IC152 Lec 17

Feb 2021

These slides are partly made with $\LaTeX$

Statistics

- Develop measures to summarize a dataset
- ***Statistics*** are quantities whose values are determined by the data
- Eg: sample mean, sample median, sample mode measure the centre of a dataset
- Sample standard deviation: measures variation
- Sample correlation: to meausre pairwise relationships

sample mean

Data points $x_1, x_2, \ldots x_n$
Sample mean is defined as

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

$$y_i = x_i + c \Rightarrow \bar{y} = \bar{x} + c$$

Shifting by a constant

Motorbike accidents 1976

Classification of accident	Interpretation
0	No head injury
1	Minor head injury
2	Moderate head injury
3	Severe, not life-threatening
4	Severe and life-threatening
5	Critical, survival uncertain at time of accident
6	Fatal

Classification	Frequency of driver with helmet	Frequency of driver without helmet
0	248	227
1	58	135
2	11	33
3	3	14
4	2	3
5	8	21
6	1	6
	331	439

Sample mean of head injury severity for helmeted riders:

$$\frac{0 \times 248 + 1 \times 58 + 2 \times 11 + 3 \times 3 + 4 \times 2 + 5 \times 8 + 6 \times 1}{331} = 0.432$$

For non-helmeted riders:

$$\frac{0 \times 227 + 1 \times 135 + 2 \times 33 + 3 \times 14 + 4 \times 3 + 5 \times 21 + 6 \times 6}{331} = 0.902$$

Data indicates that riders with helmets suffered lesser than riders without

Expected value

Deviations

Deviations from the mean: $x_i - \bar{x}$

HW: Show that: $$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Sample median

- Order the values from smallest to the largest. For odd n, sample median = middle value. For even n, sample median = average of the two middle values.
- Sample mean is affected by extreme values. Sample median is not.

Example use

- Flat-rate income tax for city. How much income to expect?
- Middle-class housing project. How many citizens can afford this?
- Both sample mean and sample median are useful statistics.

Sample mode

- Most frequently occuring value
- 8, 10, 6, 4, 10, 12, 14, 10

  sample mode is 10

Sample variance and sample std

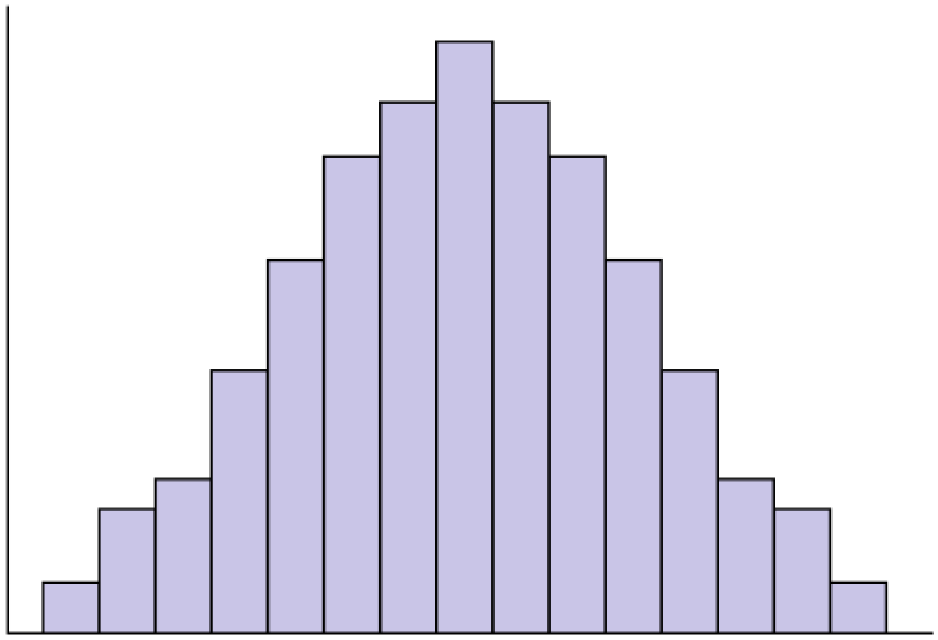
- A: 1,2,5,6,6 B: -40,0,5,20,35
- A and B have same sample mean, but B has more spread than A

Data points $x_1, x_2, \ldots x_n$
Sample variance s^2 is defined as

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

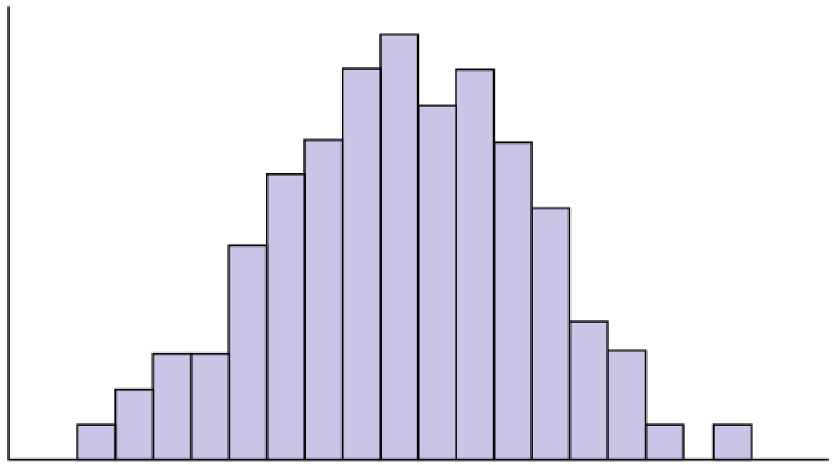
Sample standard deviation = positive square root of sample variance

Normal datasets

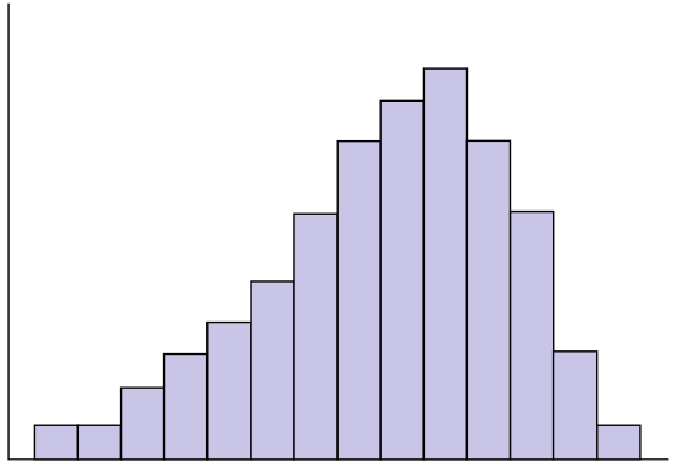


A dataset is ***normal*** if its histogram:

- Is highest at the middle interval
- Is bell-shaped
- Is symmetric about its middle interval



Approximately normal



Skewed

Empirical rule for normal data

1. Approximately 68% of the data lie within

$$\bar{x} \pm s$$

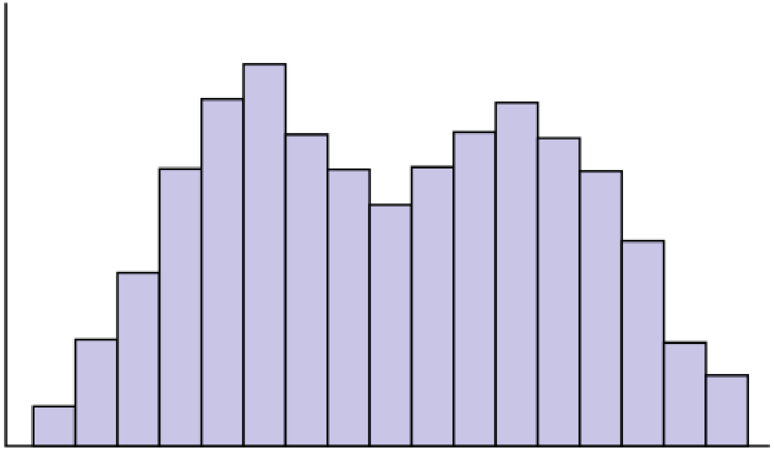
2. Approximately 95% of the data lie within

$$\bar{x} \pm 2s$$

3. Approximately 99.7% of the data lie within

$$\bar{x} \pm 3s$$

Bimodal data

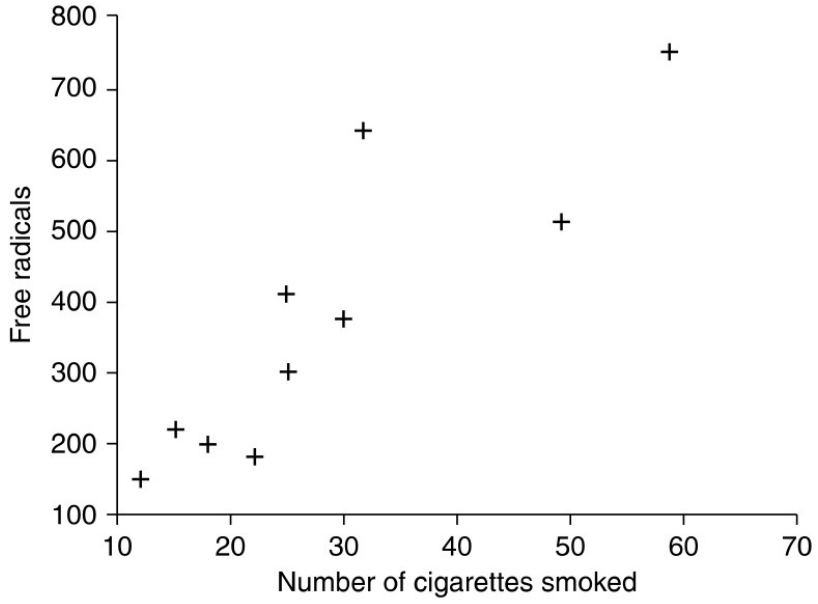


Superposition of two normal histograms
Eg. weights of men and women

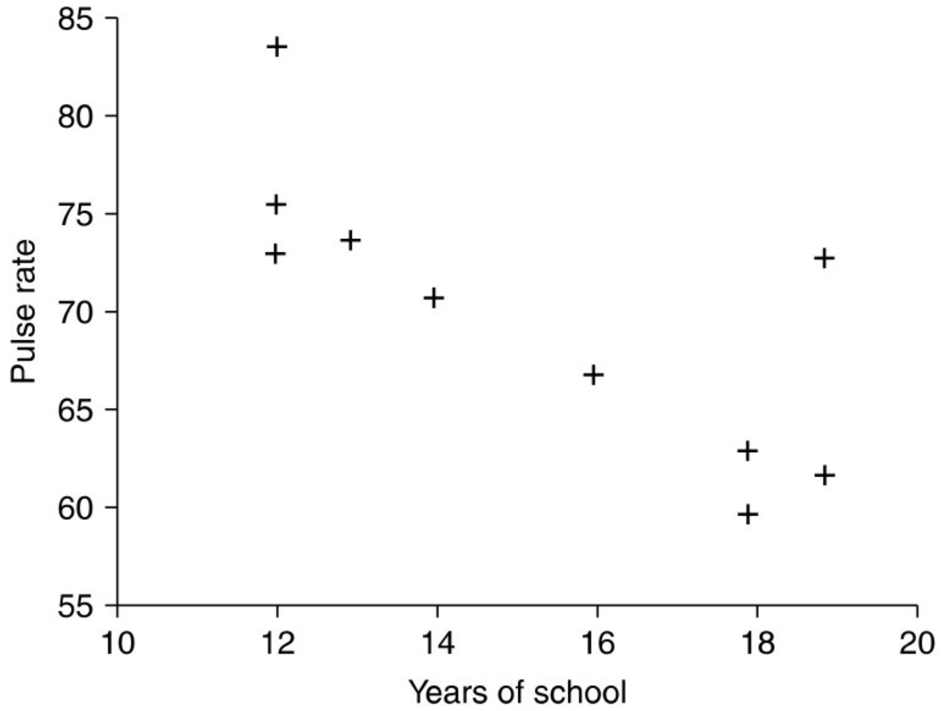
Sample correlation coefficient

- Paired data (xi,yi)
- How does increase in x affect the y?

Person	Number of cigarettes smoked	Free radicals
1	18	202
2	32	644
3	25	411
4	60	755
5	12	144
6	25	302
7	50	512
8	15	223
9	22	183
10	30	375



	Person									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Years of school	12	16	13	18	19	12	18	19	12	14
Pulse rate	73	67	74	63	73	84	60	62	76	71



If $x_i - \bar{x}$ and $y_i - \bar{y}$ have the same sign, then their product $(x_i - \bar{x})(y_i - \bar{y})$ will be positive. Thus when large x_i values are associated with large y_i values, and small x_i values with small y_i values, then $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ will be a large positive number.

Standarize the sum $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ by dividing with $n-1$, and then dividing by the product of the two sample deviations.

Let s_x and s_y be the sample standard deviations of x_i and y_i. The sample correlation coefficient r_{xy} for the data pairs (x_i, y_i) is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Suppose, instead of 2 variables x and y, we had a vector of variables

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

Now there are correlations between pairwise components of $\mathbf{x}$: $r_{x_1 x_2}$, $r_{x_1 x_3}$, $r_{x_2 x_3}$, $r_{x_1 x_1}$, $r_{x_2 x_2}$, $r_{x_3 x_3}$.

Correlation measures association, not causation

- Strong negative correlation between number of years in school and resting pulse rate
- Does this that imply more years in school reduces the pulse rate?
- Association is not causation
- Eg. more time in school, more aware of healthy lifestyle, or has a job which gives time for exercise

False causality examples (from Wikipedia):

- The faster a windmill rotaes, more wind is observed. Therefore, windmills cause winds.
- Children that watch a lot of TV are violent. Therefore, TV makes children more violent.
- Sleeping with one's shoes on is strongly correlated with waking up with a headache. Therefore, sleeping with one's shoes on causes headache. Missing factor: going to bed drunk.
- As ice cream sales increase, the rate of drowning deaths increases sharply. Therefore, ice cream consumption causes drowning. Missing factor: summer weather.