STAT22000 Autumn 2013 Lecture 2

Yibi Huang

October 2, 2013

1.2 Numerical Descriptions of Data

- Mean and Median
- Five Number Summary, IQR
- Boxplots
- Standard Deviation (SD)

Mean

The **mean** of a set of observations, $x_1, x_2, \ldots, x_n$, is the arithmetic average of the observations:

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Say the age of 9 individuals are

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
43	35	43	33	38	53	64	27	34

The mean age of the 9 people is given by:

$$\overline{x} = \frac{43+35+43+33+38+53+64+27+34}{9} = \frac{370}{9} = 41.11.$$

Median

For a list of numbers, the **median** is a number such that half of the list are smaller than it and half of the list are larger than it. How to find the median of a list of numbers $x_1, x_2, \ldots, x_n$?

1. Sort the list from the smallest to the largest:

$$x_{\min} = x_{(1)} \le x_{(2)} \le x_{(3)} \le \cdots \le x_{(n-1)} \le x_{(n)} = x_{\max},$$

where $x_{(i)}$ is the ith smallest number in the list

2. $$\text{the median} = \begin{cases} x_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \dfrac{x_{(n/2)} + x_{(n/2+1)}}{2} & \text{if } n \text{ is even} \end{cases}$$

Example 1: Say the age of 9 individuals are

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
43	35	43	33	38	53	64	27	34

sorted from the smallest to the largest:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$
27	33	34	35	38	43	43	53	64

The median is $x_{((n+1)/2)} = x_{(5)} = 38$.

Example 2: Say the age of 10 individuals are

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
43	35	43	33	38	53	64	27	34	27

sorted from the smallest to the largest:

$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
27	27	33	34	35	38	43	43	53	64

The median is

$$\frac{x_{(n/2)} + x_{(n/2+1)}}{2} = \frac{x_{(5)} + x_{(6)}}{2} = \frac{35+38}{2} = 36.5.$$

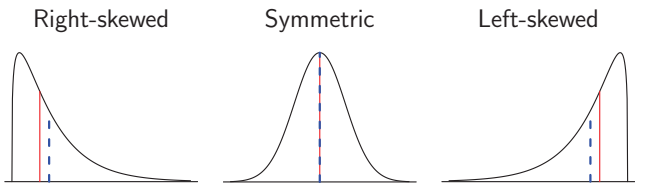
Finding Mean and Median in R

```
> x = c(43,35,43,33,38,53,64,27,34)
> mean(x)
[1] 41.11111
> sort(x)    # to sort x from min to max
[1] 27 33 34 35 38 43 43 53 64
> median(x)
[1] 38

> # vector y is vector x append by a number 27
> y = c(x,27)
> y
 [1] 43 35 43 33 38 53 64 27 34 27
> mean(y)
[1] 39.7
> sort(y)
 [1] 27 27 33 34 35 38 43 43 53 64
> median(y)
[1] 36.5
```


Mean vs. Median (1)

- The more symmetric the distribution, the closer the mean to the median.
  If exactly symmetric, then the mean = the median.
- In a skewed distribution, the mean is pulled toward the longer tail.



(solid line — median, dashed line — mean)

Mean vs. Median (2)

The **mean** is the point m that the sum of the squared distance to all data values is minimized

$$\sum_{i=1}^{n}(x_i - m)^2$$

The **median** is the point m that the sum of the absolute distance to all data values is minimized

$$\sum_{i=1}^{n}|x_i - m|$$

Robustness of the Median

Consider the list $-2, -1, 0, 0, 2, 4$. If the number '2' in the list is miss recorded as 20,

- The mean is increased by $\frac{20-2}{6} = 3$.
- The median is unaffected.

Median is more resistent, i.e., less sensitive to extreme values or outliers than the mean. We say the median is more **robust**.

- Example: Housing sales price in Hyde Park

	Mean	Median
Jun – Aug, 2011	\$525,384	\$227,000
Jun – Aug, 2013	\$423,528	\$291,750

Source: http://www.trulia.com/home_prices/Illinois/Chicago-heat_map/

Quartiles, IQR, Five-Number Summary

- **Quartiles** divide data into 4 even parts
  - **first quartile Q_1**
    = median of all observations below the median
  - **second quartile Q_2** = median
  - **third quartile Q_3**
    = median of all observations above the median
- **Interquartile Range:** $IQR = Q_3 - Q_1$
- **Five-Number Summary**:

  $\min = x_{(1)}, \quad Q_1, \quad \text{Median}, \quad Q_3, \quad x_{(n)} = \max$

Example 1

For the 9 numbers: 43, 35, 43, 33, 38, 53, 64, 27, 34

43	27	⎫	
35	33	⎬ ← median of this half	$= \frac{33+34}{2} = 33.5 = Q_1$
43	34	⎭	
33	35		
38	$\xrightarrow{\text{sort}}$ 38	← overall median	$= Q_2$
53	43	⎫	
64	43	⎬ ← median of this half	$= \frac{43+53}{2} = 48 = Q_3$
27	53	⎭	
34	64		

- IQR $= Q_3 - Q_1 = 48 - 33.5 = 14.5$
- Five number summary: 27, 33.5, 38, 48, 64

Example 2

For the 10 numbers: 43, 35, 43, 33, 38, 53, 64, 27, 34, 27

43	27	⎫	
35	27		
43	33	⎬ ← median of this half	$= 33 = Q_1$
33	34		
38	$\xrightarrow{\text{sort}}$ 35	← overall median	$= \frac{35+38}{2} = 36.5 = Q_2$
53	38		
64	43		
27	43	⎬ ← median of this half	$= 43 = Q_3$
34	53		
27	64	⎭	

- IQR $= Q_3 - Q_1 = 43 - 33 = 10$
- Five number summary: 27, 33, 36.5, 43, 64

Finding Quarters in R (1)

In fact, there are several formulas for quartiles, varying from book to book, software to software.
E.g., for the 9 numbers in Example 1

```
> x = c(43,35,43,33,38,53,64,27,34)
```

the formula in [IPS7e] gives $Q_1 = 33.5$, $Q_3 = 48$, but R gives $Q_1 = 34$, $Q_3 = 43$.

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.00   34.00   38.00   41.11   43.00   64.00
> fivenum(x)
[1] 27 34 38 43 64


> IQR(x)
[1] 9
```

Finding Quarters in R (2)

Sometimes even different commands in R give different quartiles.

E.g., for the 10 numbers in Example 2, the formula in [IPS7e] gives $Q_1 = 33$, $Q_3 = 43$, but

```
> y = c(43,35,43,33,38,53,64,27,34,27)
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  27.00   33.25   36.50   39.70   43.00   64.00
> fivenum(y)
[1] 27.0 33.0 36.5 43.0 64.0
> IQR(y)
[1] 9.75
```

Don't worry about the formula. Just keep in mind that

quartiles divide data into 4 even parts

For describing data, just report the values that your software gives.

Boxplot (aka. Box and Whiskers Plot)

25	6	6.1
24	5	5.6
23	4	5.3
22	3	4.9
21	2	4.7
20	1	4.5
19	6	4.2
18	5	4.1
17	4	3.9
16	3	3.8
15	2	3.7
14	1	3.6
13		3.4
12	6	3.3
11	5	2.9
10	4	2.8
9	3	2.5
8	2	2.3
7	1	2.3
6	6	2.1
5	5	1.5
4	4	1.9
3	3	1.6
2	2	1.2
1	1	0.6

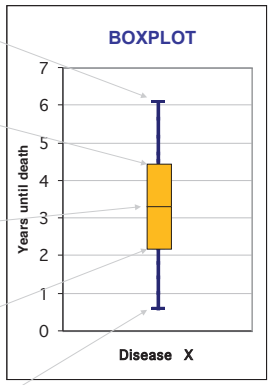
Largest = max = 6.1

Q_3= third quartile = 4.35

M = median = 3.4

Q_1= first quartile = 2.2

Smallest = min = 0.6



Five-number summary: min Q_1 M Q_3 max

1.5 IQR Rule for Suspected Outliers

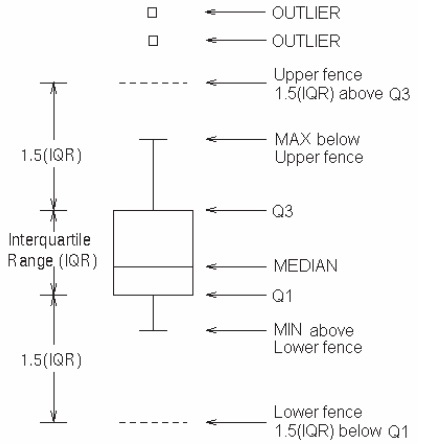
An observation is a *suspected* outlier if it lies more than $1.5 \times$ IQR below Q_1 above above Q_3.

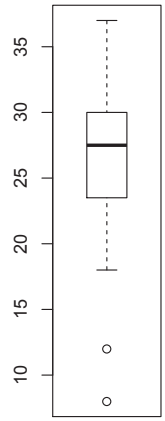
How to deal with suspected outliers:

- First, investigate their causes. If it is a *mistake* (like recording error or mistake in experiments), you can
  - correct outliers if possible
  - delete them
- If no clear reason to drop outliers, you may
  - use resistant methods, e.g., report the median rather than the mean, so the conclusions are less affected by outliers
  - analyze the data both w/ and w/o the outliers and then see how much the result is affected by the outliers
  - ...

Modified Boxplot and Whiskers Plot

Most softwares give modified boxplot rather than boxplot.





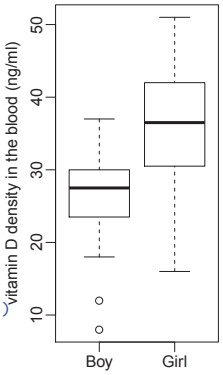
Side by Side Boxplots

Just like back-to-back stemplots, boxplots of related distribution are often placed side-by-side for comparison. E.g., for the vitamin D levels of teenager boys and girls data in Lecture 1

```
> vboys = c(18,28,28,28,37,31,24,29,8,
  27,24,12,21,32,27,24,23,33,31,29)
> vgirls = c(16,43,38,48,42,23,36,35,37,
  34,25,28,26,43,51,33,40,35,41,42)

> # boxplot for boys only
> # the plot is in the previous page
> boxplot(vboys)

> # side-by-side boxplots for boys and girls
> boxplot(vboys,vgirls,names=c("Boy","Girl"),
ylab="vitamin D density in the blood (ng/ml)")
```



Variance and Standard Deviation

Suppose there are n observations $x_1, x_2, \ldots, x_n$.
The (sample) **variance** of the n observations is:

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$$

This is (approximately) the average of the squared distances of the observations from the mean.
The (sample) **standard deviation** (SD) is:

$$s = \sqrt{s^2} = \sqrt{\text{Variance}}$$

Example: Find the variance and SD of the list 1, 2, 2, 7.

- The average of the list is $\frac{1+2+2+7}{4} = 3$

$$\text{Variance} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(1-3)^2 + (2-3)^2 + (2-3)^2 + (7-3)^2}{4-1} = \frac{22}{3}$$

- $\text{SD} = \sqrt{\text{Variance}} = \sqrt{\frac{22}{3}} \approx 2.708$

An alternative way to find the variance is

$$\text{Variance} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1} = \frac{1^2 + 2^2 + 2^2 + 7^2 - 4 \times 3^2}{4-1} = \frac{1+4+4+49-36}{3} = \frac{22}{3}.$$

Why Squared Distances? The exact average of the distances of the observations from the mean is the **mean absolute deviation** (MAD)

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

MAD is less commonly used than SD because the absolute value function is not as easy to work with algebraically as the square function, e.g., the absolute value function is not differentiable.

Why Dividing by $n-1$? Not n? Note the sum of the deviations is always zero

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

Thus, if you know any $n-1$ of the deviations, the last deviation can be determined from the others. The number of "freely varying" deviations, $n-1$ in this case, is called the **degrees of freedom**.

Properties of Standard Deviation (SD)

- SD measures spread about the mean and should be used only when the mean is the measure of center.
- When SD $= 0$, what are the observations look like?
- and what if SD < 0?
- SD is NOT resistant to outliers.
- SD has the same units of measurement as the original observations, while the varianceis in the square of these units.

Choosing Among Summary Statistics

- Unimodal, symmetric distribution w/o outliers
  - Use mean and SD
- Unimodal, skewed distribution (w/ or w/o outlier)
  - Use 5-number summary, Boxplots
- Multimodal (i.e. clustered) distribution
  - Use histograms or stemplots

Always check the plot of your data: numerical measures of center and spread can not describe the entire shape of distribution.

Linear Transformations $y = a + bx$

- Adding a constant to each observation: $y_i = x_i + a$
  - $\bar{y} = \bar{x} + a$, $s_y = s_x$
  - changes mean but not SD
- Multiplying each observation by a constant: $y_i = bx_i$
  - $\bar{y} = b\bar{x}$, $s_y = |b|s_x$.
  - scales both mean and SD
- General linear transformation: $y_i = a + bx_i$
  - $\bar{y} = a + b\bar{x}$
  - $\text{SD}_y = |b| \times \text{SD}_x$
  - Linear transformations do not change the basic shape of a distribution (skew or symmetric, number of modes), just change the center and spread.

Exercises

Find the mean and SD for each of the following lists of numbers.

(a) 1, 2, 2, 7

Mean $= 3$, SD $= \sqrt{\frac{22}{3}} \approx 2.708$ (See slide 2-19)

(b) 21, 22, 22, 27 (= list (a) +20)

Mean $= 3 + 20 = 23$, SD $= 2.708$ (unchanged)

(c) 10, 20, 20, 70 (= list (a) $\times 10$)

Mean $= 3 \times 10 = 30$, SD $= 2.708 \times 10 = 27.08$

(d) -1, -2, -2, -7 (= list (a) $\times(-1)$)

Mean $= 3 \times (-1) = -3$, SD $= 2.708 \times |-1| = 2.708$