Descriptive statistics and EDA, chapter 3.1

Grethe Hystad
University of Arizona

August 19, 2012

Statistics

Definition

Statistics is a mathematical science that studies the collection, analysis, interpretation, and presentation of data.

- In testing for a new drug, one compares the new drug with the old one on the same individual. Is the new drug better?
- Do smaller class sizes improve student's learning?
- Do children who play soccer, do better in school?
- Are certain types of cancer related to the Tsjernobyl accident?
- Do higher SAT scores lead to higher graduation rate?
- Do male hook-billed kites have larger tail and wing length than female hook-billed kites?

In the news

- Recent Los Angeles times article:
  "No yolk: eating the whole egg as dangerous as smoking?"
- The research article is published in Atherosclerosis, "Egg yolk consumption and carotid plaque."
- The research article's conclusion: "Our findings suggest that regular consumption of egg yolk should be avoided by persons at risk of cardiovascular disease. This hypothesis should be tested in a prospective study with more detailed information about diet, and other possible confounders such as exercise and waist circumference."
- The LA times wrote instead: "We believe our study makes it imperative to reassess the role of egg yolks, and dietary cholesterol in general, as a risk factor for coronary heart disease."

Statistics

- Any set of data contains information about some group of individuals. These individuals are representative chosen from a population under study.
- **The individuals** are the objects described by a data set. They may be people, animals, or objects.
- **A variable** is a characteristic of an individual like for example, college major, income, genetic disease.
- **A categorical variable** partition the individuals into groups or categories.
- **A quantitative variable** has numerical values for which arithmetic operations such as addition and differences make sense.

Quantitative variables-histogram

- The distribution of a quantitative variable tells us what values the variables takes and how often it takes these values.
- A **histogram** is a graph of the distribution of the quantitative variable.

Creating a histogram

- Given a sample of size n of data.
- Divide the range of data into non-overlapping intervals of equal length. (does not always have to be of equal length).
- Count the number of observations in each class.
- The histogram is constructed by drawing a rectangle representing the frequencies or relative frequencies.

How to construct the intervals in the histogram

- Find the range R=maximum-minimum value of the data.
- Divide the data into k-intervals.
- The intervals are non-overlapping, usually of equal length and cover the range R.
- Rule of thumb: $k \approx \sqrt{n}$ when $n \leq 400$.
- Each interval begins and ends halfway between two possible values of the observations.
- The first interval should begin about as much below the smallest value as the last interval ends above the largest.
- The intervals are called **class-intervals** and the boundaries are called **cut-points**.
- The **class mark** is the midpoint of a class-interval.

How to construct the histogram

We consider class-intervals of equal length:

- **Frequency histogram:** Draw a rectangle for each class with
  - **base**=the length of the class interval
  - **height**=the frequency (number of observations) of the class.

How to construct the histogram

We consider class-intervals of equal length:

- **Frequency histogram:** Draw a rectangle for each class with
  - **base**=the length of the class interval
  - **height**=the frequency (number of observations) of the class.
- **Relative frequency histogram:** Draw a rectangle for each class with
  - **base**=the length of the class interval
  - **area**= the relative frequency $\frac{f_i}{n}$ of the observations of the class. Here f_i is the frequency of the i^{th} class.
  - That is the **height** of the rectangles are given by the value of the function

$$h(x) = \frac{f_i}{n(c_i - c_{i-1})}$$

for $c_{i-1} < x \leq c_i$ and $i = 1, 2, ..., k$ for intervals (c_0, c_1), (c_1, c_2),...,(c_{k-1}, c_k).

Histogram, example

Example

Suppose the scores of a calculus test in a class of 16 students where distributed as follows:
$49, 50, 52, 52, 61, 67, 73, 73, 82, 82, 90, 94, 99, 99, 99, 100$.
Draw a frequency histogram and a relative frequency histogram.
Determine the fractions of students that score below 74.

Histogram, example

Solution

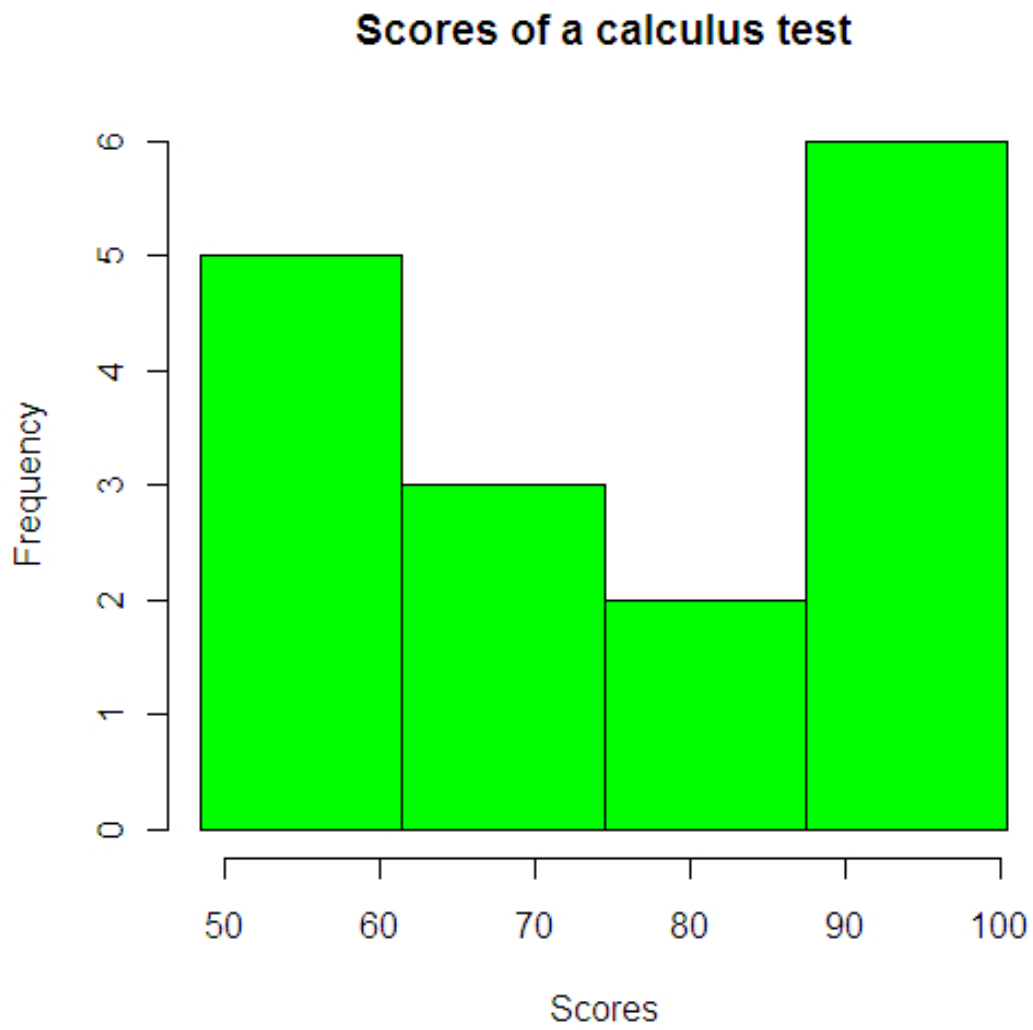
- *Range*=$100 - 49 = 51$.
- *Number of intervals:* $k = \sqrt{16} = 4$
- *Length of intervals:* $\frac{51}{4} = 12.75 \approx 13 = c_i - c_{i-1}$.
- *Now* $(13)(4) = 52$, *so we start the interval* $\frac{52-51}{2} = 0.5$ *below the smallest value.*
- $n(c_i - c_{i-1}) = (16)(13) = 208$.

Solution continue

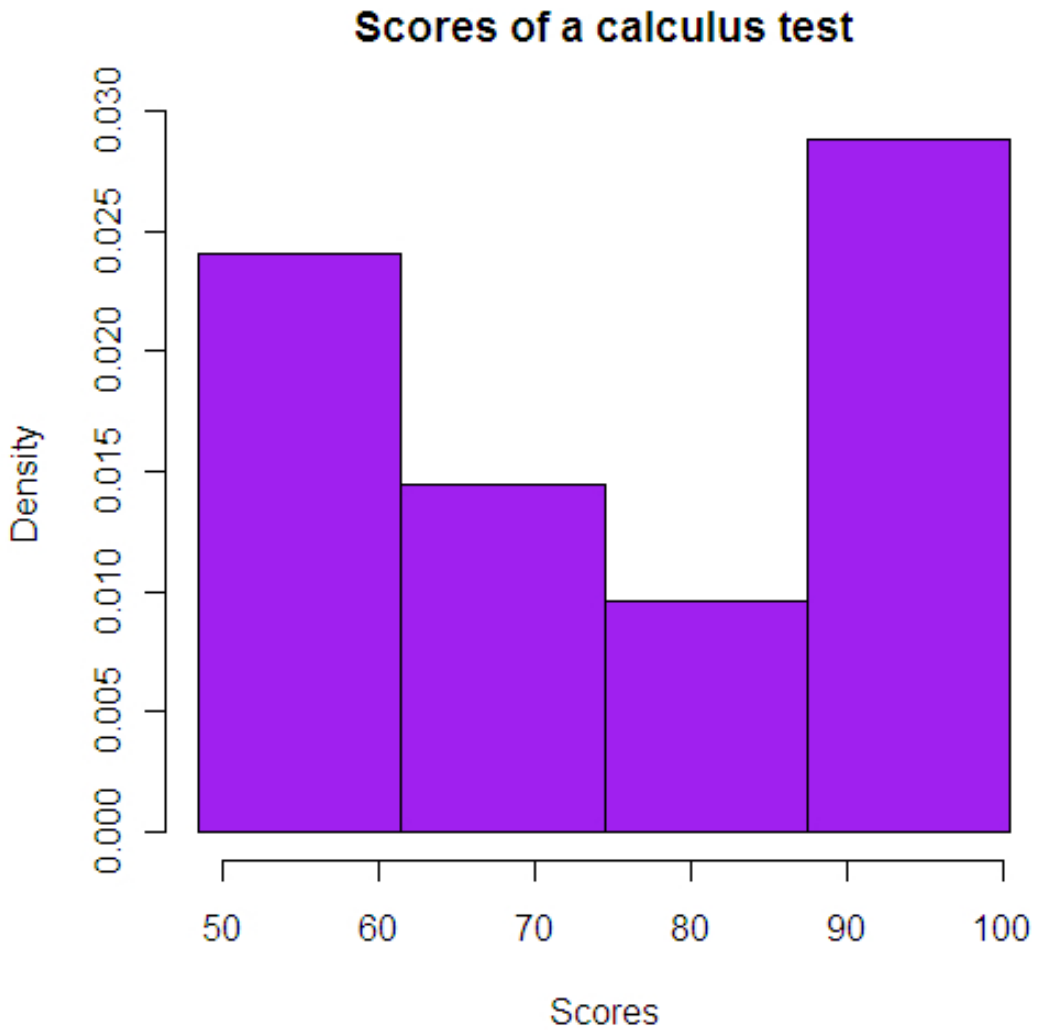
Solution

Class interval	*Frequency,* f_i	$h(x)$	*Class marks*
$(48.5, 61.5]$	5	$\frac{5}{208} \approx 0.024$	55
$(61.5, 74.5]$	3	$\frac{3}{208} \approx 0.014$	68
$(74.5, 87.5]$	2	$\frac{2}{208} \approx 0.010$	81
$(87.5, 100.5]$	6	$\frac{6}{208} \approx 0.029$	94

Solution continue, frequency histogram



Solution continue, relative frequency histogram



Solution continue, relative frequency histogram

Notice that $\left(\frac{5}{208}+\frac{3}{208}+\frac{2}{208}+\frac{6}{208}\right)\cdot 13=1$.

The fraction of students that score below 74 is $\frac{8}{16}=0.5$.

Usage of the histogram:

- Look for overall pattern in the data and deviations from the patterns.

Overall patterns:

- the shape (symmetry, number of peaks, skewness)
- the center (sample mean, median)
- the spread (standard deviation, variance, quantiles)

Deviations from the patterns:

- outliers
- gaps

Histogram

Skewness:

- A distribution is skewed to the right if the right tail of the histogram is much longer than the left tail. Similar definition for left skewness.

Measuring center:

- the sample mean
- the median

Sample mean

The sample mean:

- For n observations, $x_1, x_2, ..., x_n$, the sample mean is the arithmetic average,

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Sample mean

The sample mean:

- For n observations, $x_1, x_2, ..., x_n$, the sample mean is the arithmetic average,

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- Alternatively: If x occurs $r(x)$ times, we have

$$\bar{x} = \frac{1}{n}\sum_{x} xr(x) = \sum_{x} xp(x),$$

where $p(x) = \frac{r(x)}{n}$, where $p(x)$ is the proportions of observations of X.

Sample mean, example

Example

Suppose the scores of a calculus test in a class of 16 students where distributed as follows:
$49, 50, 52, 52, 61, 67, 73, 73, 82, 82, 90, 94, 99, 99, 99, 100$.
Determine the mean, $\bar{x}$.

Solution

Add the numbers together and divide by $n = 16$.

$$\bar{x} = \frac{49 + 50 + 52 + 52 + \cdots + 99 + 99 + 99 + 100}{16}$$

$$= \frac{(1)(49) + (1)(50) + (2)(52) + (1)(61) + \cdots + (3)(99) + (1)(100)}{16}$$

$$\approx 76.38$$

Median

The median:

- The median is the midpoint of the distribution after the data are arranged in order from smallest to largest,
  $x_1, x_2, ..., x_k, ..., x_n$.
  x_k is called the k^{th}-order statistics.

Median

The median:

- The median is the midpoint of the distribution after the data are arranged in order from smallest to largest,
  $x_1, x_2, ..., x_k, ..., x_n$.
  x_k is called the k^{th}-order statistics.
- If the number of observations, n, is odd, the median is the center, $x_{(\frac{n+1}{2})}$.

Median

The median:

- The median is the midpoint of the distribution after the data are arranged in order from smallest to largest,
  $x_1, x_2, ..., x_k, ..., x_n$.
  x_k is called the k^{th}-order statistics.
- If the number of observations, n, is odd, the median is the center, $x_{(\frac{n+1}{2})}$.
- If the number of observations, n, is even, the median is the average of the two center observations, $\frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$.

Median, example

Example

Suppose the scores of a calculus test in a class of 16 students where distributed as follows:
$49, 50, 52, 52, 61, 67, 73, 73, 82, 82, 90, 94, 99, 99, 99, 100$.
Determine the median.

Solution

The number of observations, $n = 16$, *is even, so the median is*

$$\frac{1}{2}(x_{(\frac{16}{2})} + x_{(\frac{16}{2}+1)}) = \frac{1}{2}(x_8 + x_9) = \frac{1}{2}(73 + 82) = 77.5.$$

Median and mean

- If the distribution is exactly symmetric, the mean and the median are the same.
- The mean is usually larger than the median if the distribution is skewed to the right.
- The mean is usually smaller than the median if the distribution is skewed to the left.

Steam and leaf display

In the previous example consider the number 49:

- 4 is the steam
- 9 is the leaf

```
 4 | 9
 5 | 022
 6 | 17
 7 | 33
 8 | 22
 9 | 04999
10 | 0
```

Figure: Ordered Steam and leaf display

Measuring spread, the variance

The sample variance:

- $$Var(x) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$
- Notice that we divide by $(n-1)$ instead of by n. More about this later.
- We have $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$. It follows that $(n-1)$ of the terms can vary freely and the last one depends on those $(n-1)$ terms.
- $(n-1)$ is called the degree of freedom of the variance or the standard deviation.

Measuring spread, the standard deviation

The sample standard deviation:

- $$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$
- s measures the spread about the mean.
- $s = 0$, when no spread.
- s gets larger as the observations become more spread out about their mean.
- s has the same units as the originally observations.

The variance, example

Example

Suppose the scores of a calculus test in a class of 16 students where distributed as follows:
$49, 50, 52, 52, 61, 67, 73, 73, 82, 82, 90, 94, 99, 99, 99, 100.$
Find the variance and the standard deviation.

Solution

$$var(x)$$
$$= \frac{1}{15}[(49 - 76.38)^2 + (50 - 76.38)^2 + \cdots + (100 - 76.38)^2] \approx 375.58$$

$$s = \sqrt{375.58} = 19.38.$$

Order Statistics

- If $0 < p < 1$, the $(100p)^{th}$ sample percentile has approximately (np) sample observations less than it and $n(1-p)$ observations greater than it.

Definition

- If $(n+1)p$ is an integer, take the $(100p)^{th}$ sample percentile as the $(n+1)p^{th}$ order statistics.
- If $(n+1)p$ is not an integer, but can be written a sum of an integer,r, plus a fraction, $\frac{a}{b}$, then define the $(100p)^{th}$ sample percentile as

$$\tilde{\pi}_p = x_r + \frac{a}{b}(x_{r+1} - x_r) = (1 - \tfrac{a}{b})x_r + \frac{a}{b}x_{r+1}.$$

Order Statistics, example

- The median is the 50^{th} percentile (2 quartile) with $p = 0.5$

Example

Suppose the scores of a calculus test in a class of 16 students where distributed as follows:
$49, 50, 52, 52, 61, 67, 73, 73, 82, 82, 90, 94, 99, 99, 99, 100$.
Determine the first quartile (25^{th} sample percentile), the second quartile (median), and the third quartile (75^{th} sample percentile).

Solution

First quartile: $p = 0.25$ *and* $(n+1)p = 17(0.25) = 4.25$.
Hence

$$\tilde{q}_1 = \tilde{\pi}_{0.25} = 0.75x_4 + 0.25x_5 = (0.75)(52) + (0.25)(61) = 54.25.$$

Order Statistics, example continue

Solution

Second quartile:

$$\tilde{q}_2 = \tilde{\pi}_{0.5} = 0.5x_8 + 0.5x_9 = (0.5)(73) + (0.5)(82) = 77.5.$$

Third quartile:
$p = 0.75$ *and* $(n+1)p = 17(0.75) = 12.75.$
Hence

$$\tilde{q}_3 = \tilde{\pi}_{0.75} = 0.25x_{12} + 0.75x_{13} = (0.25)(94) + (0.75)(99) = 97.75.$$

Five-number summary

The five-number summary

minimum first quartile median third quartile maximum

- The five-number summary is a better representation of the data than the mean and the standard deviation for skewed distributions or for distributions with strong outliers.
- R-code:

```
> x=c(49,50,52,52,61,67,73,73,82,82,90,94,99,99,99,100)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  49.00   58.75   77.50   76.38   95.25  100.00
```

Box-and-whisker diagram (Boxplot)

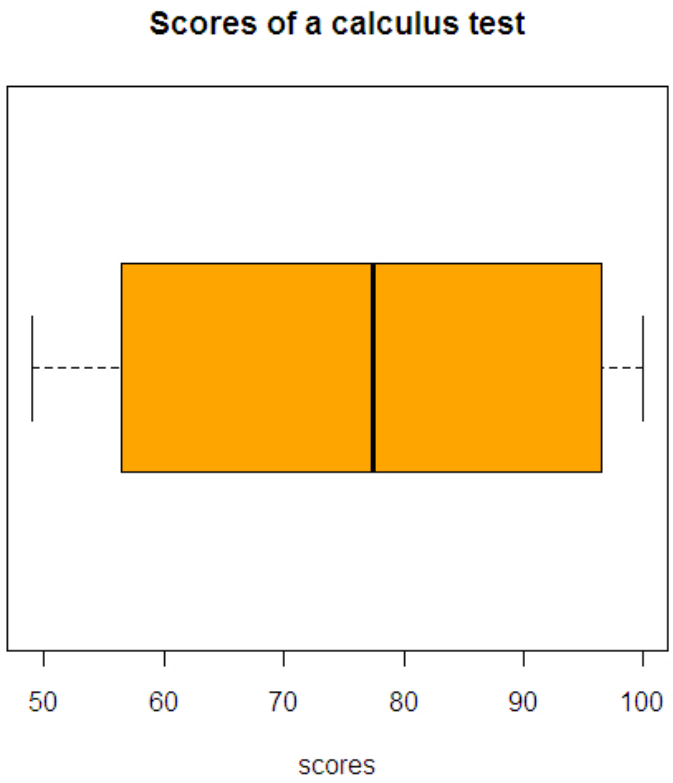
Definition

Interquartile range (IQR): $\tilde{q}_3 - \tilde{q}_1$.

Horizontal boxplot:

- The data is displayed on the horizontal axis.
- Draw a rectangular box with the left side drawn at $\tilde{q}_1$ and the right side at $\tilde{q}_3$ with a vertical line segment at $\tilde{q}_2$.
- Draw a left whisker as a horizontal line segment from the minimum to the midpoint of the left side of the box.
- Draw a right whisker as a horizontal line segment from the midpoint of right side of the box to the maximum.
- The length of the box is equal to IQR.
- The left whisker represents the first quarter of the data.
- The right whisker represents the fourth quarter of the data.

Boxplot, previous example



The longer whisker is to the left, so the data is slightly skewed to the left.

Outliers

- An outlier is an observation that deviates markedly from the other observations in the sample.
- An outlier can occur by chance. (Use statistics that are robust to outliers.)
- An outlier can result from measurements error.
- An outlier can indicate a heavy-tailed distribution.
- An observation is an outlier if it is $1.5 \cdot IQR$ above the third quartile or $1.5 \cdot IQR$ below the first quartile

Outliers

Example

Suppose the scores of a calculus test in a class of 16 students where distributed as follows:
$49, 50, 52, 52, 61, 67, 73, 73, 82, 82, 90, 94, 99, 99, 99, 100$.
Determine if there are outliers.

Solution

Using the quartiles computed in the R:
$IQR = \tilde{q}_3 - \tilde{q}_1 = 95.25 - 58.75 = 36.5$
$1.5 \cdot IQR = 54.75$
$\tilde{q}_1 - (1.5 \cdot IQR) = 58.75 - 54.75 = 4.$
$\tilde{q}_3 + (1.5 \cdot IQR) = 95.25 + 54.75 = 120$ *so there are no outliers in the data.*

Outliers

- $\bar{x}$ and s are sensitive to outliers.
- The median and the five-number summary are resistant to outliers. (These measures does not change with changes in the extreme observations).
- Truncated mean: Discard the lower and upper $(p \cdot 100)$ percent of the data and take the average of the rest. p is the fraction of observations to be discarded from the lower and upper end and is a number between 0 and 0.5.
- To compute the truncated mean in R: $mean(x, trim = p)$.
- Median: Truncated mean with $p = 0.5$.

The following table shows the average math SAT scores in 2010 for states in the south-west.

State	Average math SAT score
Arizona	525
California	516
Colorado	572
Nevada	501
New Mexico	549
Oklahoma	568
Texas	505
Utah	559

(Source: Public Agenda)

The following table shows the average math SAT scores in 2010 for states in the north-east.

State	Average math SAT score
Connecticut	514
Maine	467
Massachusetts	526
New Hampshire	524
Rhode Island	495
Vermont	521
New Jersey	514
New York	499
Pennsylvania	501

(Source: Public Agenda)

Boxplot, R-code

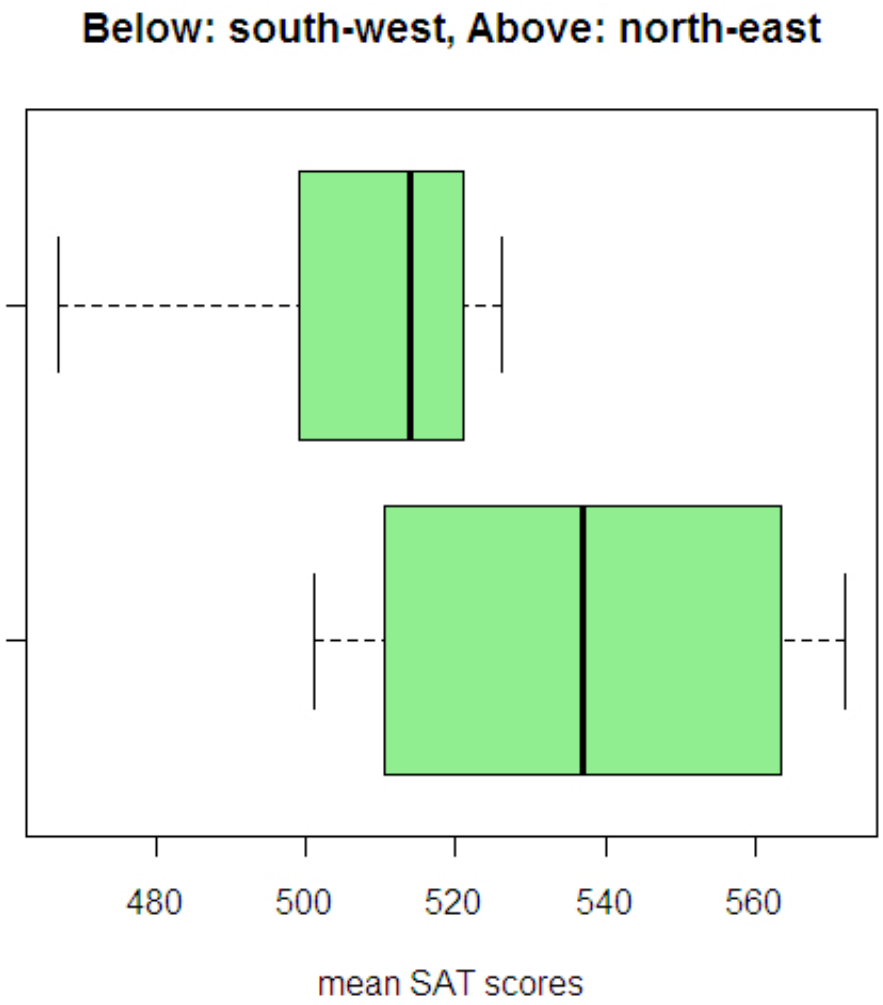
R-code:

$>$ $sw = c(525, 516, 572, 501, 549, 568, 505, 559)$

$>$ $ne = c(514, 467, 526, 524, 495, 521, 514, 499, 501)$

$>$ boxplot(sw,ne,horizontal=TRUE,main="Below: south-west, Above: north-east", xlab="mean SAT scores", col="lightgreen")

SAT-scores


Below: south-west, Above: north-east
480
500
520
540
560
mean SAT scores