Sta 111 - Summer II 2017
Probability and Statistical Inference

1. Introduction & exploratory data analysis

Lu Wang

Duke University, Department of Statistical Science

June 27, 2017

Outline

1. Course outline

2. Exploratory data analysis

1. Always start your exploration with a visualization
2. When describing numerical distributions discuss shape, center, spread, and unusual observations
3. Robust statistics are not easily affected by outliers and extreme skew
4. Use box plots to display quartiles, median, and outliers

Course outline

General information at *https://wangronglu.github.io/sta111*

- Mathematics behind statistics
  - Ch 2 - Probability: basics of probability, conditional probability, Bayes' Theorem
  - Ch 3 - Distributions: binomial and normal distributions.
- Statistical inference
  - Ch 4 - Framework for inference: CLT, confidence intervals, hypothesis testing.
  - Midterm 1
  - Ch 5 - Statistical inference for numerical variables
  - Ch 6 - Statistical inference for categorical variables
  - Midterm 2
- Modeling
  - Ch 7 - Introduction to linear regression: bivariate correlation, introduction to modeling.
  - Ch 8 - Multiple and logistic regression: more advanced modeling with multiple predictors and binary response.
  - Final Exam

Data matrix

Data collected on students in a statistics class on a variety of variables:

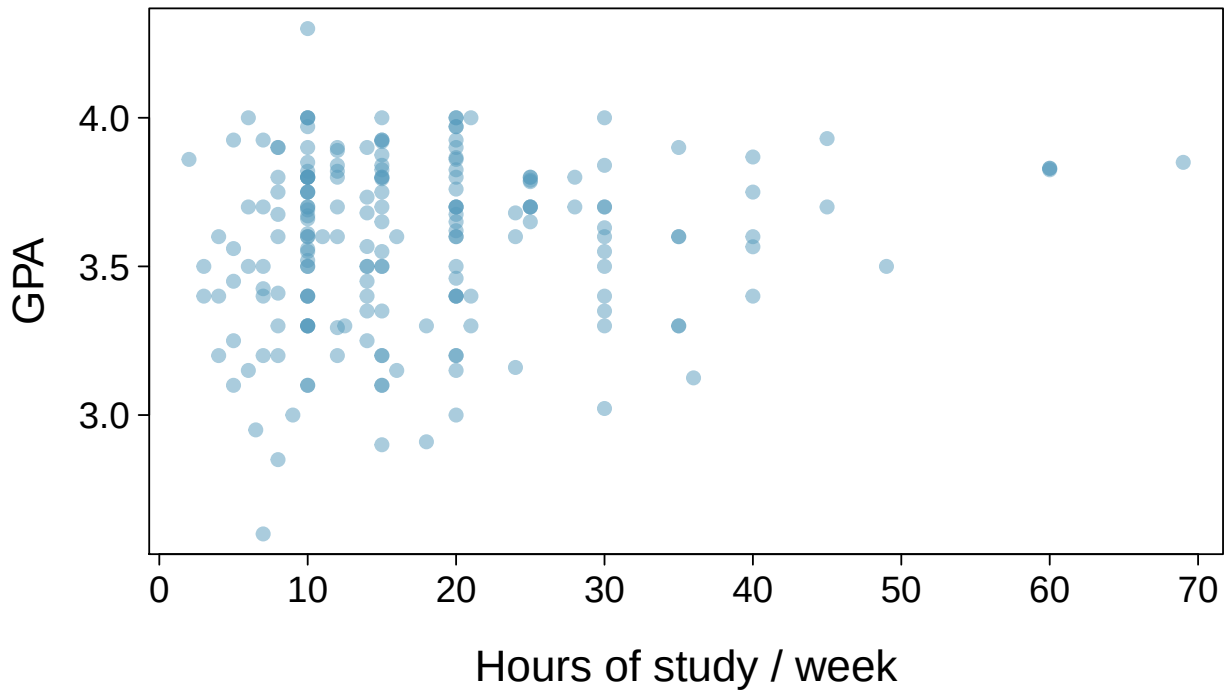
variable ↓

Stu.	gender	intro_extra	···	dread
1	male	extravert	···	3
2	female	extravert	···	2
3	female	introvert	···	4
4	female	extravert	···	2
⋮	⋮	⋮	⋮	⋮
86	male	extravert	···	3

← *observation*

Relationships among variables

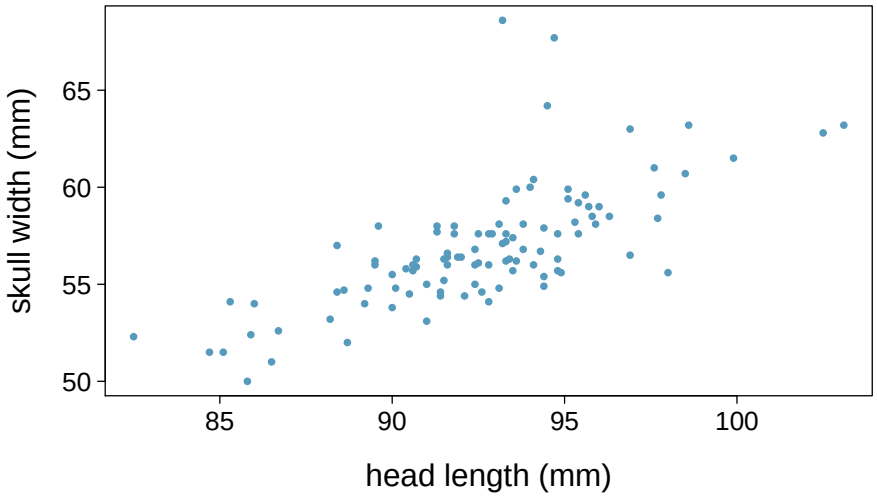
Scatterplots are useful for visualizing the relationship between two numerical variables.
Does there appear to be a relationship between GPA and number of hours students study per week?



Can you spot anything unusual about any of the data points?

Relationships among variables

Based on the scatterplot on the right, which of the following statements is correct about the head and skull lengths of possums?



(a) There is no relationship between head length and skull width, i.e. the variables are independent.

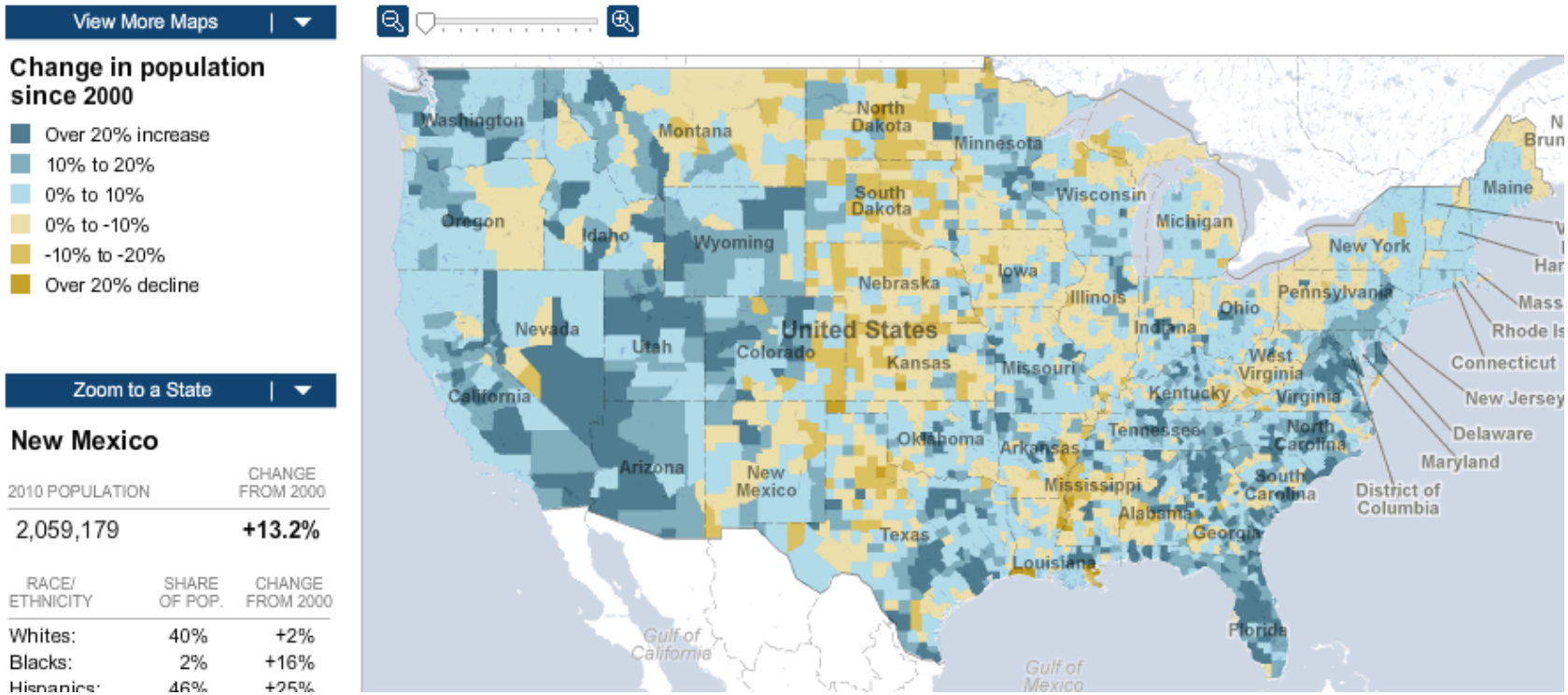
(b) Head length and skull width are positively associated.

(c) A longer head causes the skull to be wider.

(d) A wider skull causes the head to be longer.

Intensity map

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



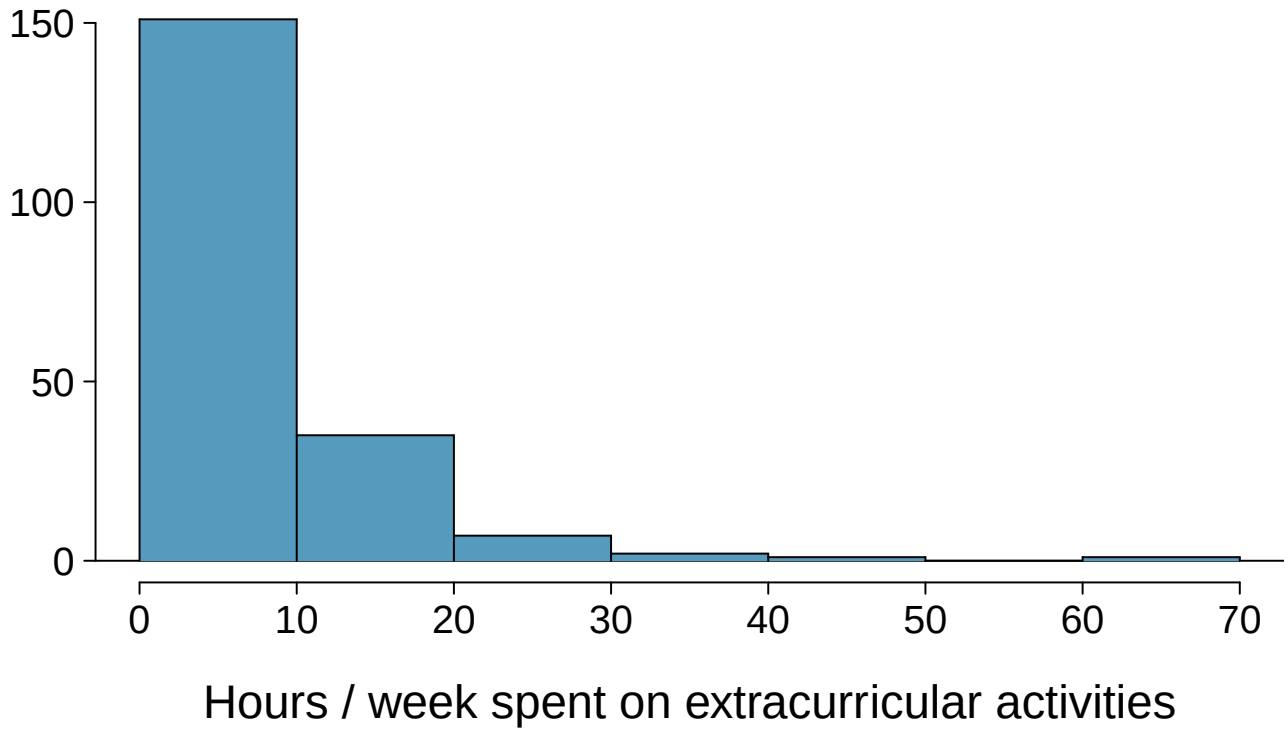
http://projects.nytimes.com/census/2010/map

Describing distributions of numerical variables

- Shape: skewness, modality
- Unusual observations: observations that stand out from the rest of the data that may be suspected outliers
- Center: an estimate of a *typical* observation in the distribution (mean, median, mode, etc.)
  - Notation: μ: population mean, $\bar{x}$: sample mean
- Spread: measure of variability in the distribution (standard deviation, IQR, range, etc.)

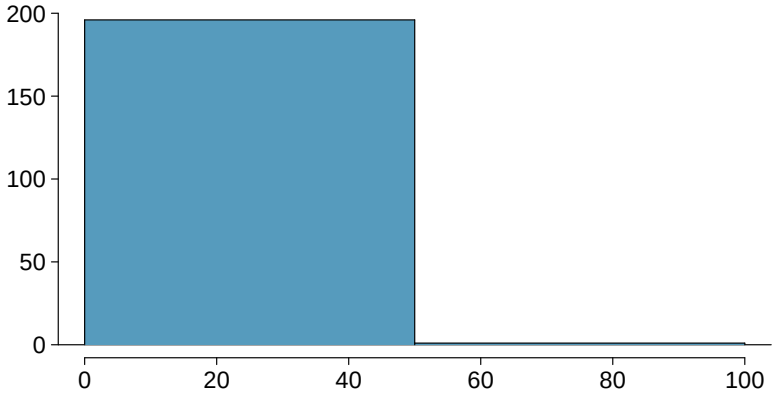
Histograms - Extracurricular hours

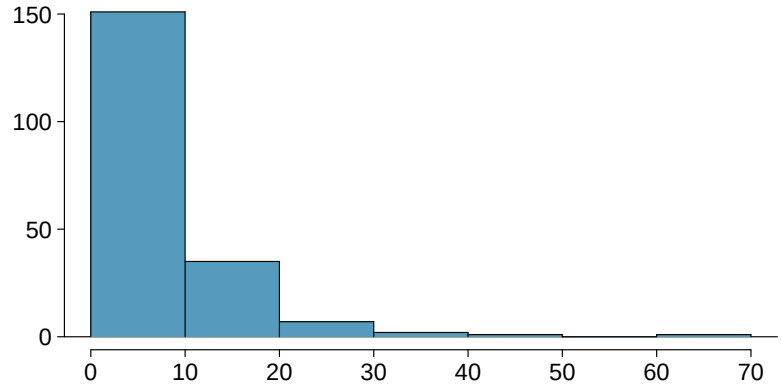
- Histograms provide a view of the *data distribution*. Higher bars represent where the data are relatively more common.
- Histograms are especially convenient for describing the *shape* of the data distribution.
- The chosen *bin width* can alter the story the histogram is telling.

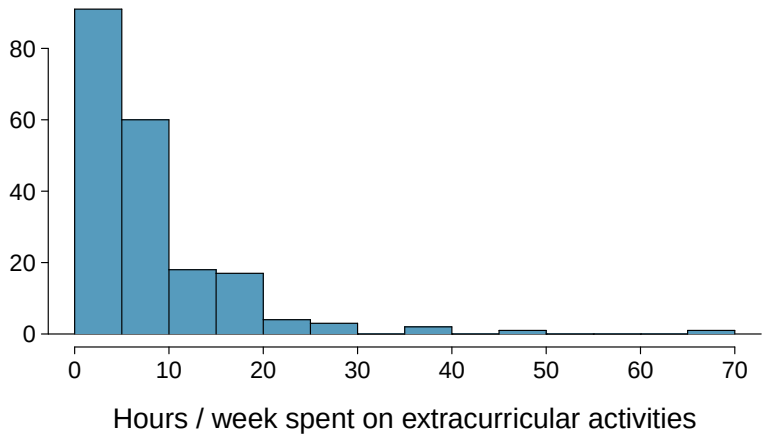


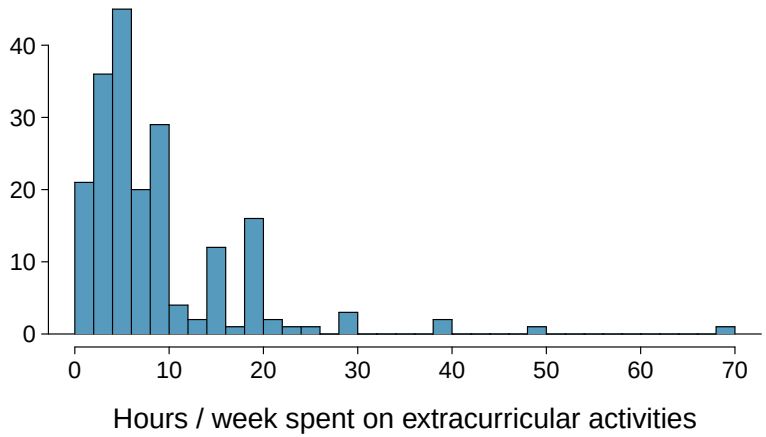
Bin width

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



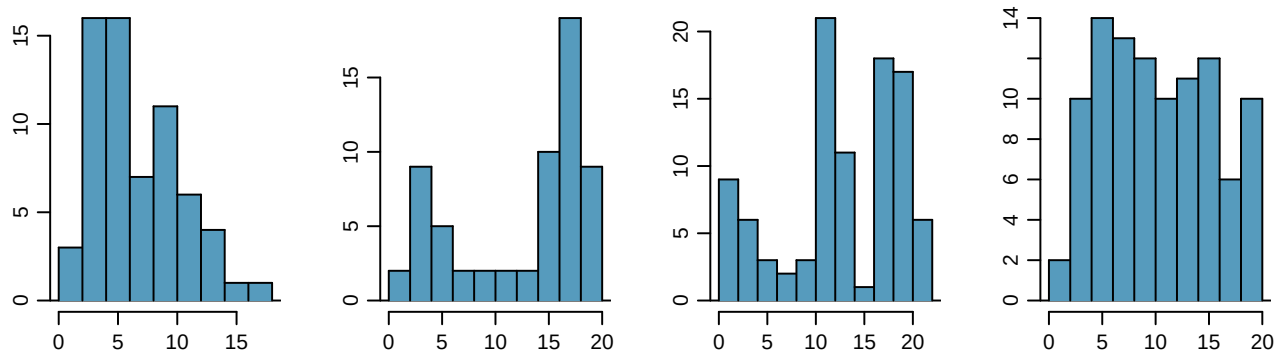






Shape of a distribution: modality

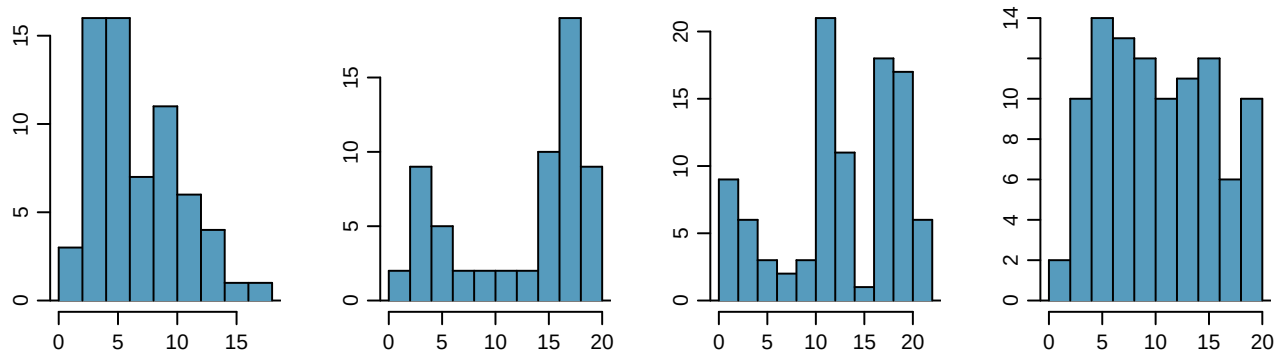
Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?



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: modality

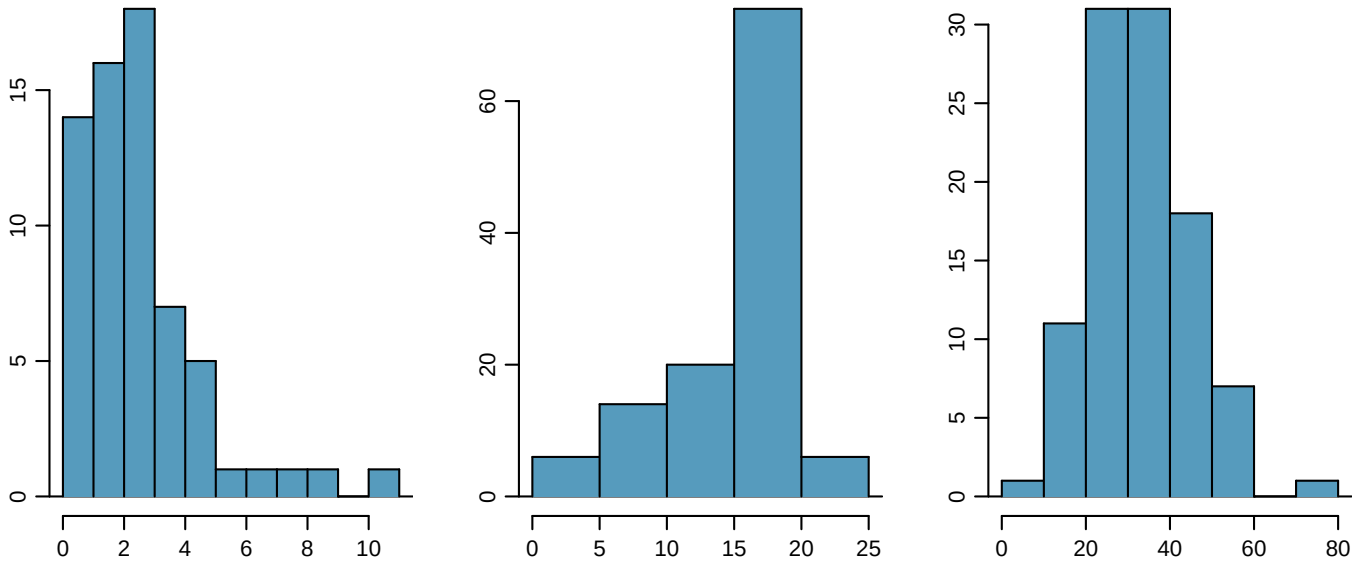
Does the histogram have a single prominent peak (*unimodal*), several prominent peaks (*bimodal/multimodal*), or no apparent peaks (*uniform*)?



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

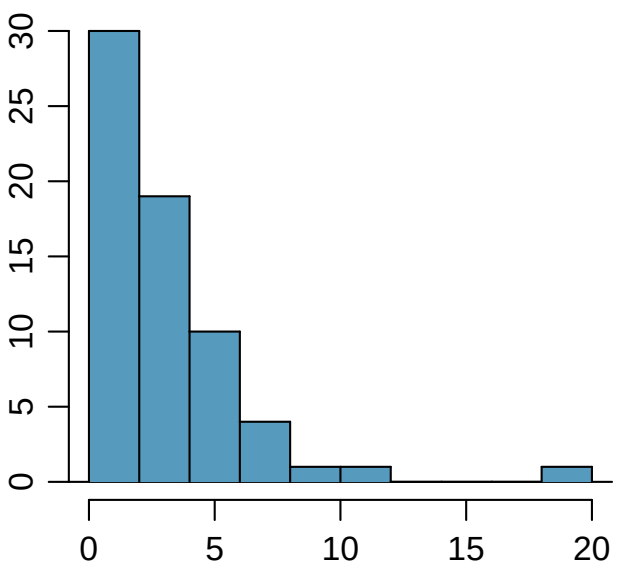
Is the histogram *right skewed*, *left skewed*, or *symmetric*?

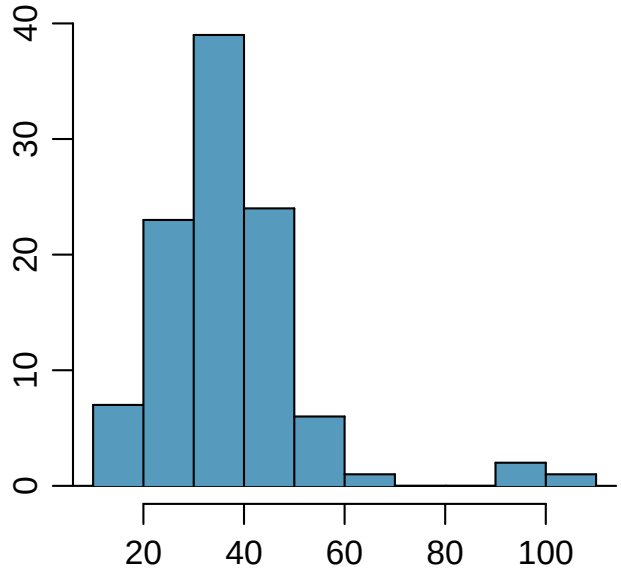


Note: Histograms are said to be skewed to the side of the long tail.

Shape of a distribution: unusual observations

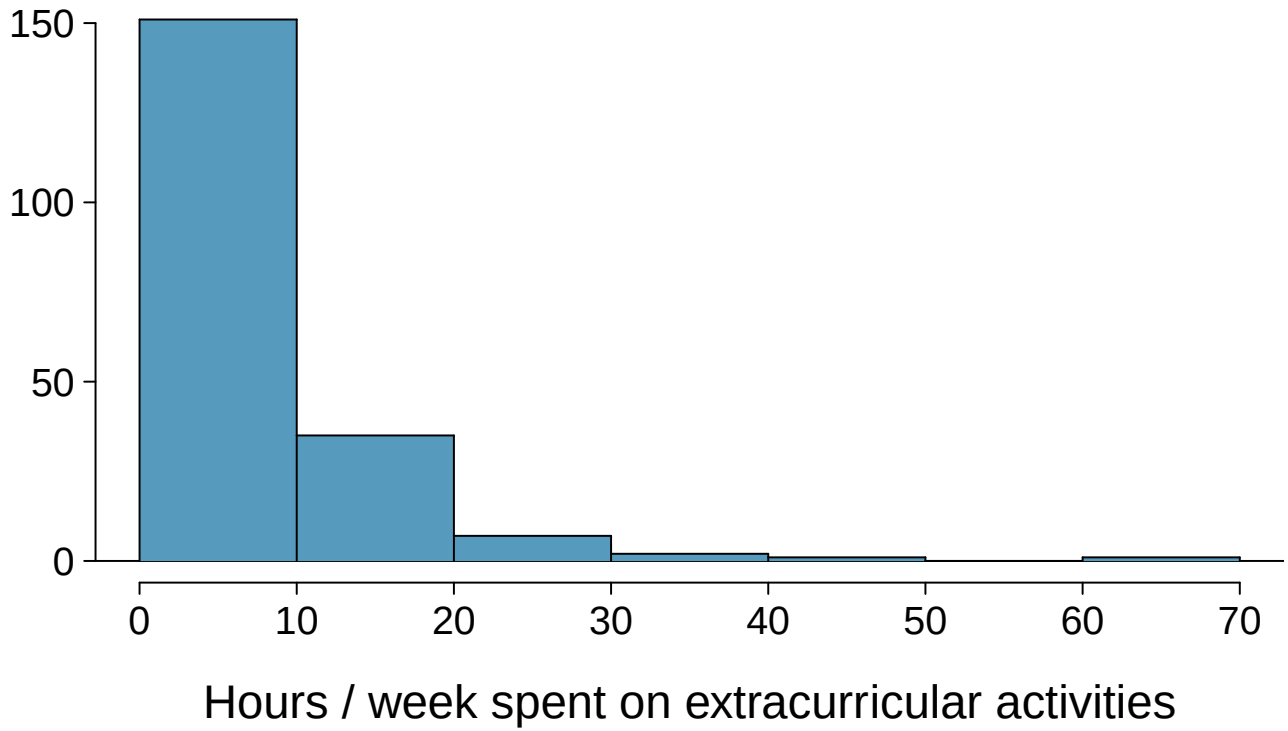
Are there any unusual observations or potential *outliers*?





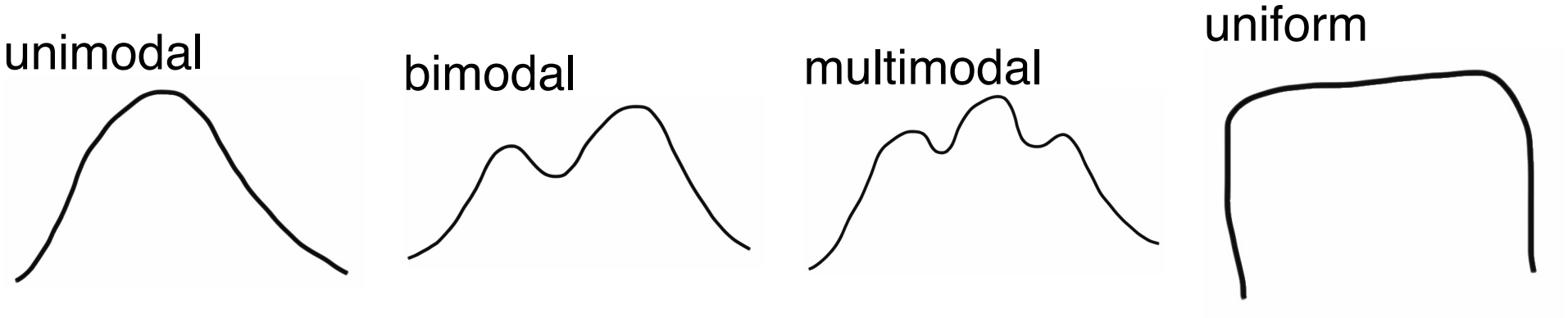
Extracurricular activities

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



Commonly observed shapes of distributions

- modality



- skewness

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)

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.

Median

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

$$0, 1, \mathit{2}, 3, 4$$

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

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

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

How do the mean and median of the following two datasets compare?
Dataset 1: 30, 50, 70, 90
Dataset 2: 30, 50, 70, 1000

(a) $\bar{x}_1 = \bar{x}_2$, $median_1 = median_2$

(b) $\bar{x}_1 < \bar{x}_2$, $median_1 = median_2$

(c) $\bar{x}_1 < \bar{x}_2$, $median_1 < median_2$

(d) $\bar{x}_1 > \bar{x}_2$, $median_1 < median_2$

(e) $\bar{x}_1 > \bar{x}_2$, $median_1 = median_2$

Standard deviation and variance

- Most commonly used measure of variability is the *standard deviation*, which roughly measures the average deviation from the mean.
  - Notation: σ: population standard deviation, s: sample standard deviation

- Calculating the standard deviation, for a population (rarely, if ever) and for a sample:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{n}} \qquad s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

  - The *standard deviation* has the same unit as the data.

- Square of the standard deviation is called the *variance*.

Why divide by $n - 1$ instead of n when calculating the sample standard deviation?

Lose a "degree of freedom" for using an estimate (the sample mean $\bar{x}$) in estimating the sample variance/standard deviation.

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.

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.

$$IQR = Q3 - Q1$$

True / False: The range is always at least as large as the IQR for a given dataset.

(a) Yes

(b) No

Is the range or the IQR more robust to outliers?

Robust statistics

- Mean and standard deviation are easily affected by extreme observations since the value of each data point contributes to their calculation.
- Median and IQR are more robust.
- 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.

Box plot

A *box plot* visualizes the median, the quartiles, and suspected outliers. An *outlier* is defined as an observation more than $1.5 \times$ IQR away from the quartiles.