M2S1 - Numerical data

Professor Jarad Niemi

STAT 226 - Iowa State University

August 29, 2018

Outline

- Summary statistics
  - Measures of location
    - Mean
    - Median
    - Quartiles
    - Minimum/maximum
  - Measures of spread
    - Range
    - Interquartile range
    - Variance
    - Standard deviation
  - Robustness
- Boxplots

Numerical variables

Definition

A numerical, or quantitative, variable take numerical values for which arithmetic operations such as adding and averaging make sense.

Examples:

- height/weight of a person
- temperature
- time it takes to run a mile
- currency exchange rates
- number of webpage hits in an hour

For numerical variables, we also consider whether the variable is a count and whether or not that count has a technical upper limit.

Toyota Sienna Gas Mileage data set

	date	fuel	cost	miles	ethanol	octane	mpg
248	2018-07-02	13.185	35.59	291.0	0	87	22.07053
249	2018-07-05	14.865	35.66	326.4	0	87	21.95762
250	2018-07-11	17.542	49.10	370.9	0	87	21.14354
251	2018-07-13	17.563	47.40	366.1	10	87	20.84496
252	2018-07-19	12.895	33.90	239.5	10	87	18.57309
253	2018-07-19	6.664	18.12	146.6	0	87	21.99880
254	2018-07-19	7.894	22.10	190.8	0	87	24.17026
255	2018-07-22	10.322	27.86	197.3	10	87	19.11451
256	2018-07-22	6.859	18.24	145.5	10	87	21.21300
257	2018-07-22	6.778	18.43	147.7	0	87	21.79109
258	2018-07-23	7.449	18.99	154.3	10	87	20.71419
259	2018-07-28	8.762	24.09	157.2	10	87	17.94111
260	2018-08-07	12.043	33.23	259.4	10	87	21.53948
261	2018-08-10	11.388	31.08	231.0	10	87	20.28451
262	2018-08-10	6.455	17.42	147.1	0	87	22.78854

Summary statistics

Definition

A summary statistic is a numerical value calculated from the sample.

- Measures of location: mean, median, quartiles, minimum/maximum
- Measures of spread: range, interquartile range, variance, standard deviation

Sample mean

Definition

The sample mean of a set of observations $y_1, y_2, \ldots, y_n$ is the arithmetic average of all observations:

$$\overline{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

where $\sum$ is the summation sign.

Example

The number of sick days employees took during the past year in a small local business is 0,1,2,0,4,0,1,2,3,2. The sample mean of these observations is

$$\overline{y} = \frac{0+1+2+0+4+0+1+2+3+2}{10} = 1.5 \text{ days.}$$

Sample mean is not robust

Example

The number of sick days employees took during the past year in a small local business is 0,1,2,0,4,0,1,2,3,60. The sample mean of these observations is

$$\overline{y} = \frac{0+1+2+0+4+0+1+2+3+60}{10} = 7.3 \text{ days.}$$

Definition

A summary statistic is robust if the value of the statistic does not change very much with a (possibly large) change in a small number of observations.

The sample mean is not robust.

Sample median

Definition

The sample median corresponds to the value of the data that is in the middle when all observations are ordered from smallest to largest. If there are two such observations, their arithmetic average is the median.

Example

The number of sick days employees took during the past year in a small local business is 0,1,2,0,4,0,1,2,3,2. The ordered observations are 0,0,0,1,1,2,2,2,3,4 and the median is

$$\frac{1+2}{2} = 1.5 \text{ days.}$$

Sample median is robust

Example

The number of sick days employees took during the past year in a small local business is 0,1,2,0,4,0,1,2,3,60. The ordered observations are 0,0,0,1,1,2,2,3,4,60 and the median is

$$\frac{1+2}{2} = 1.5 \text{ days.}$$

The sample median is robust.

Quartiles

Definition

The sample quartiles (Q1,Q2,Q3) are the 3 numbers that divide the ordered observations into 4 equally sized groups, i.e. each group contains 25% of all observations.

- The first quartile, Q1, is the 25th percentile and the median of the observations below the sample median.
- The second quartile, Q2, is the 50th percentile and the sample median.
- The third quartile, Q3, is the 75th percentile and the median of the observations above the sample median.

Example

The number of sick days employees took during the past year in a small local business is 0,1,2,0,4,0,1,2,3,2. The ordered observations are 0,0,0,1,1,2,2,2,3,4. The second quartile (median) is 1.5 days, the first quartile is 0 days, and the third quartile is 2 days.

5-number summary

Definition

A (typical) 5-number summary consists of the following measures

Minimum Q1 Median Q3 Maximum

Example

The number of sick days employees took during the past year in a small local business is 0,1,2,0,4,0,1,2,3,2. The ordered observations are 0,0,0,1,1,2,2,2,3,4. The 5-number summary is 0, 0, 1.5, 2, and 4 days.

Let software find this for you

For the Toyota Sienna miles per gallon data set, we have

```
mean(mpg)

[1] 19.31347

min(mpg); max(mpg)

[1] 8.508946
[1] 39.08611

quantile(mpg, c(.25,.5,.75), type=2)

     25%      50%      75%
17.35947 19.29787 21.33436

summary(mpg)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  8.509  17.359  19.298  19.313  21.334  39.086
```

Measures of spread

Measures of location:

- Mean
- Median
- Quartiles
- Minimum/maximum

Measures of spread:

- Range
- Interquartile range
- Variance
- Standard deviation

Measures of spread are 0 if the data are all identical and increase as the data become more variable.

Range

Definition

The range is the maximum minus the minimum.

Example

The number of sick days employees took during the past year in a small local business is 0,1,2,0,4,0,1,2,3,2. The minimum is 0 days, the maximum is 4 days, and the range is 4 - 0 = 4 days.

Interquartile range

Definition

The interquartile range is Q3 minus Q1.

Example

The number of sick days employees took during the past year in a small local business is 0,1,2,0,4,0,1,2,3,2. The Q1 is 0 days, Q3 is 2 days, and the interquartile range is 2 - 0 = 2 days.

Sample variance

Definition

The sample variance is

$$s^2 = \frac{(y_1 - \overline{y})^2 + (y_2 - \overline{y})^2 + \cdots + (y_n - \overline{y})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

The units are squared.

Example

The number of sick days employees took during the past year in a small local business is 0,1,2,0,4,0,1,2,3,2. The sample mean is 1.5 and the sample variance is

$$s^2 = \frac{(0 - 1.5)^2 + (1 - 1.5)^2 + \cdots (2 - 1.5)^2}{10 - 1} = 1.8\overline{3} \text{ days}^2.$$

Sample standard deviation

Definition

The sample standard deviation is the square root of the sample variance, i.e.

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

The units are normal.

Example

The number of sick days employees took during the past year in a small local business is 0,1,2,0,4,0,1,2,3,2. The sample variance is $1.8\overline{3}$ and the sample standard deviation

$$s = \sqrt{1.8\overline{3}} \approx 1.354 \text{ days.}$$

Let software find this for you

For the Toyota Sienna miles per gallon data set, we have

```
diff(range(mpg))

[1] 30.57717

diff(quantile(mpg, c(.25,.75), type=2))

     75%
3.974883

var(mpg)

[1] 8.871431

sd(mpg)

[1] 2.978495
```

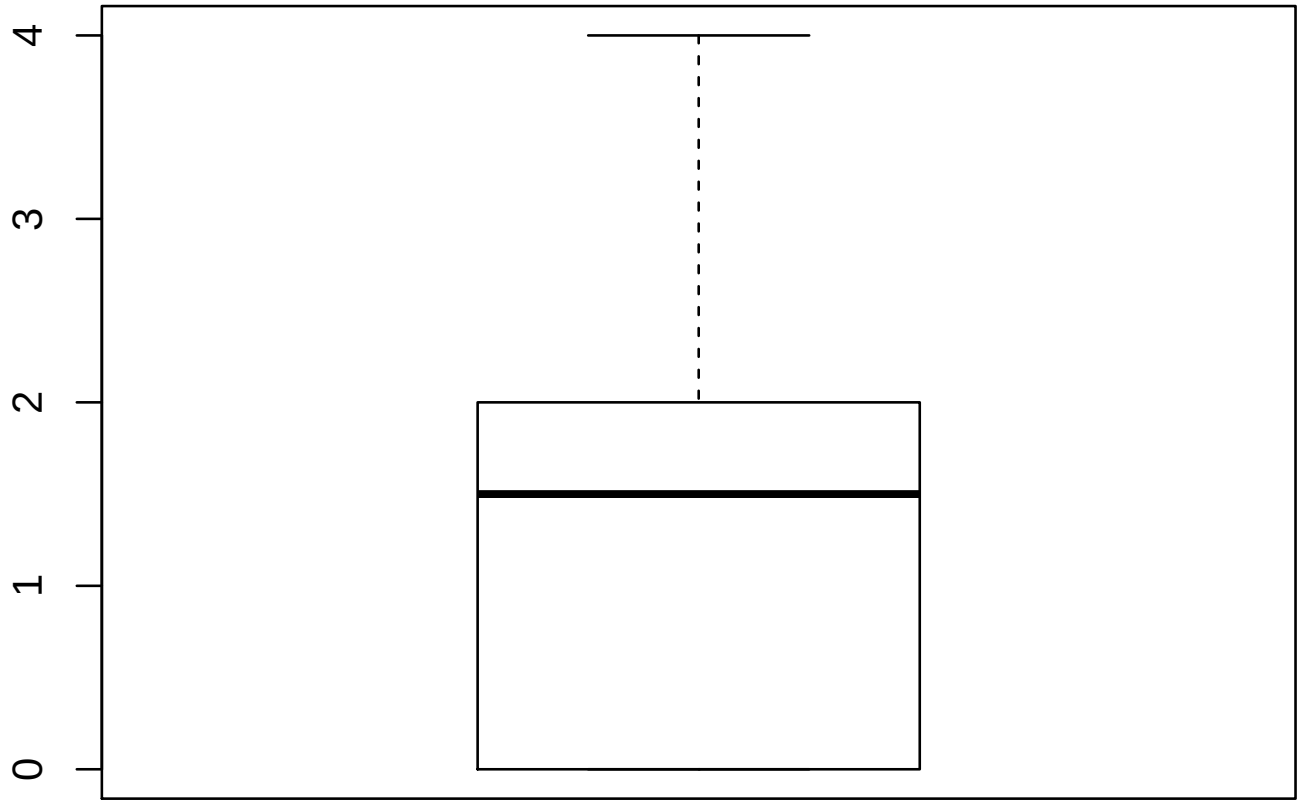
Boxplot

Definition

A boxplot is a graphical representation of the 5-number summary. A boxplot is typically constructed like this

- A box with endpoints at Q1 and Q3 with a line in the middle at Q2 (median).
- Whiskers that extend out to
  - Q1-1.5IQR on the low side and
  - Q3+1.5IQR on the high side.
- Dots for points beyond these whiskers.

Sick days boxplot



Miles per gallon boxplot