Examine Bivariate Data

Two Categorical Variables

Contingency Table

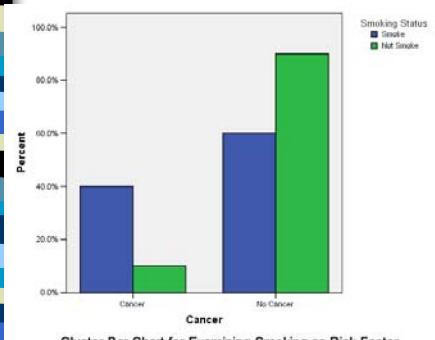
	Cancer	No cancer	Row Total
Smoker	20	30	50
Non- Smoker	5	45	50
Column Total	25	75	100

Odds of smoker to have cancer: **20/30 = 6/9**

Odds of nonsmoker to have cancer: **5/45 = 1/9**

Odds Ratio = (6/9)/(1/9) = 6

Cluster bar chart



Cluster Bar Chart for Examining Smoking as Risk Factor

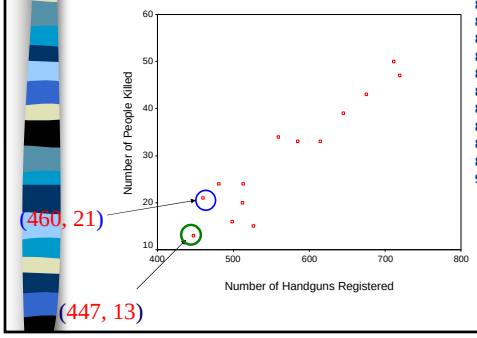
Relation Between Two Quantitative Variables

Is there relation between "**number of handguns registered**" in the area and "**number of people killed by guns**"?

Year	NGR(x_i)	Nkill(y_i)
77	447	13
78	460	21
79	481	24
80	498	16
81	513	24
82	512	20
83	526	15
84	559	34
85	585	33
86	614	33
87	645	39
88	675	43
89	711	50
90	719	47

Scatter Plot

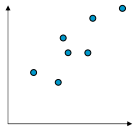
Year	NGR(x_i)	Nkill(y_i)
77	447	13
78	460	21
79	481	24
80	498	16
81	513	24
82	512	20
83	526	15
84	559	34
85	585	33
86	614	33
87	645	39
88	675	43
89	711	50
90	719	47

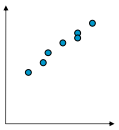


Information in Scatter Plot

- Form (Straight line or curve relation)
- Direction (Positive or negative relation)
- Strength (Strong or weak relation)
- Outliers

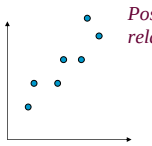
Correlation



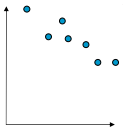


Stronger linear trend

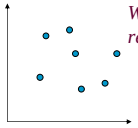
Correlation Coefficient



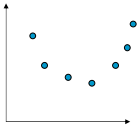
Positive relation



Negative relation



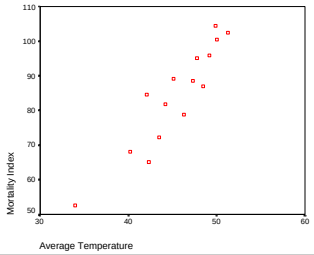
Weak relation



Curve relation

Two Quantitative variables

Average annual temperature and the mortality index for a type of breast cancer in women in certain region of Europe.

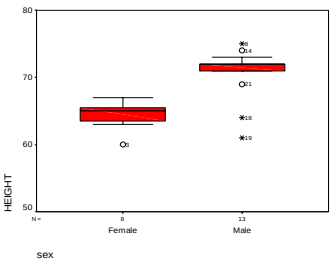


Two Quantitative variables

Data:

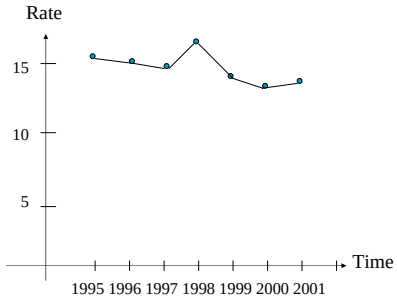
Temperature	Mortality Index
34	52
40	68
42	63
42	83
43	72
44	81
45	89
46	77
47	88
48	94
49	86
50	95
51	105
51	100
52	102

A Categorical & A Quantitative Variables
Side-by-side Boxplot

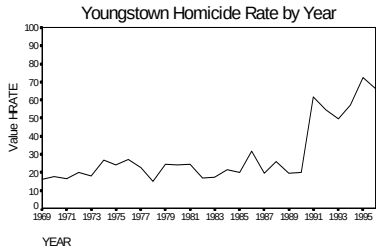


Time Plot

Time	Rate
1995	15.2
1996	15.1
1997	14.9
1998	16.2
1999	14.3
2000	13.2
2001	13.5



Time Plot

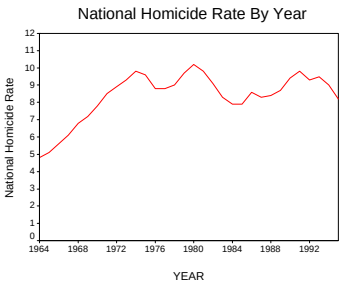


Chapter 3 - 13

Time Plot

Chapter 3 - 14

Time Plot

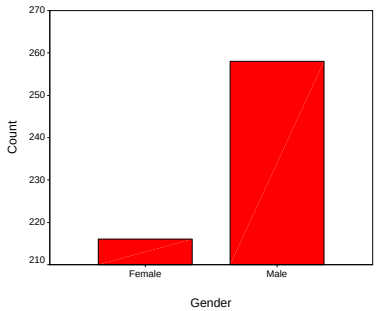


Chapter 3 - 15

Incorrect and Misleading Chart

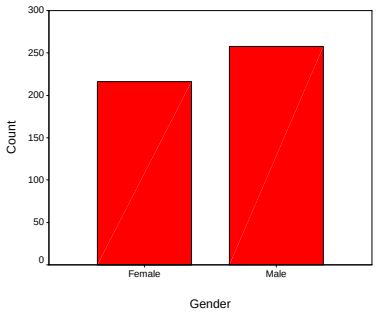
Chapter 3 - 16

Misleading Chart



Chapter 3 - 17

Chart in Original Scale



Chapter 3 - 18

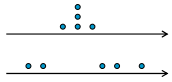
Data Description

Describe Distribution with Numbers

Example: Birth weights (in lb) of 5 babies born from two groups of women under different care programs.

Group 1: 7, 6, 8, 7, 7

Group 2: 3, 4, 8, 9, 11



Measure of Central Tendency

Describing Center

Measure of Central Tendency

Mean: the average value of the data.

If the values of a sample of n observations are denoted by $x_1, x_2, ..., x_n$, their **sample mean** is

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

* If the data were for the whole population then the result from this calculation would be called the **population mean**, and the notation for it is μ.

Example: Birth weights (in lb) of 5 babies born from a group of women under certain diet.

7, 6, 8, 7, 7

Sol:

mean = (7 + 6 + 8 +7 + 7) / 5 = 35/5 = 7

[near the center of the data set]

Example: (number of hysterectomies performed by 15 male doctors)

27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

=> mean = **41.33**

Median: of a data set is

- the data value exactly in the middle of its ordered list if the number of pieces of data is odd,
- the mean of the two middle data values in its ordered list if the number of pieces of data is even.

[median is not influenced by outliers and is best for nonsymmetric distribution]

Example: (number of times visited class website by 15 students)
27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

ordered list => 20, 25, 25, 27, 28, 31, 33, **34**, 36, 37, 44, 50, 59, 85, 86
median = **34**

Example: (Birth weights for 6 infants.)

5, 7, 6, 8, 5, 9

ordered list => 5, 5, 6, 7, 8, 9

median = (6+7) / 2 = ***6.5***

Mode: of a data set is the observation that occurs most frequently.

Example 1: (number of times visited class website by 15 students)
27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

ordered list => 20, **25, 25**, 27, 28, 31, 33, 34, 36, 37, 44, 50, 59, 85, 86

Mode = **25**

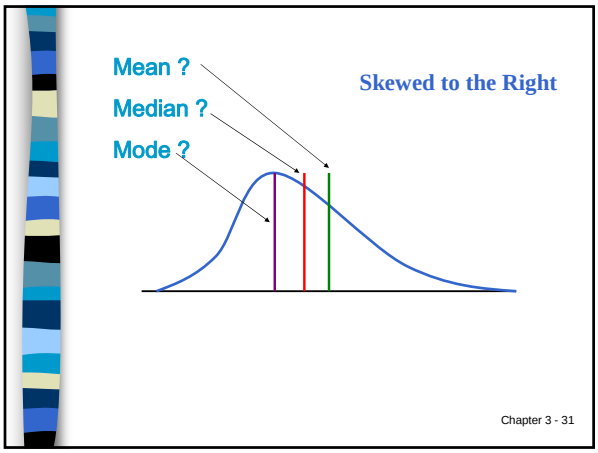
Example 2: (Blood type of 15 students)
A, B, A, A, O, AB, A, A, B, B, O, O, A, A, A

Mode = A

A – 8
B – 3
O – 3
AB – 1

What is a Modal class?

Class	Frequency	Relative Freq.	Cumulative R.F.
90< - 110	2	2/22 = .091	2/22
110< - 130	2	2/22 = .091	4/22
130< - 150	4	4/22 = .182	8/22
150< - 170	2	2/22 = .091	10/22
170< - 190	**7**	**7/22 = .318**	**17/22**
190< - 210	3	3/22 = .136	20/22
210< - 230	1	1/22 = .045	21/22
230< - 250	0	0/22 = .000	21/22
250< - 270	0	0/22 = .000	21/22
270< - 290	1	1/22 = .045	22/22
Total	**22**	**1.000**	



Midrange

- The average of the lowest and the highest values in the data set.

$$\text{Midrange} = \frac{\text{Lowest Value} + \text{Highest Value}}{2}$$

Example: (number of times visited class website by 15 students)
27, 50, 33, 25, 86, 25, 85, 31, 37, 44, 20, 36, 59, 34, 28

Lowest value = 20
Highest value = 86

Midrange = (20 + 86) / 2 = **53**

Weighted Mean

Example: (Grade point average)
A student received 3 A's, 5 B's, 2 C's.

Class (grade point, x)	Frequency (weight, w)
4	3
3	5
2	2

$$\text{Average grade point} = \frac{3 \times 4 + 5 \times 3 + 2 \times 2}{3 + 5 + 2} = \frac{31}{10} = 3.1$$

weight

Weighted Mean

$$\text{Weighted mean} = \frac{w_1 \cdot x_1 + w_2 \cdot x_2 + \ldots + w_k \cdot x_k}{w_1 + w_2 + \ldots + w_k} = \frac{\Sigma\, w \cdot x}{\Sigma\, w}$$

Where $w_1, w_2, \ldots$ are the weights and $x_1, x_2, \ldots$ are the values (or class midpoint or class mark).

Mean Estimation

Class	Frequency (w)	Class Mark (x)	w · x
90 - < 110	1	100	1x100
110 - < 130	2	120	2x120
130 - <150	3	140	3x140
150 - < 170	1	160	1x160
Total	**7**		**920**

$$\text{Estimated mean} = \frac{920}{7} = \mathbf{131.43}$$