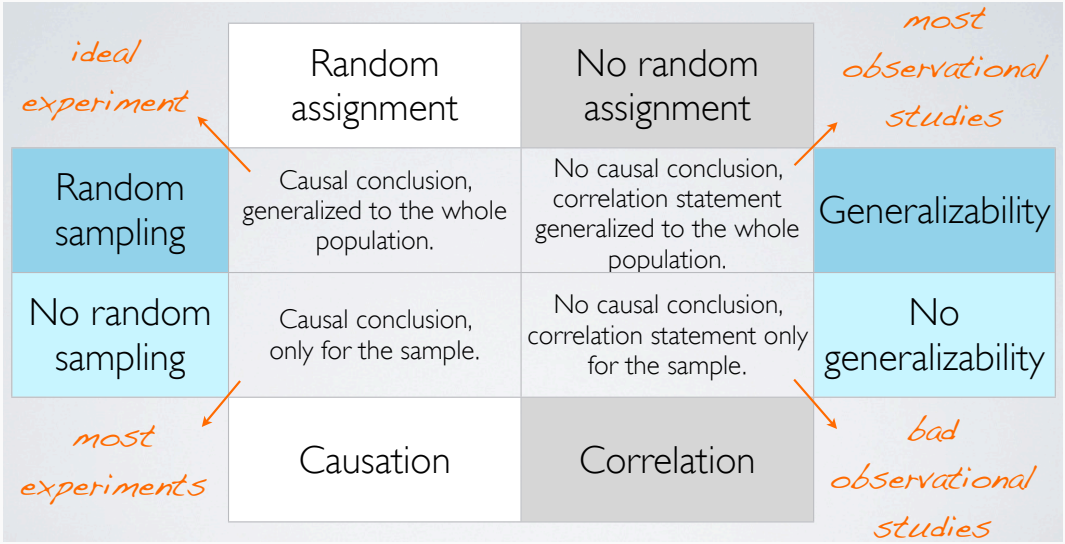
Difference between blocking and explanatory variables

- Factors are conditions we can impose on the experimental units.
- Blocking variables are characteristics that the experimental units come with, that we would like to control for.
- Blocking is like stratifying, except used in experimental settings when randomly assigning, as opposed to when sampling.

More experimental design terminology...

- ***Placebo:*** fake treatment, often used as the control group for medical studies
- ***Placebo effect:*** experimental units showing improvement simply because they believe they are receiving a special treatment
- ***Blinding:*** when experimental units do not know whether they are in the control or treatment group
- ***Double-blind:*** when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

Random assignment vs. random sampling



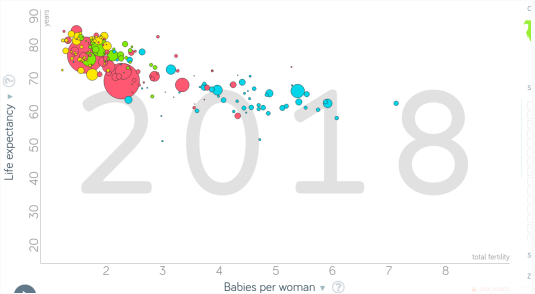
Examining numerical data

Scatterplots

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?

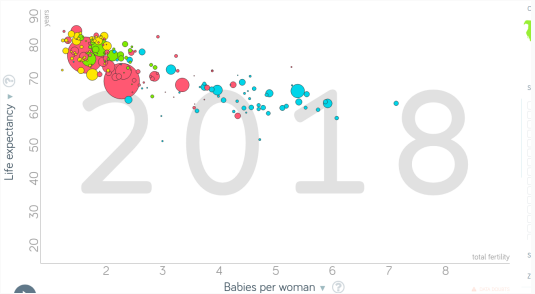


Source: `http://www.gapminder.org/world`

Scatterplots

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

- 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?

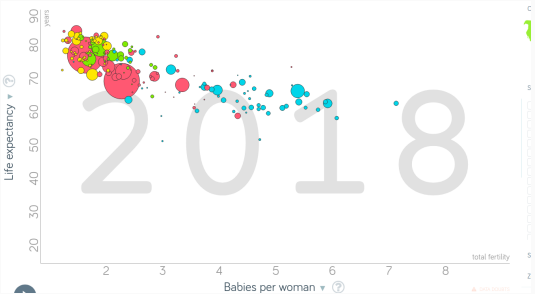


Source: `http://www.gapminder.org/world`

Scatterplots

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

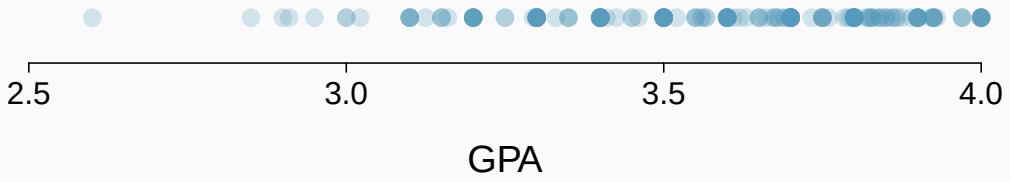
- Do life expectancy and total fertility appear to be ***associated*** or ***independent***?
  - They appear to be linearly and negatively associated: as fertility increases, life expectancy decreases.
- Was the relationship the same throughout the years, or did it change?
  - The relationship changed over the years.



Source: `http://www.gapminder.org/world`

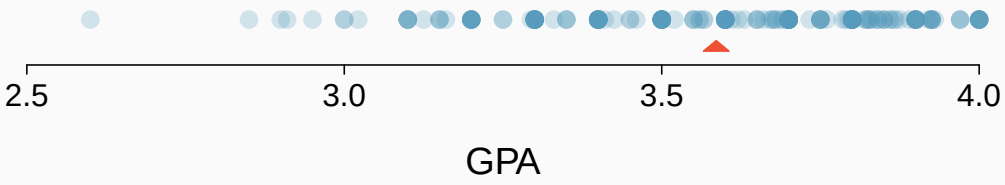
Dot plots

Useful for visualizing one numerical variable. Darker colors represent areas where there are more observations.



How would you describe the distribution of GPAs in this data set? Make sure to say something about the center, shape, and spread of the distribution.

Dot plots & mean



- The ***mean***, also called the ***average*** (marked with a triangle in the above plot), is one way to measure the center of a ***distribution*** of data.
- The mean GPA is 3.59.

Mean

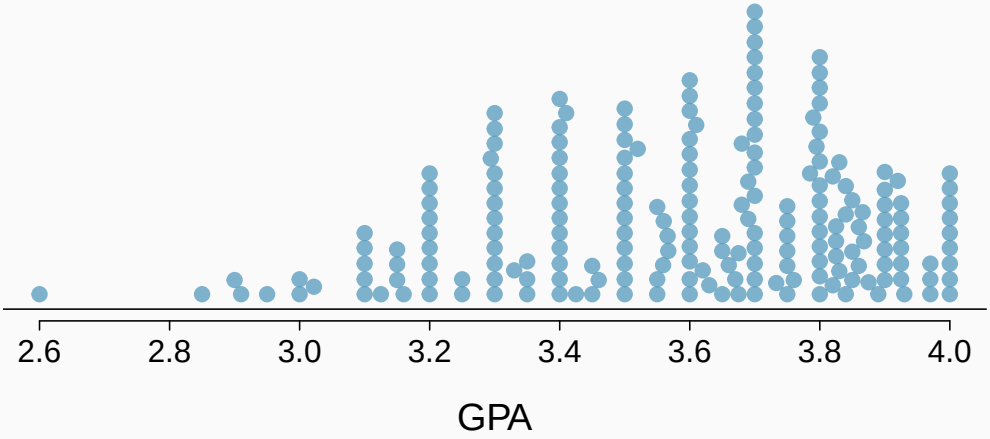
- The ***sample mean***, denoted as $\bar{x}$, can be calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

  where $x_1, x_2, \cdots, x_n$ represent the ***n*** observed values.
- The ***population mean*** is also computed the same way but is denoted as μ. It is often not possible to calculate μ since population data are rarely available.
- The sample mean is a ***sample statistic***, and serves as a ***point estimate*** of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

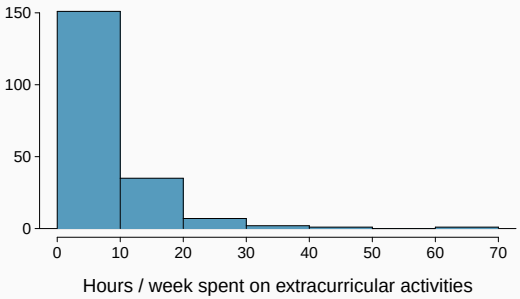
Stacked dot plot

Higher bars represent areas where there are more observations, makes it a little easier to judge the center and the shape of the distribution.



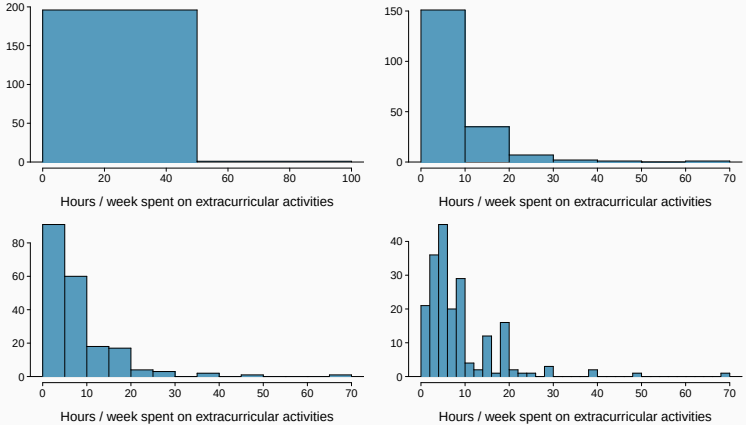
Histograms - Extracurricular hours

- Histograms provide a view of the ***data density***. Higher bars represent where the data are relatively more common.
- Histograms are especially convenient for describing the ***shape*** of the data distribution.
- The chosen ***bin width*** can alter the story the histogram is telling.



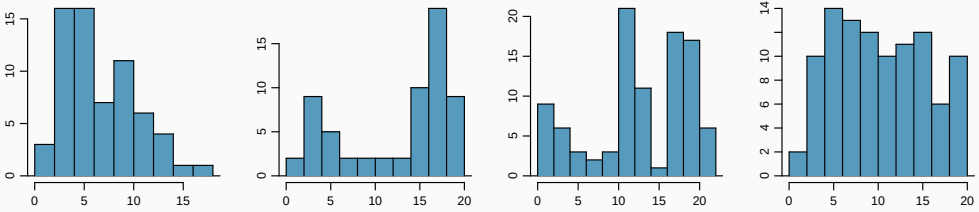
Bin width

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



Shape of a distribution: modality

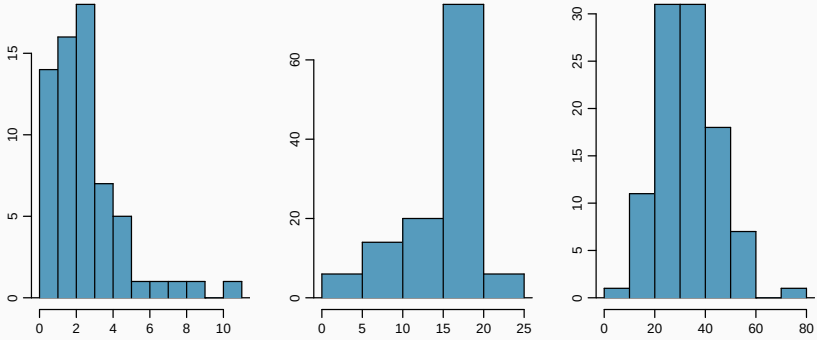
Does the histogram have a single prominent peak (***unimodal***), several prominent peaks (***bimodal/multimodal***), or no apparent peaks (***uniform***)?



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.

Shape of a distribution: skewness

Is the histogram ***right skewed***, ***left skewed***, or ***symmetric***?



Note: Histograms are said to be skewed to the side of the long tail.

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)

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)

Variance

Variance is roughly the average squared deviation from the mean.

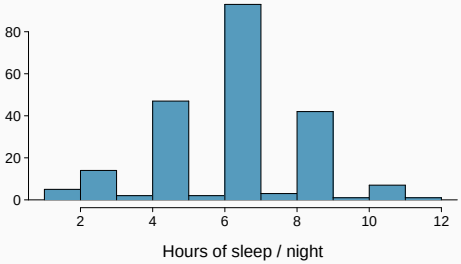
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Variance

Variance is roughly the average squared deviation from the mean.

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

- The sample mean is $\bar{x} = 6.71$, and the sample size is $n = 217$.

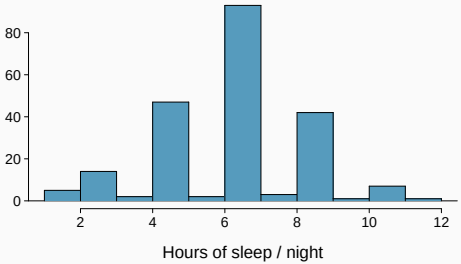


Variance

Variance is roughly the average squared deviation from the mean.

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

- The sample mean is $\bar{x} = 6.71$, and the sample size is $n = 217$.
- The variance of amount of sleep students get per night can be calculated as:



$$s^2 = \frac{(5-6.71)^2 + (9-6.71)^2 + \cdots + (7-6.71)^2}{217-1} = 4.11 \ hours^2$$

Variance (cont.)

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

Variance (cont.)

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

The ***standard deviation*** is the square root of the variance, and has the same units as the data.

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

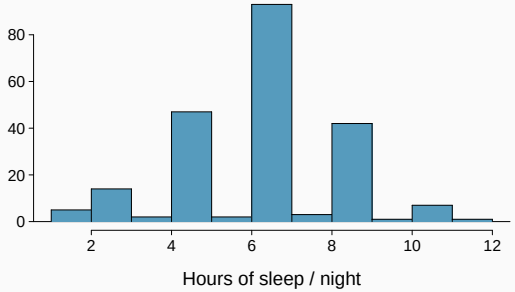
Standard deviation

The ***standard deviation*** is the square root of the variance, and has the same units as the data.

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

- The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \text{ hours}$$



Standard deviation

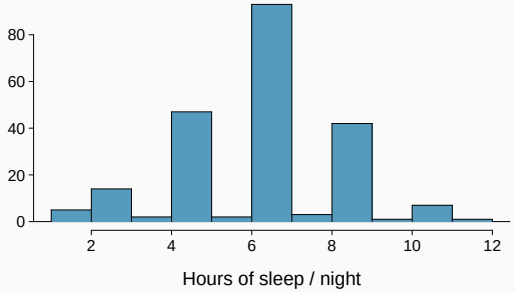
The ***standard deviation*** is the square root of the variance, and has the same units as the data.

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

- The standard deviation of amount of sleep students get per night can be calculated as:

$$s = \sqrt{4.11} = 2.03 \text{ hours}$$

- We can see that all of the data are within 3 standard deviations of the mean.



Median

- The ***median*** is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

- If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

- Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50^{th} ***percentile***.

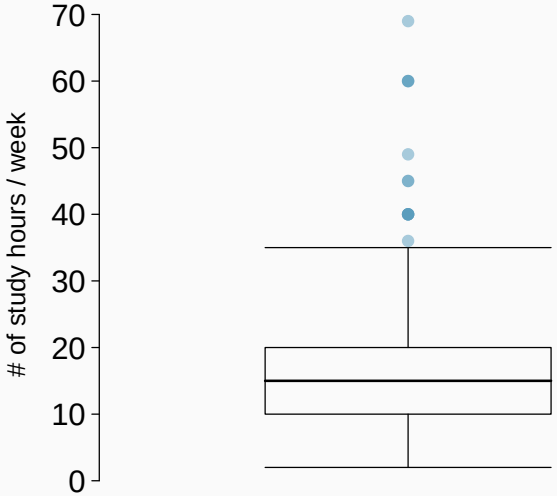
Q1, Q3, and IQR

- The 25^{th} percentile is also called the first quartile, ***Q1***.
- The 50^{th} percentile is also called the median.
- The 75^{th} percentile is also called the third quartile, ***Q3***.
- Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the ***interquartile range***, or the ***IQR***.

$$IQR = Q3 - Q1$$

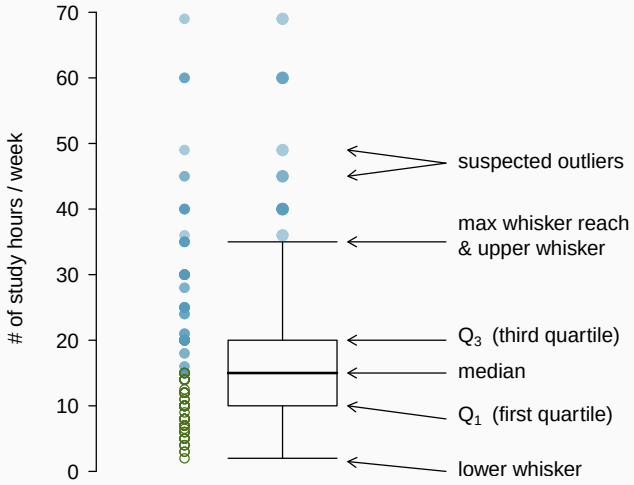
Box plot

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



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.



Whiskers and outliers

- ***Whiskers*** of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$

$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

Whiskers and outliers

- ***Whiskers*** of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$
$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

$$\text{IQR} : 20 - 10 = 10$$
$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$
$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

Whiskers and outliers

- ***Whiskers*** of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$
$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

$$\text{IQR} : 20 - 10 = 10$$
$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$
$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

- 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 (cont.)

Why is it important to look for outliers?

Outliers (cont.)

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.