MA331 Intermediate Statistics

Lecture 02 Descriptive Statistics [1]

Xiaohu Li

Department of Mathematical Sciences
Stevens Institute of Technology
Hoboken, New Jersey 07030

Weeks 01-02

[1] Based on Chapters 2 and 3.

1. Numerical measures of location: mean and median

For a data $x_1, \cdots, x_n$,

✌ the mean is the arithmetic average of the data, it measures the algebraic center of all observations.

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

✌ the order statistic is the increasing arrangement of the data, denoted as

$$x_{1,n} \le x_{2,n} \le \cdots \le x_{n,n}.$$

✌ the median gives the middle point of the data, i.e.,

$$x_{(n+1)/2,n} \quad \text{for odd } n \qquad \text{or} \qquad \frac{x_{n/2,n} + x_{n/2+1,n}}{2} \quad \text{for even } n,$$

which is the geometric center of all observations.

Query 1: In R mean(x) and median(x) outputs the mean and median of the sample x, respectively. What does order(x) tell us? How do you get the order statistics of x?

2. Numerical measures of location: an example

TABLE 1.8 Fuel economy (miles per gallon) for model year 2001 cars

Minicompact cars			Two-seater cars		
Model	City	Highway	Model	City	Highway
Audi TT Coupe	22	31	Acura NSX	17	24
BMW 325CI Convertible	19	27	Audi TT Roadster	22	30
BMW 330CI Convertible	20	28	BMW Z3 Coupe	21	28
BMW M3 Convertible	16	23	BMW Z3 Roadster	20	27
Jaguar XK8 Convertible	17	24	BMW Z8	13	21
Jaguar XKR Convertible	16	22	Chevrolet Corvette	18	26
Mercedes-Benz CLK320	20	28	Dodge Viper	11	21
Mercedes-Benz CLK430	18	24	Ferrari Modena	11	16
Mitsubishi Eclipse	22	30	Ferrari Maranello	8	13
Porsche 911 Carrera	17	25	Honda Insight	61	68
Porsche 911 Turbo	15	22	Honda S2000	20	26
			Lamborghini Diablo	10	13
			Mazda Miata	22	28
			Mercedes-Benz SL500	16	23
			Mercedes-Benz SL600	13	19
			Mercedes-Benz SLK320	21	27
			Plymouth Prowler	17	23
			Porsche Boxster	19	27
			Toyota MR2	25	30

- The mean highway mileage for 19 two-seaters $\frac{24+30+\cdots+30}{19} = 25.8\text{mpg}$.
- Mean is sensitive to the outlier or extremal observations. Revising 68mpg as 30 reduces the mean mileage to 23.8mpg.
- The median highway mileage for 19 two-seaters

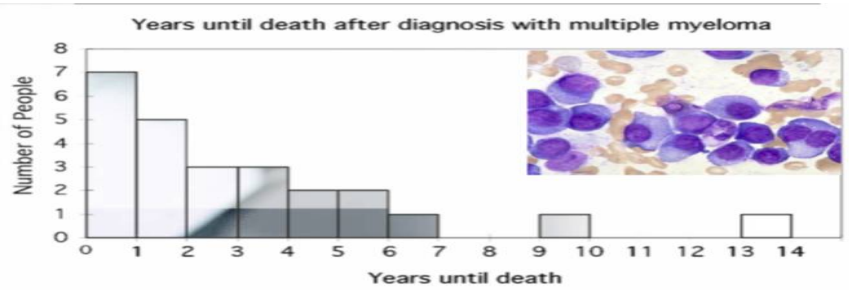
$x_{1,19}$	$\cdots$	$x_{8,19}$	$x_{9,19}$	$x_{10,19}$	$x_{11,19}$	$x_{12,19}$	$\cdots$	$x_{19,19}$
13	$\cdots$	23	24	26	26	27	$\cdots$	30

- Median is less sensitive to the outlier. Revising 30mpg as 32 doesn't change the median mileage.

3. Numerical measures of location: symmetry

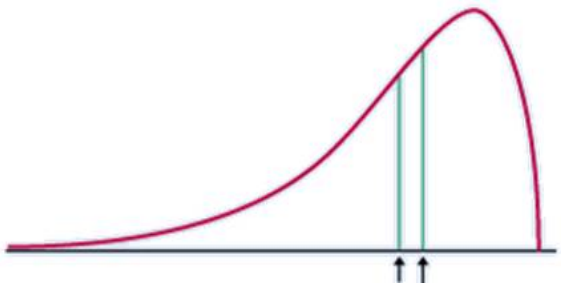
✍ Mean and median are equal only if the distribution is symmetric.

(a) Symmetric mean = median =3.4

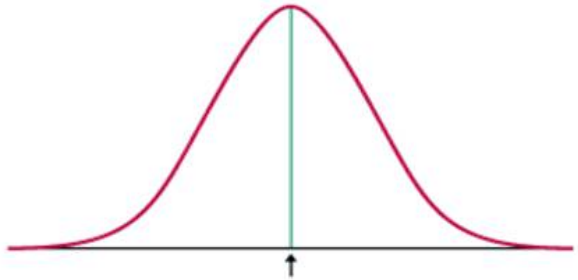


(b) Right skewed mean =3.4>2.5=median

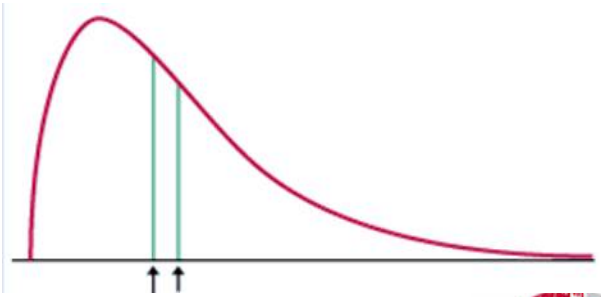
✍ For a distribution curve, mean = median only if it is symmetric.



(c) Left skewed: mean<median



(d) Symmetry: mean = median



(e) Right skewed: median<mean

4. Numerical measures of spread: percentile and quartile

✍ For a random variable X the distribution function $F(x)$ tells the probability for $X \leq x$. Correspondingly, the quantile function $F^{-1}(p)$ defines the point x such that

$$F(x) = \mathrm{P}(X \leq x) \leq p \qquad \text{and} \qquad \mathrm{P}(X > x) > 1 - p, \qquad p \in (0, 1).$$

✍ The p-th percentile for a sample: $100 \cdot p\%$ of the observations fall at or below it. For example, the median is the 50-th percentile.

✍ Quartiles divide the sample into four parts of equal percentage.

- the 1st quartile Q_1: the 25-th percentile (median of the lower half of data);
- the 2nd quartile Q_2: 50-th percentile;
- the 3rd quartile Q_3: the 75-th percentile (median of the upper half of data).

✍ Inter-quartiles range $IQR = Q_3 - Q_1$ also serves as a measure of spread.

✍ 1.5×IQR rule for outliers: An observation is a suspected outlier if it falls $1.5 \cdot IQR$ away outside the interval $[Q_1, Q_3]$.

5. Numerical measures of spread: an example

The **first quartile, Q_1,** is the value in the sample that has 25% of the data at or below it ($\Leftrightarrow$ it is the median of the lower half of the sorted data, excluding M).

M = median = 3.4

The **third quartile, Q_3,** is the value in the sample that has 75% of the data at or below it ($\Leftrightarrow$ it is the median of the upper half of the sorted data, excluding M).

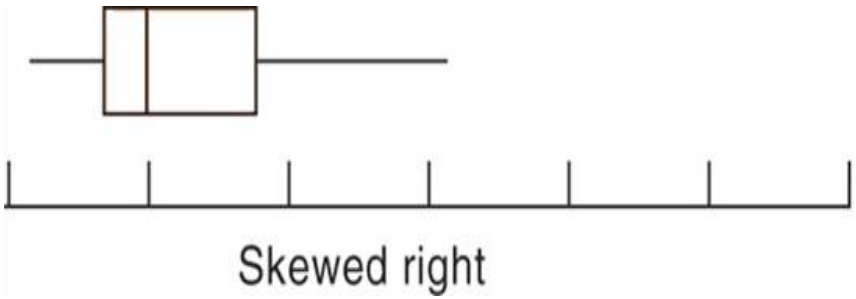
1	1	0.6
2	2	1.2
3	3	1.6
4	4	1.9
5	5	1.5
6	6	2.1
7	7	2.3
8	1	2.3
9	2	2.5
10	3	2.8
11	4	2.9
12	5	3.3
13		3.4
14	1	3.6
15	2	3.7
16	3	3.8
17	4	3.9
18	5	4.1
19	6	4.2
20	7	4.5
21	1	4.7
22	2	4.9
23	3	5.3
24	4	5.6
25	5	6.1

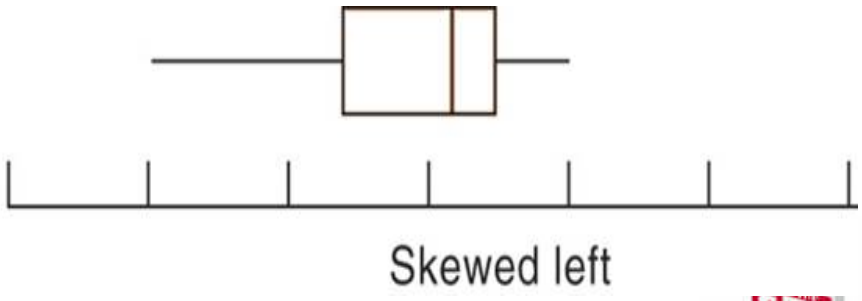
Q_1= first quartile = 2.2

Q_3= third quartile = 4.35

6. Five-number summary and box plot

- Five numbers: Minimum, Q_1, Median, Q_3 and Maximum.
- Box plot is a visual displaying of the five-number summary.
  - Central box: Q_1 to Q_3.
  - Line inside box: Median.
  - Extended straight lines: lowest to highest observation, except outliers.
  - Outliers marked as circles or stars.
- Read the symmetry from box plot.





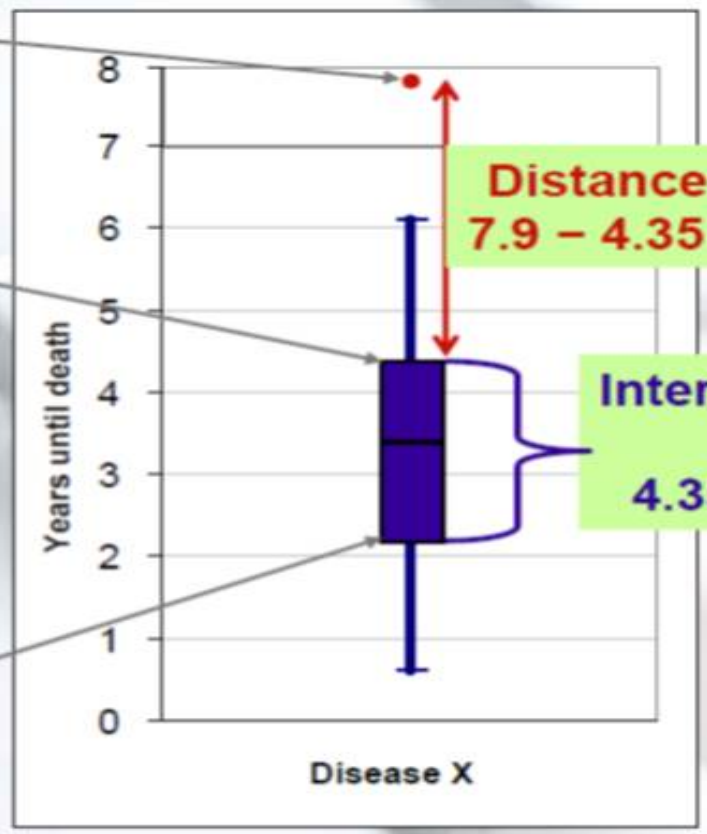
QUERY 2: In R the function boxplot(x) produces the box plot of the sample x.

7. Five-number summary and box plot

25	6	7.9
24	5	6.1
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

$Q_3 = 4.35$

$Q_1 = 2.2$



Distance to Q_3
$7.9 - 4.35 = 3.55$

Interquartile range
$Q_3 - Q_1$
$4.35 - 2.2 = 2.15$

Individual #25 has a value of 7.9 years, which is 3.55 years above the third quartile. This is more than 3.225 years, 1.5 * IQR. Thus, individual #25 is a suspected outlier.

8. Numerical measure of spread: standard deviation

☞ Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$ average squared distance from the mean.

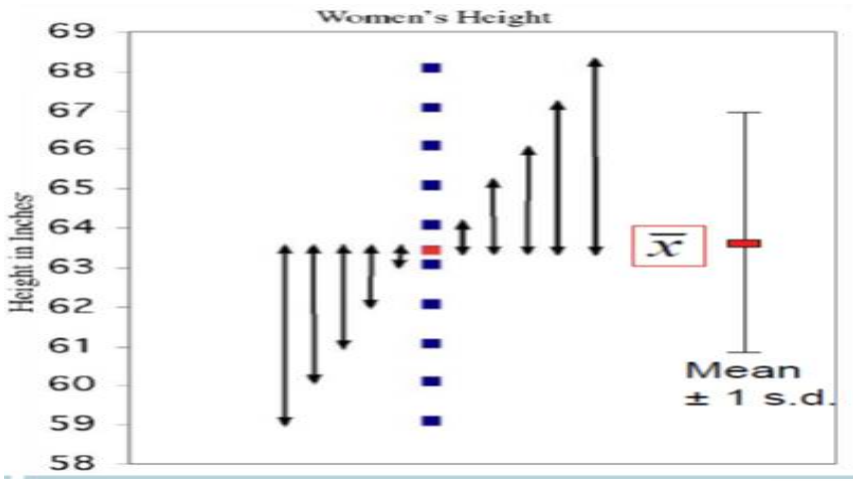
☞ Standard deviation $s = \sqrt{s^2}$ measures the variation around mean, a larger s implies larger variation, and $s = 0$ mean no variation (How can it occur?).

☞ Variance and standard deviation are not robust.

Women height

i	x_i	$\bar{x}$	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$
1	59	63.4	-4.4	19.0
2	60	63.4	-3.4	11.3
3	61	63.4	-2.4	5.6
4	62	63.4	-1.4	1.8
5	62	63.4	-1.4	1.8
6	63	63.4	-0.4	0.1
7	63	63.4	-0.4	0.1
8	63	63.4	-0.4	0.1
9	64	63.4	0.6	0.4
10	64	63.4	0.6	0.4
11	65	63.4	1.6	2.7
12	66	63.4	2.6	7.0
13	67	63.4	3.6	13.3
14	68	63.4	4.6	21.6
	Mean 63.4		Sum 0.0	Sum 85.2

(f) Variance $s^2 = \frac{85.2}{14-1} = 6.55$



(g) Sum of squared distances

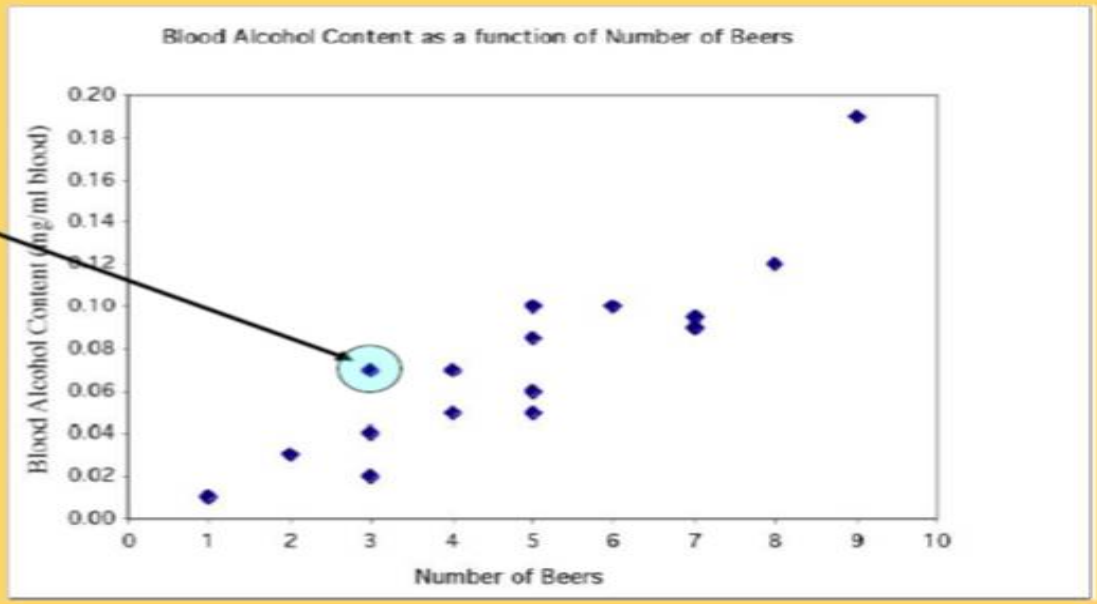
Query 3: Why not use n instead of $n-1$ in the variance?

9. Relation between two variables - scatter plot

✐ Association: for a data with two or more variables of each individual being observed, usually some values of one variable tend to occur more often with certain values of the other variable. This is also called as statistical dependence.

✍ Scatter plot is a visual presentation of the association.

Student	Beers	BAC
1	5	0.1
2	2	0.03
3	9	0.19
6	7	0.095
7	3	0.07
9	3	0.02
11	4	0.07
13	5	0.085
4	8	0.12
5	3	0.04
8	5	0.06
10	5	0.05
12	6	0.1
14	7	0.09
15	1	0.01
16	4	0.05

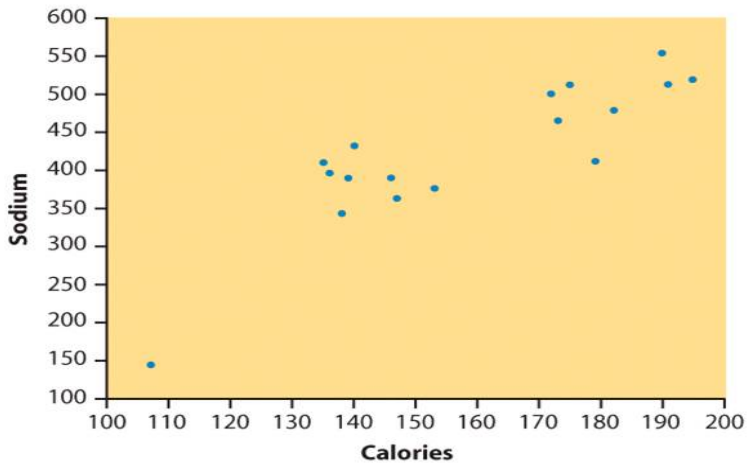


10. Response variable vs explanatory variables

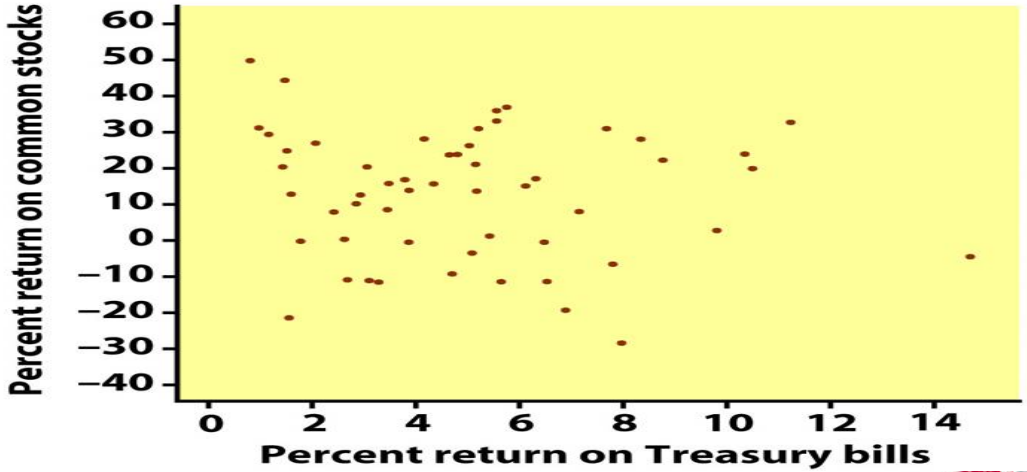
✏ The response variable measures or records an outcome of a study, and explanatory variables explain the change in the response variable. For example, life expectancy, weight, height and smoking habits etc.

✍ Typically, the explanatory variable is plotted horizontally and the response is plotted vertically.

✍ Some plots don't have clear explanatory and response variables.



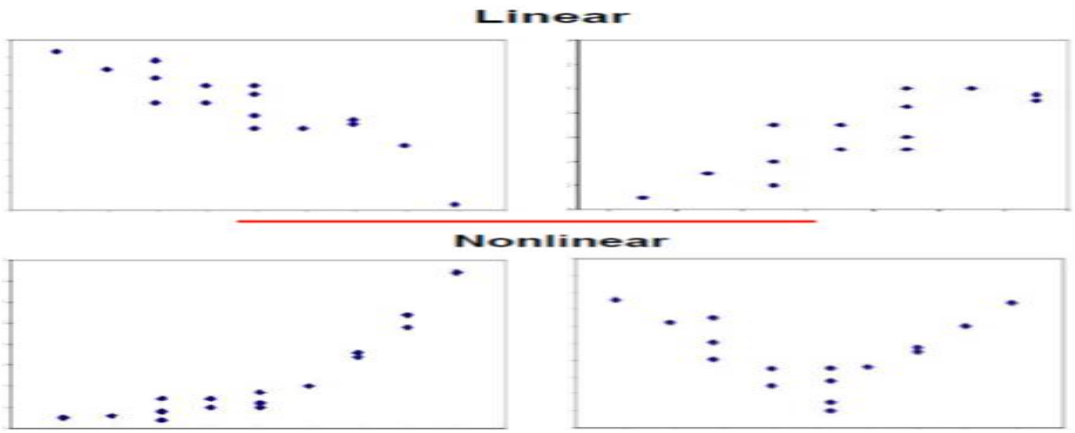
(h) calories explain sodium amounts?

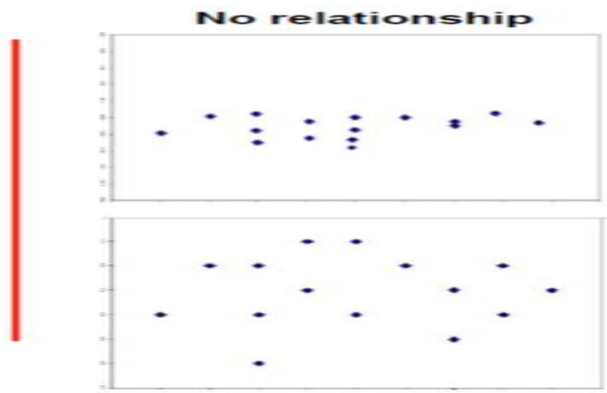


(i) return on treasury bills explains that on stocks?

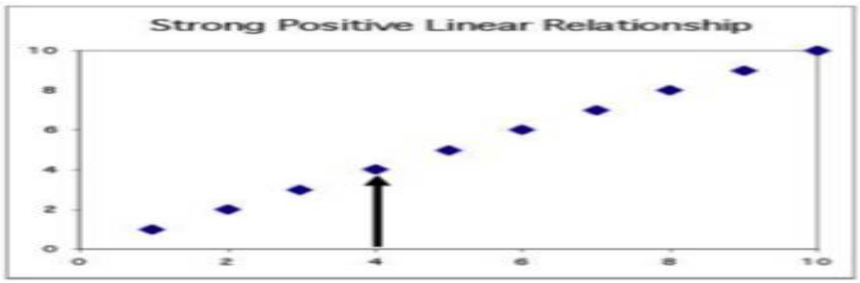
11. Interpreting scatter plots – form, direction, strength

✏ Form (linear, curved, clusters, no pattern) and direction (positive, negative, no direction).

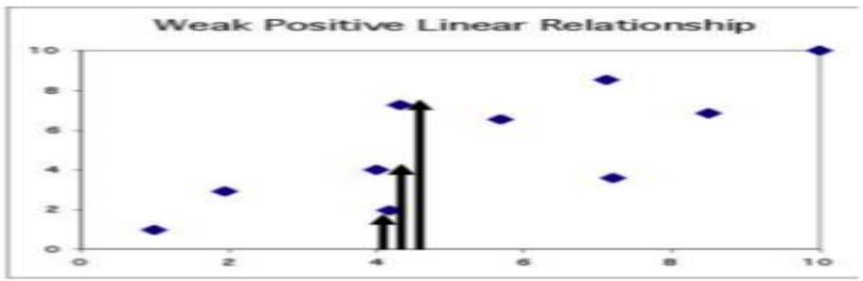




✏ Strength: how closely the points fit the 'form'.



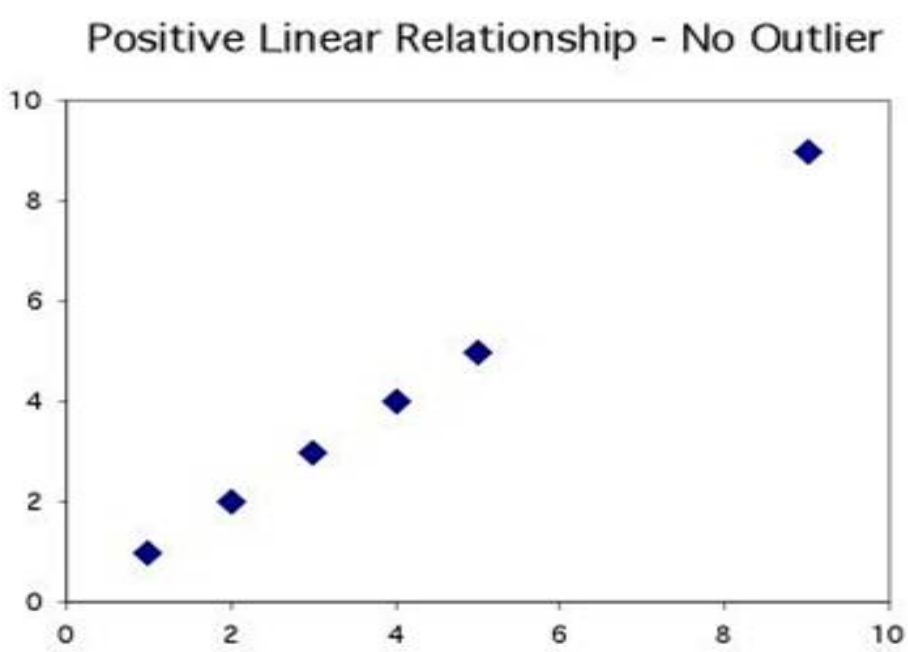
With a strong relationship, you can get a pretty good estimate of y if you know x.

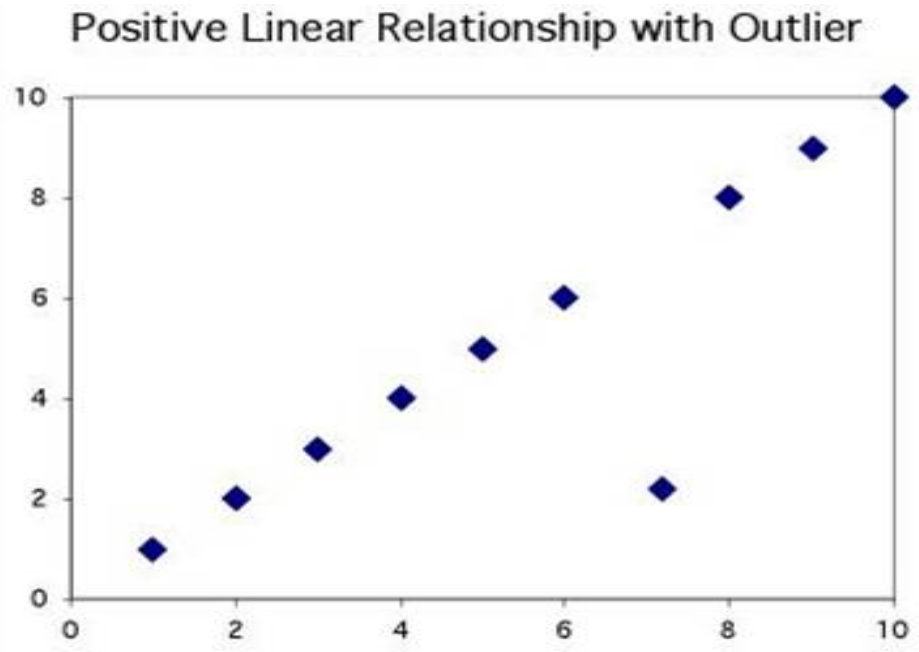


With a weak relationship, for any x you might get a wide range of y values.

12. Interpreting scatter plots – outliers

✏ An outlier is a data value that has a very low probability of occurrence (i.e., it is unusual or unexpected).





✏ In a scatter plot, outliers are points that fall outside of the overall pattern of the relationship.

13. Measure of association - correlation coefficient

✏The linear association is quite common in practice, and it can be conveniently modeled by using linear regression.

✏ For paired data, (x_i, y_i), $i = 1, \cdots, n$, the Pearson's correlation coefficient

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

measures the strength and direction of a linear relationship between two x and y, where s_x and s_y are standard deviations of x and y, respectively.

✍ Since all both physical units are discarded when we standardize x_i by

$$\frac{x_i - \bar{x}}{s_x}, \qquad \frac{y_i - \bar{y}}{s_y} \quad i = 1, \cdots, n,$$

the r has no unit and thus is scale invariant, i.e., changing unit incurs no change in the coefficient.

Query 4: Check R references for the command to obtain the correlation coefficient.

14. Correlation coefficient – interpretation

✍ Let $\boldsymbol{x} = (x_1 - \bar{x}, \cdots, x_n - \bar{x})'$ and $\boldsymbol{y} = (y_1 - \bar{y}, \cdots, y_n - \bar{y})'$. Then,

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\boldsymbol{x}'\boldsymbol{y}}{\sqrt{\boldsymbol{x}'\boldsymbol{x}} \cdot \sqrt{\boldsymbol{y}'\boldsymbol{y}}}.$$

According to Cauchy-Schwarz inequality,

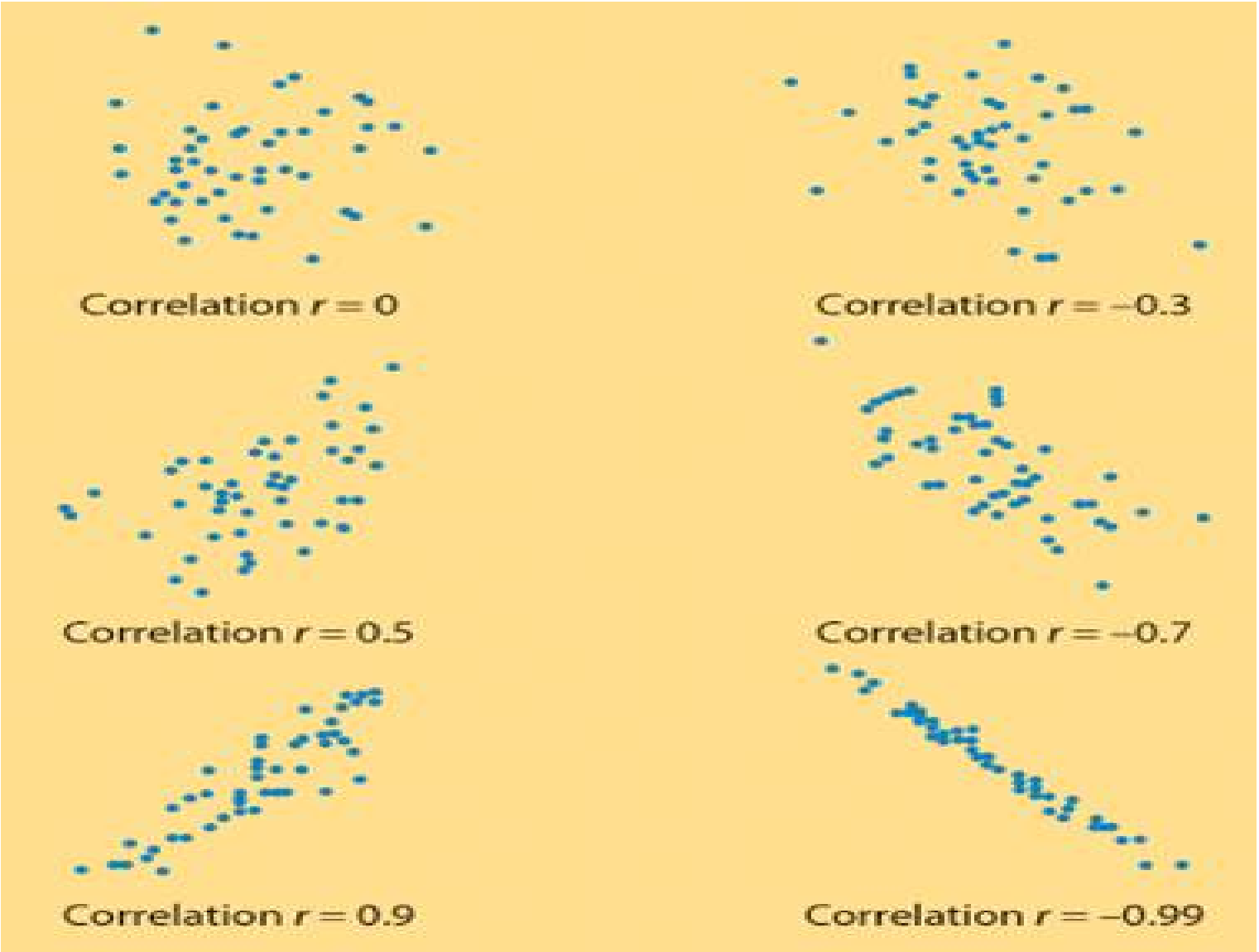
$$(\boldsymbol{x}'\boldsymbol{y})^2 \leq (\boldsymbol{x}'\boldsymbol{x}) \cdot (\boldsymbol{y}'\boldsymbol{y}).$$

Then, it holds that $r^2 \leq 1$, i.e., r ranges from -1 to +1.

✍ In the above inequality '=' holds iff $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly dependent, i.e., $\boldsymbol{y} = \lambda \boldsymbol{x}$ for some constant λ. So, r quantifies

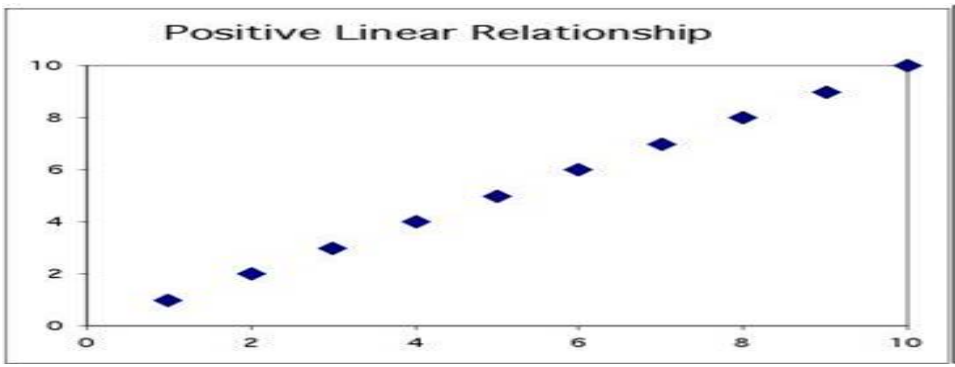
- strength: larger $|r|$ corresponding to points more closely centered around a straight line, and
- direction: positive (negative) when individuals with larger x values tend to have larger (smaller) y values.

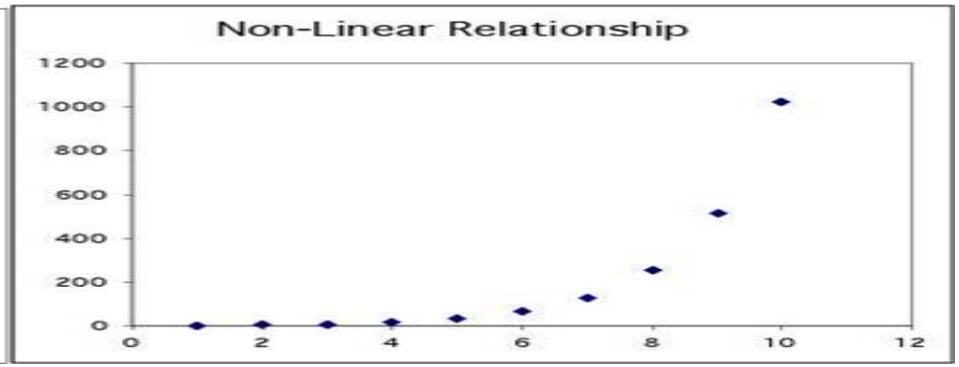
15. Correlation coefficient – interpretation



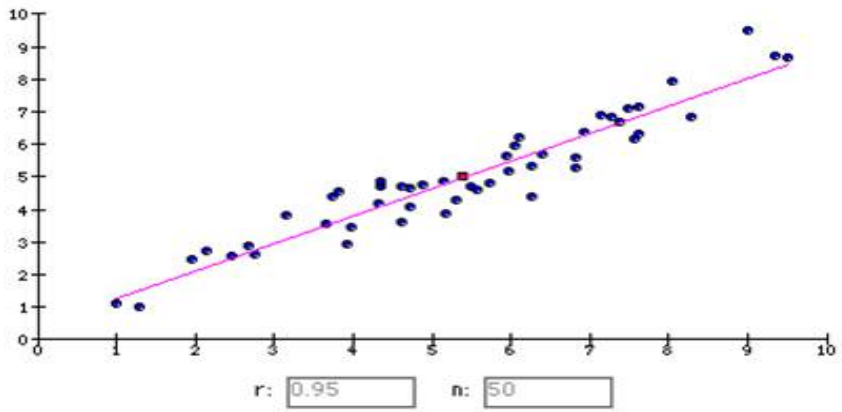
16. Correlation coefficient – interpretation

☞ The correlation coefficient for the linear association fails for the nonlinear pattern.

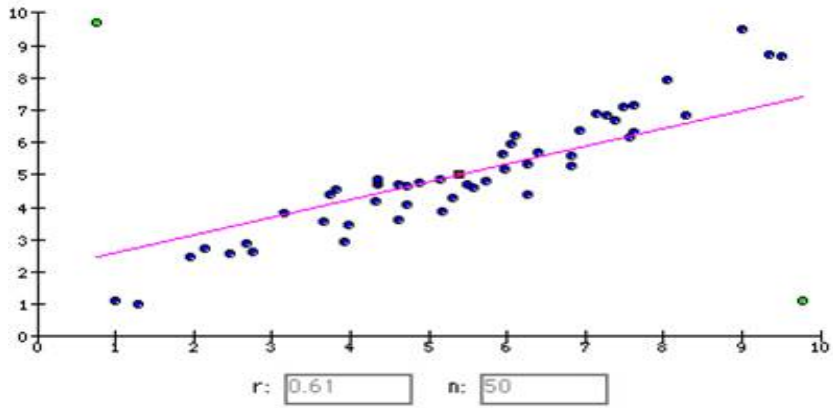




☞ Influential points: r is not resistant to outliers.



(j) without outliers



(k) with two outliers

QUERY 5: Refer to Wikipedia for Kendall's τ and Spearman's ρ, coefficients for nonlinear association.