STA13: Elementary Statistics

Lecture 5

Dmitriy Izyumin

January 19 2018

Population vs. Sample

The population of N objects has the following parameters:

Mean	$\mu = \frac{1}{N}\sum_{i=1}^{N}$
Variance	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$
Standard deviation	$\sigma = \sqrt{\sigma^2}$

A sample of n objects has the following statistics:

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

Parameters and Statistics

A parameter is a number that describes the population.

A statistic is a number that is computed from a sample.

Parameters are fixed (constant value), but typically unknown.

Statistics have many possible values (from different possible samples), and one observed value.

Statistics are used to estimate parameters.

Empirical Rule

Applies to mound-shaped, symmetric distributions and samples.

- Approximately 68% observations will be within one s.d. of the mean, i.e. in the interval $(\mu - \sigma, \mu + \sigma)$.
  Use interval $(\bar{x} - s, \bar{x} + s)$ for a sample.
- Approximately 95% observations will be within two s.d.'s of the mean, i.e. in the interval $(\mu - 2\sigma, \mu + 2\sigma)$.
  Use interval $(\bar{x} - 2s, \bar{x} + 2s)$ for a sample.
- Approximately 99.7% observations will be within three s.d.'s of the mean, i.e. in the interval $(\mu - 3\sigma, \mu + 3\sigma)$.
  Use interval $(\bar{x} - 3s, \bar{x} + 3s)$ for a sample.

Chebyshev's Rule

Applies to all distributions and samples.

For any $k > 1$, at least $(1 - \frac{1}{k^2})$ of the observations will be within k standard deviations of the mean.

- At least $\frac{3}{4}$ observations will be within two s.d.'s of the mean, i.e. in the interval $(\mu - 2\sigma, \mu + 2\sigma)$.
  Use interval $(\bar{x} - 2s, \bar{x} + 2s)$ for a sample.

- At least $\frac{8}{9}$ observations will be within three s.d.'s of the mean, i.e. in the interval $(\mu - 3\sigma, \mu + 3\sigma)$.
  Use interval $(\bar{x} - 3s, \bar{x} + 3s)$ for a sample.

Z-scores

The z-score of an observation x_i is defined as

$$z_i = \frac{x_i - \mu}{\sigma}$$

- Answers the question "How many standard deviations away from the mean is an ovservation?"
- If we start at $\bar{x}$ and take steps the size of σ, then z_i is the number of steps we need to take to get to x_i.
- Require knowledge about the population. We can use $\bar{x}$ and s to obtain t-scores.

Z-scores

The z-score of an observation x_i is defined as

$$z_i = \frac{x_i - \mu}{\sigma}$$

- Regardless of the sample, z-scores are always on the same scale.
- Z-scores with absolute value larger than 3 are suspicious (possible outliers).
- Z-scores with absolute value larger than 4 are very suspicious (almost certain outliers).

Ranks

The rank of an observation is its position if the observations are listed from smallest to largest.

- The smallest observation is ranked 1, the second smallest is ranked 2, and so on.
- Ties may get a little tricky.
- Answers the question "How many observations are less than or equal to a given observation?"

Percentiles

The k^{th} percentile is the value chosen such that $k\%$ of the observations fall below that value.

Let's say the 82^{nd} percentile for midterm scores in a STA13 class last quarter was 70 points. This means that 82% of the scores in that class were less than 70 points.

Important Percentiles

The lower quartile (or first quartile) Q1

- larger than 25% of the observations, and less than 75% of observations
- 25^{th} percentile

The median (or second quartile)

- larger than 50% of the observations, and less than 50% of observations
- 50^{th} percentile

The upper quartile (or third quartile) Q2

- larger than 75% of the observations, and less than 25% of observations
- 75^{th} percentile

Five Number Summary

A common way to summarize data is to provide these five numbers:

Min Q1 Median Q3 Max

This cuts up the data into four equal chunks:

25% of observations fall between Min and Q1,
25% of observations fall between Q1 and Median
25% of observations fall between Median and Q3
25% of observations fall between Q3 and Max

IQR

The interquartile range (IQR) is the difference between the third and first quartiles: $IQR = Q3 - Q1$.

IQR is another measure of spread, and is also used for outlier detection.

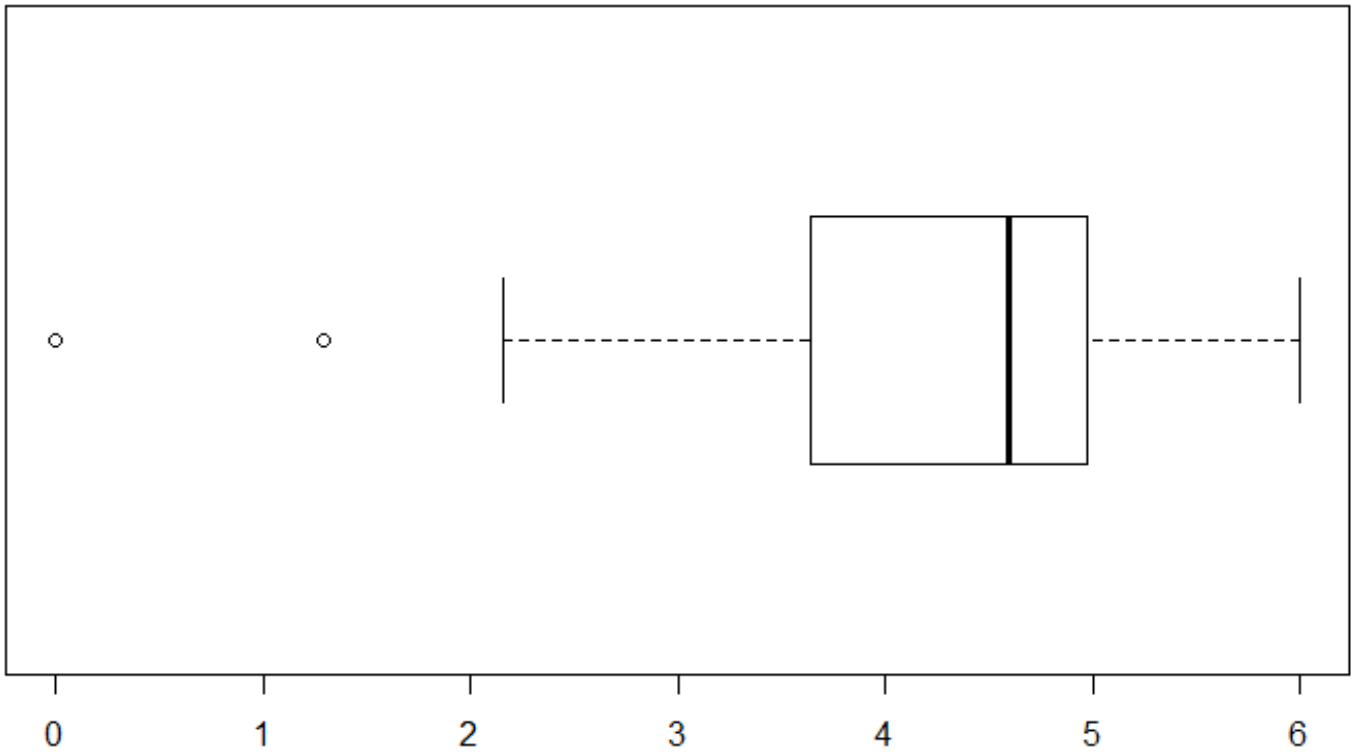
It is common to declare all observations larger than $Q3 + 1.5(\text{IQR})$ or smaller than $Q1 - 1.5(\text{IQR})$ to be outliers.

How are range and IQR similar? How are they different?

Boxplots

A boxplot is another tool for graphically summarizing the data. It is essentially a visual representation of the five number summary.

It is sometimes called a "box and whisker plot."



Boxplots

Steps for making a boxplot:

1. Draw a numberline large enough to encompass the minimum and maximum values of the data.
2. Compute the median, Q1 and Q3. Use these to construct the box.
3. Compute the IQR and the fences.
4. Draw the left whisker goes from Q1 to the smallest observation that is larger than $Q1 - 1.5(\text{IQR})$.
5. Draw the right whisker goes from Q3 to the largest observation that is smaller than $Q3 + 1.5(\text{IQR})$.
6. Mark the outliers (observations not within the fences).

Example

15 observations:

0, 1.29, 2.16, 3.50, 3.78, 3.94, 4.40, 4.60, 4.62, 4.80, 4.81, 5.13, 5.50, 5.83, 6.00

- Five number summary: $0, 3.50, 4.60, 5.13, 6$
- IQR = 5.13-3.50=1.63
  upper fence: 5.13+1.5(1.63)=7.575
  lower fence: 3.50-1.5(1.63)=1.055
- Observations above 7.575 or below 1.055 are outliers
  Outliers: 0, 1.29

Reading a Boxplot

Spread

As with the five number summary, the data are divided (roughly) into quarters. The sections of the boxplot can help us understand how tightly packed or spread out these quarter chunks of the data are.

Outliers

The presence of outliers is easily detected with a boxplot.

Reading a Boxplot

Skew

- Median is roughly in the middle of the box, and whiskers are roughly the same length - symmetric
- Median is in the right half of box, and the left whisker is longer - skewed left
- Median is in the left half of box, and the right whisker is longer - skewed right

Comparing Two Boxplots

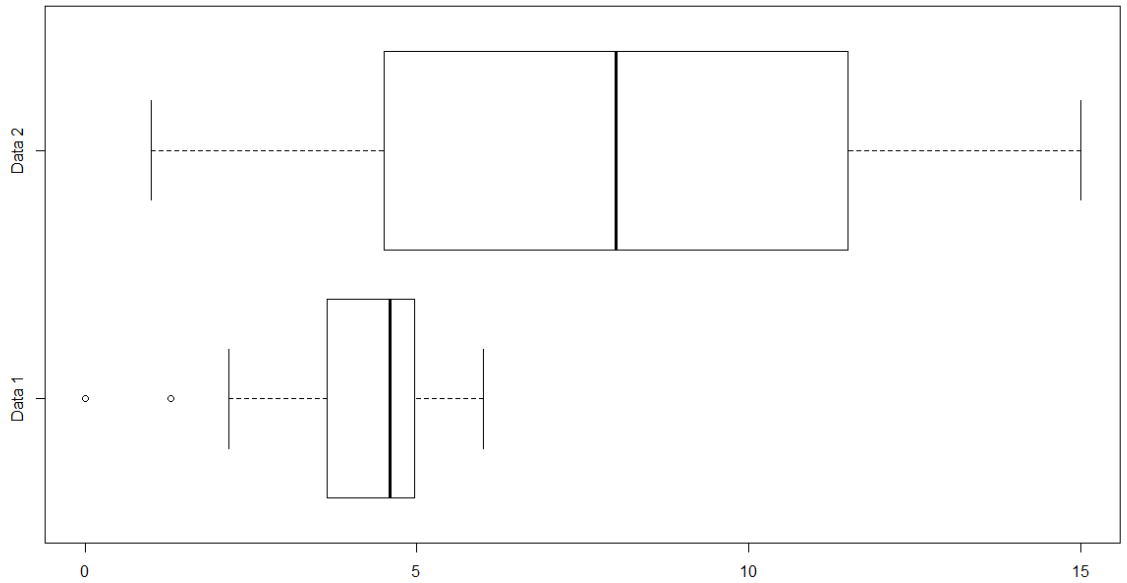
Often we want to compare two (or more) boxplots.

Data 1:

0, 1.29, 2.16, 3.50, 3.78, 3.94, 4.40, 4.60, 4.62, 4.80, 4.81, 5.13, 5.50, 5.83, 6.00

Data 2:

1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15



What can you say about the plots?

Comparing Two Boxplots

What can you say about the plots?

- Compare centers
  Data 1 is centered around 4.6, Data 1 is centered around 8
  Data 1 has a higher average value
- Compare spread
  Data 2 has a noticeably larger spread that Data 1
- Compare skew
  Data 1 is skewed left, Data 2 is symmetric
- Outliers
  Data 1 has two outliers, Data 2 has none
- Think in terms of the "quarter chunks" of data
  More than 50% of the Data 2 observations are greater than the biggest Data 1 observation