Unit 1: Introduction to Data

2. Exploratory Data Analysis

(Chapter 1.6)

1/22/2020

Quiz 1 - Data and where it comes from

A sampling metaphor

When you taste a spoonful of soup and decide the spoonful you tasted isn’t salty enough, that’s **exploratory analysis**

If you generalize and conclude that your entire soup needs salt, that’s an **inference**

For your inference to be valid, the spoonful you tasted (the **sample**) needs to be **representative** of the entire pot (the **population**)

If the soup is not well stirred, it doesn't matter how large a spoon you have, it will still not taste right. If the soup is well stirred, a small spoon will suffice to test the soup.

Thanks Mine Çetinkaya-Rundel

Key ideas

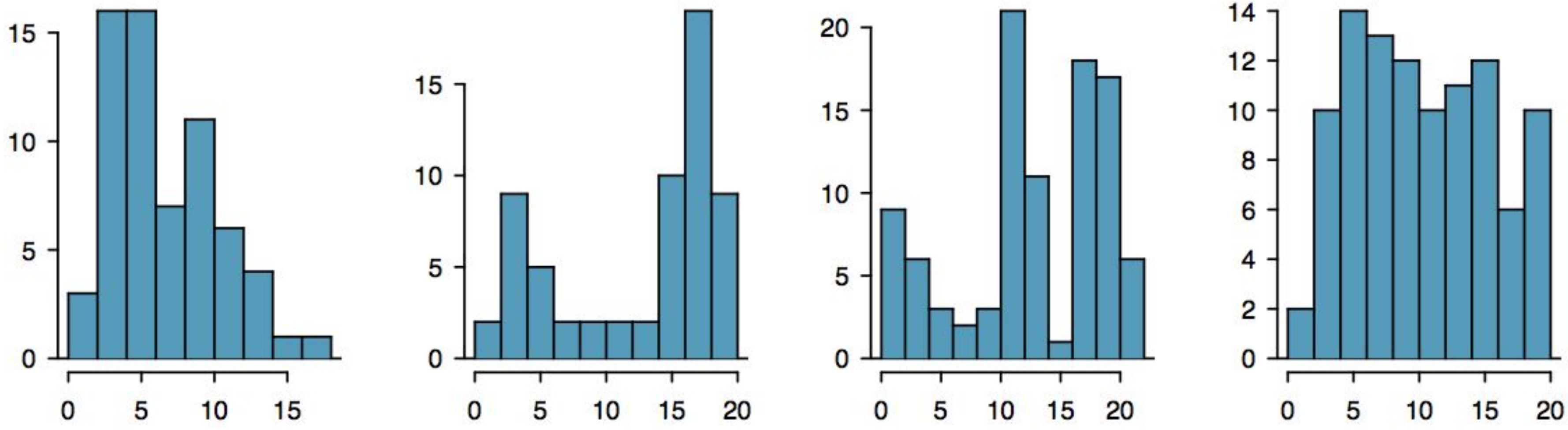
1. Always start by visualizing your data
2. Descriptive statistics compress data to make it easier to understand and communicate about
3. We generally want to talk about **shape**, **center**, and **spread**

Getting some data

1. Your height in inches
2. Your birth month (numerical)
3. Number of siblings

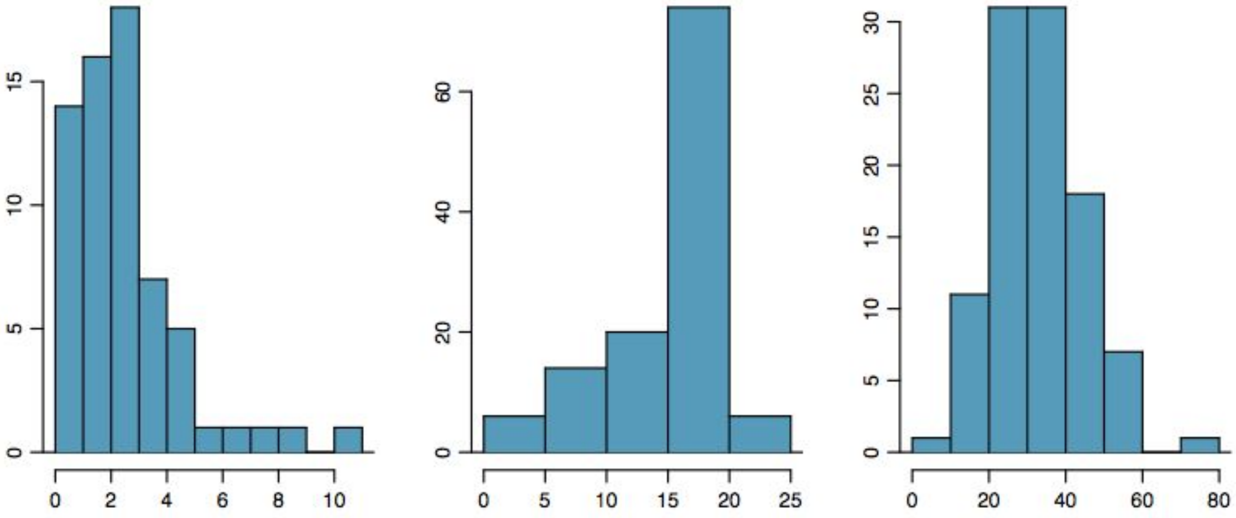
Shape of a distribution: Modality

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



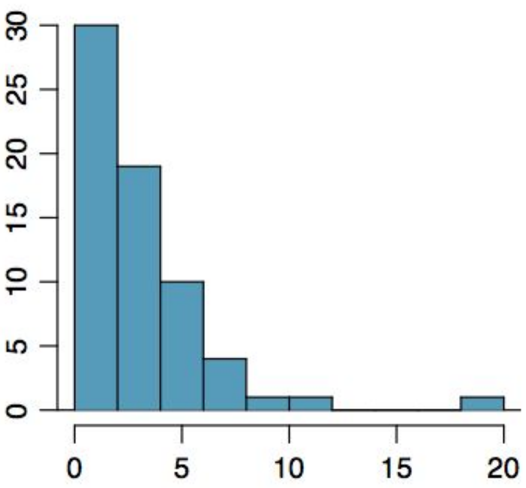
Shape of a distribution: Skewness

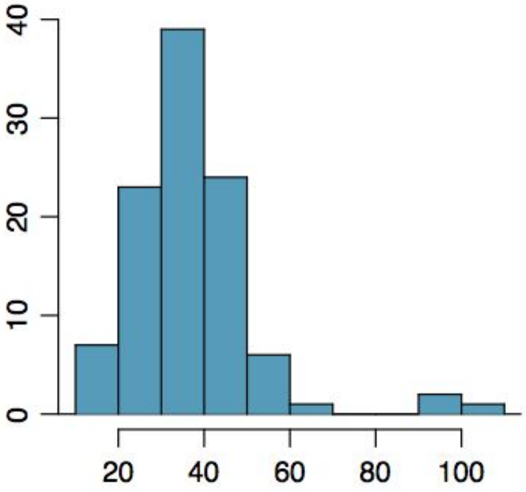
Is the histogram right-skewed, left-skewed, or symmetric?



Shape of a distribution: Outliers

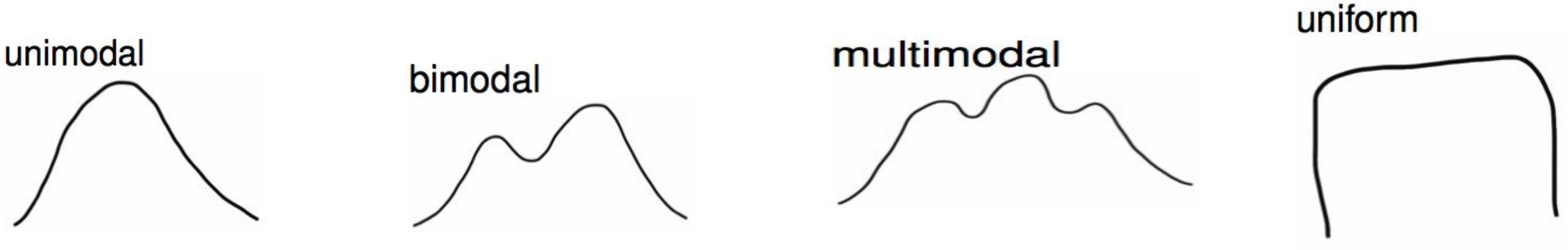
Are there any unusual observations or potential outliers?





Common shapes of distributions

Modality



Skewness

Practice Question 1

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.

Central tendency

What's the difference between .mp3 and .FLAC?
.jpeg and .png?

.mp3 and .jpeg are **lossy compression** -- they make data smaller by throwing some of it away.

Central tendency is a kind of lossy compression: **What one number is the most representative of my data**?

One measure of central tendency: The 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, …, 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 an 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.

Spread: How different is my data (on average) from the center?

The standard deviation(s) is roughly the average deviation from the mean

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

The population standard deviation is denoted σ is also computed the same way, except that you do not subtract one from the number of measurements

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

The square of the standard deviation (σ^2) is called the variance

Details of the standard deviation

Why did we divide by *n-1* instead of n when calculating the sample standard deviation (s)?

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

Why did we use the squared deviation in calculating spread?

1. To get rid of negatives so that observations equally distant from the mean are weighted equally
2. To weigh large deviations more heavily

Key ideas

1. Always start by visualizing your data
2. Descriptive statistics compress data to make it easier to understand and communicate about
3. We generally want to talk about **shape**, **center**, and **spread**