Describing datasets

IC152 Lec 12
Feb 2021

The nature of data

- Numerical findings: need to be presented consisely
- Especially needed for large datasets
- Features of the data include:
  - Range
  - Degree of symmetry
  - Concentrated or spread out
  - Where are they concentrated? Etc.
- Univariate or multivariate

An example

2, 2, 0, 0, 5, 8, 3, 4, 1, 0, 0, 7, 1, 7, 1, 5, 4, 0, 4, 0, 1, 8, 9, 7, 0,

1, 7, 2, 5, 5, 4, 3, 3, 0, 0, 2, 5, 1, 3, 0, 1, 0, 2, 4, 5, 0, 5, 7, 5, 1

Data: Number of sick leaves taken by 50 employees, over six weeks

Value	Freq
0	12
1	8
2	5
3	4
4	5
5	8
6	0
7	5
8	2
9	1

Sum of freq values = N

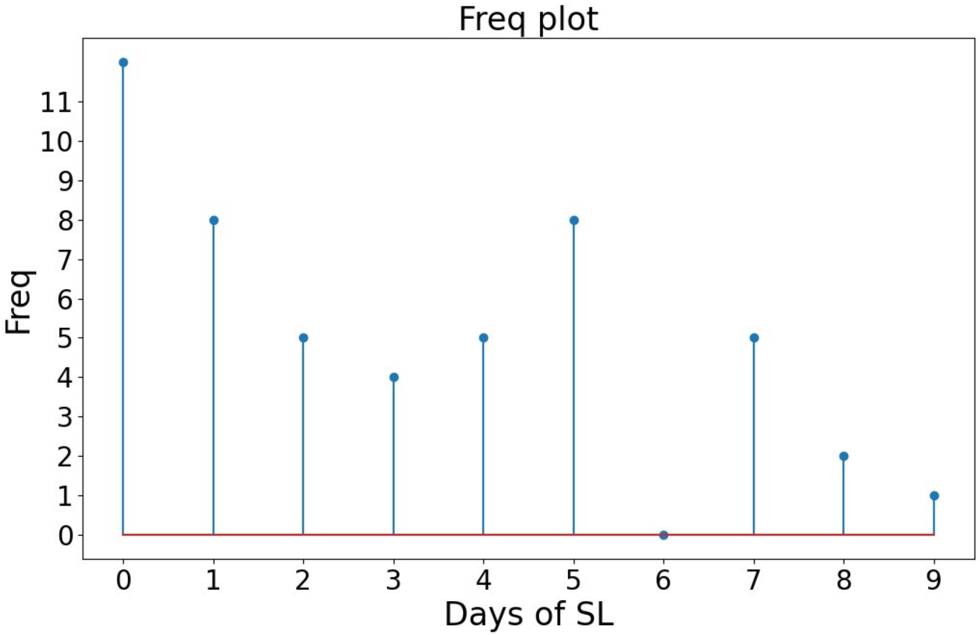
N is the total number of samples in the data

Small dataset

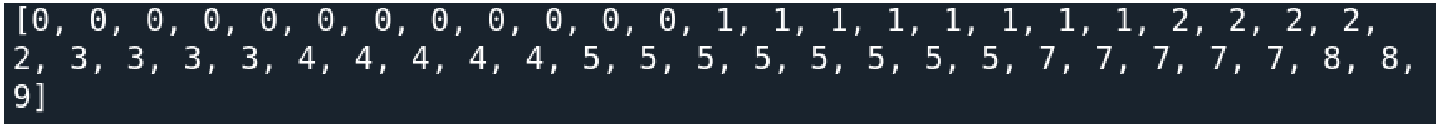
Frequency table

- How many workers had at least one day of sick leave? 50-12 = 38
- How many workers had between 3 and 5 days of sick leave? 4+5+8 = 17
- How many had more than 5 days? 8

Value	Freq
0	12
1	8
2	5
3	4
4	5
5	8
6	0
7	5
8	2
9	1



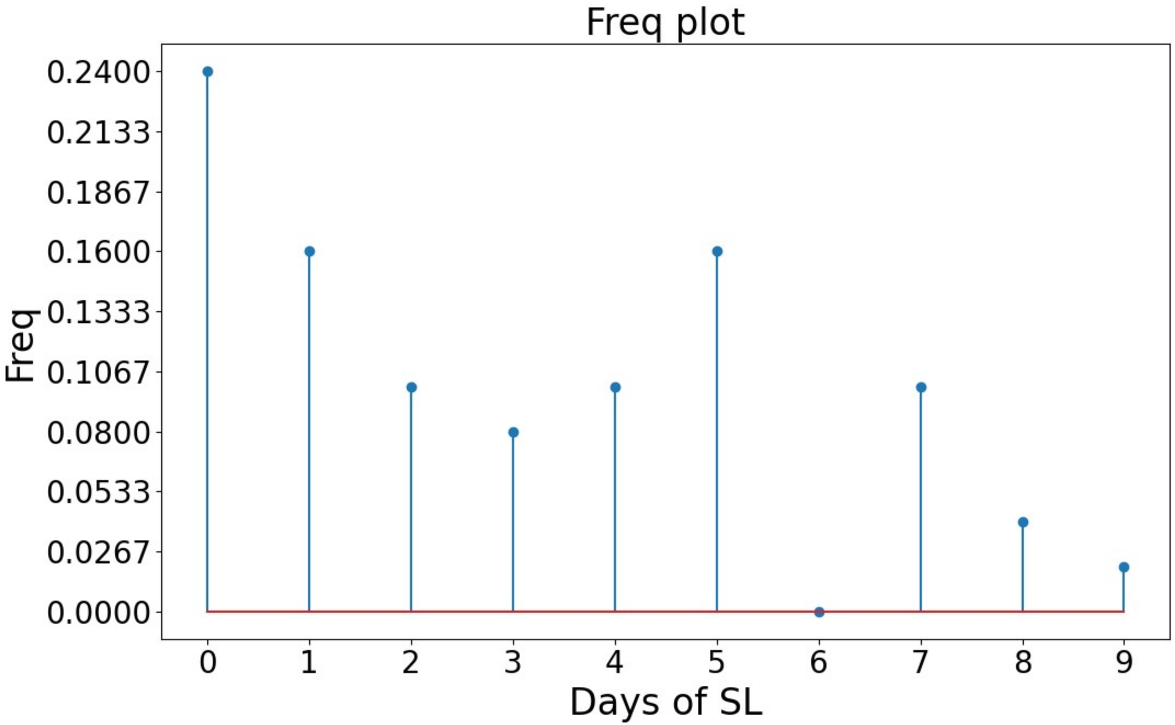
How will you construct this table and make the plot?



```
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2,
2, 3, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 7, 7, 7, 7, 7, 8, 8,
9]
```

Sorting is a common preprocessing operation

Normalized frequency plot

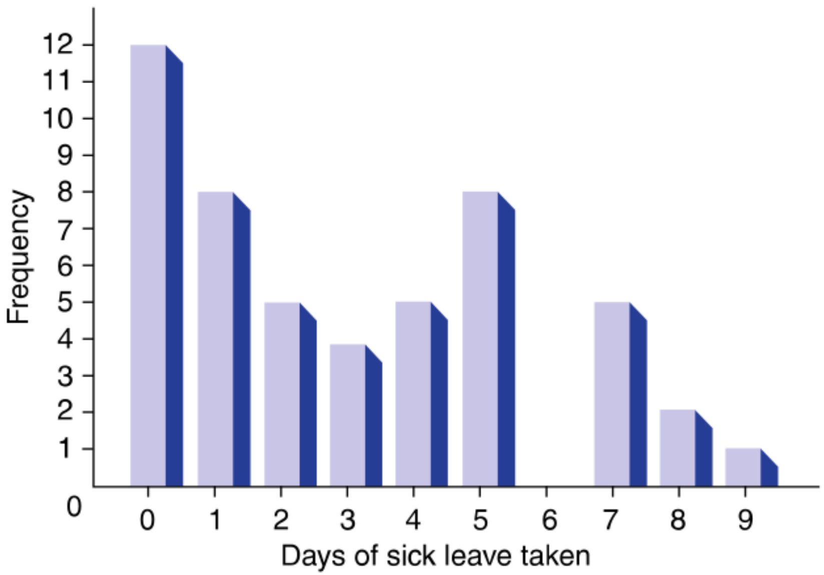


Sometimes normalized frequency is more convenient

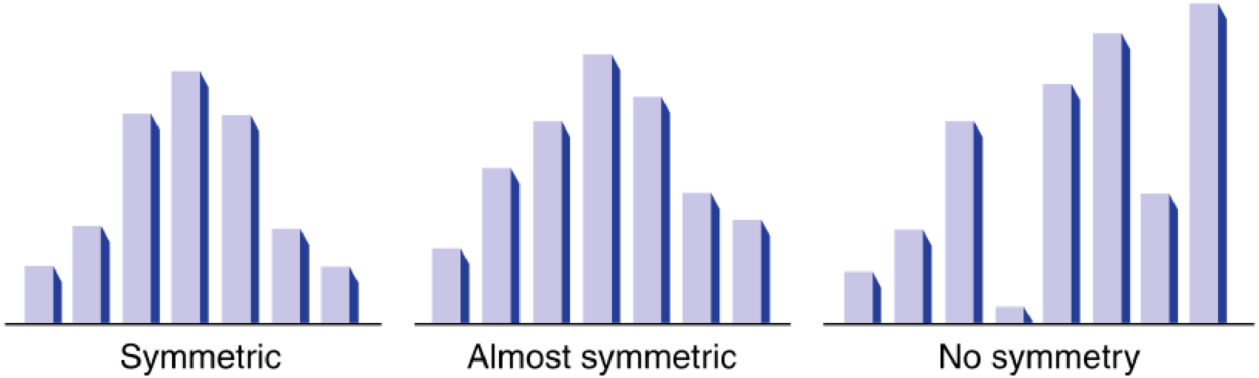
Sum of frequency values = 1

Bar plot

Bars rather than lines



Types of bar plots

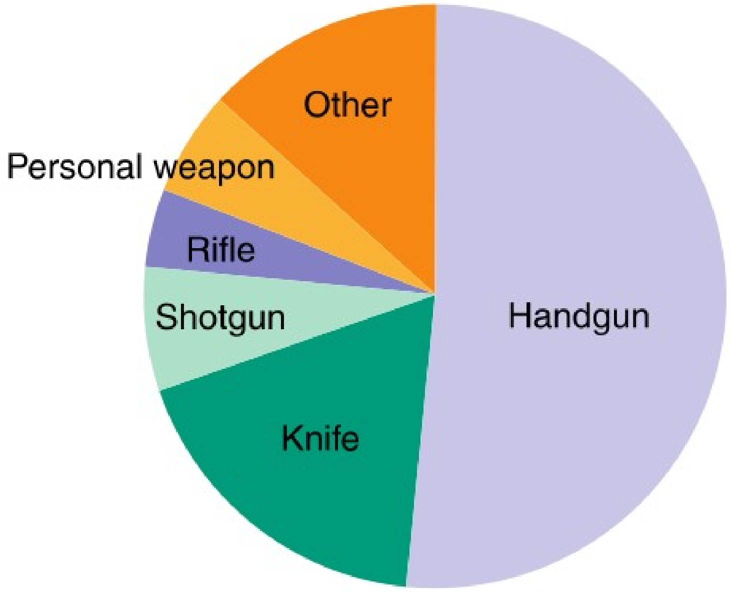


Symmetric

Almost symmetric

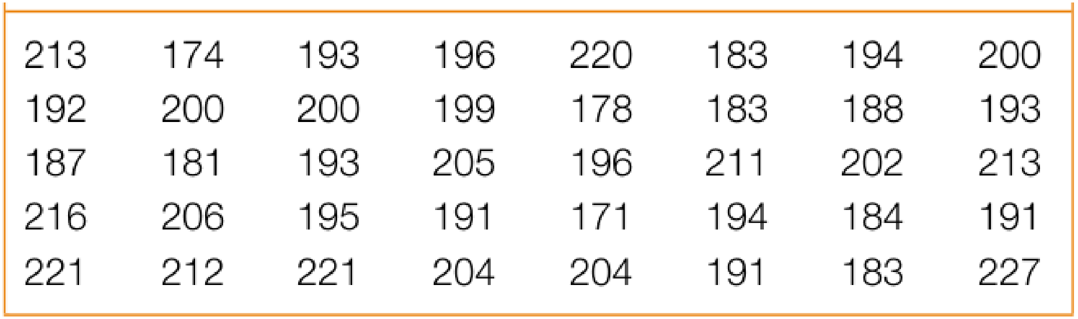
No symmetry

Pie chart: for non-numerical data
Visualization of relative frequency plot



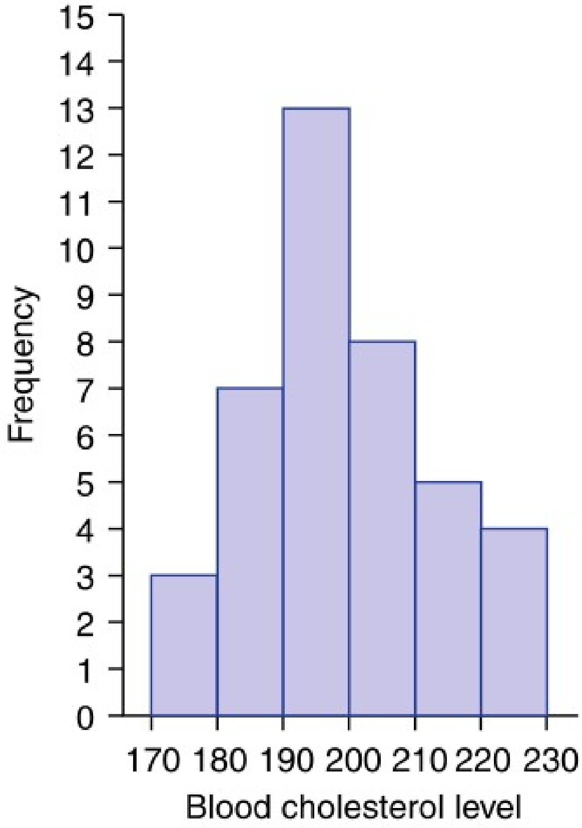
Weapons used in crimes

Grouped data and histograms



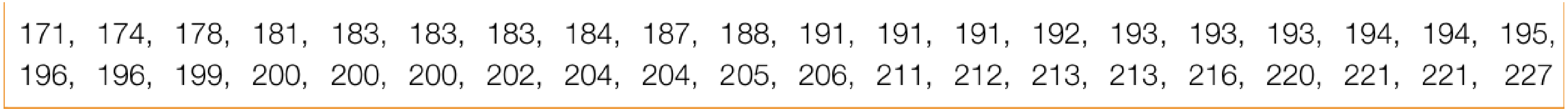
213	174	193	196	220	183	194	200
192	200	200	199	178	183	188	193
187	181	193	205	196	211	202	213
216	206	195	191	171	194	184	191
221	212	221	204	204	191	183	227

Data: Blood cholestrol levels



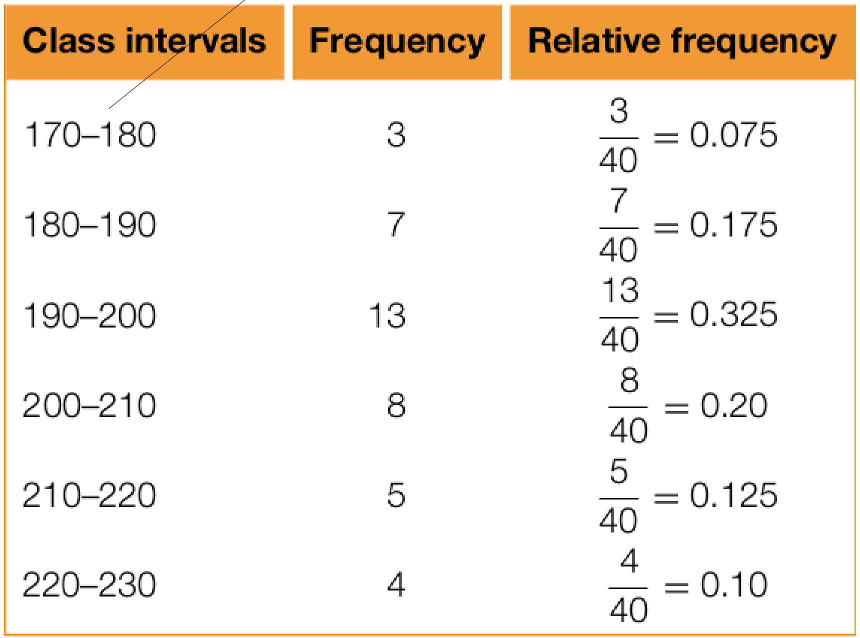
Used when the number of unique values are numerous

Sorted data

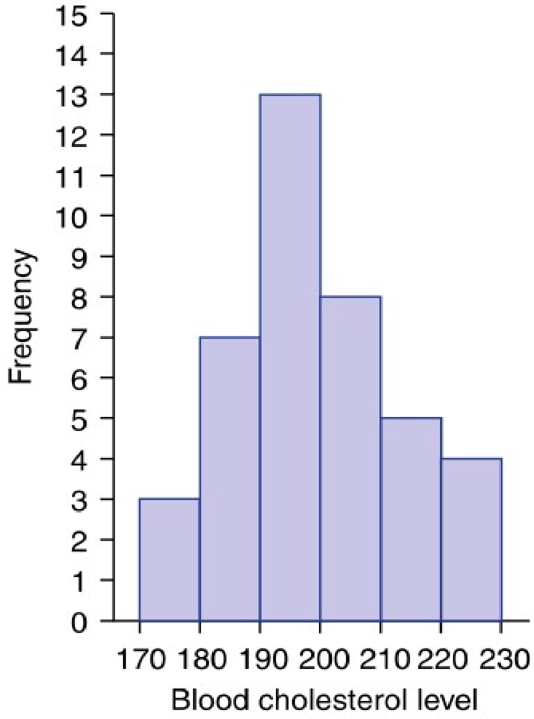


171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195,
196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, 227

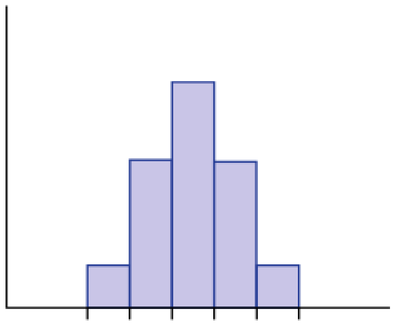
Class interval contains left-end,
but not right-end



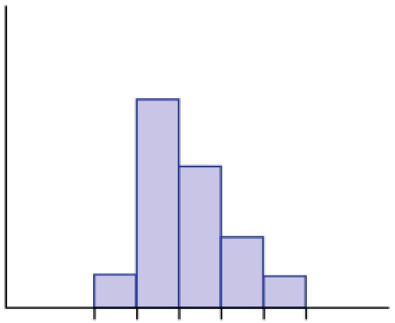
Class intervals	Frequency	Relative frequency
170–180	3	$\frac{3}{40} = 0.075$
180–190	7	$\frac{7}{40} = 0.175$
190–200	13	$\frac{13}{40} = 0.325$
200–210	8	$\frac{8}{40} = 0.20$
210–220	5	$\frac{5}{40} = 0.125$
220–230	4	$\frac{4}{40} = 0.10$



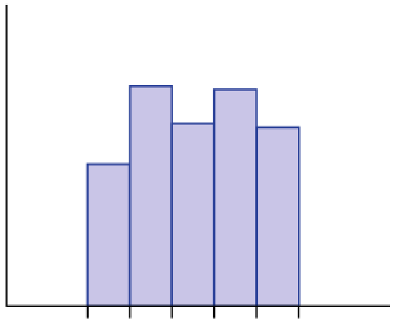
Types of histograms



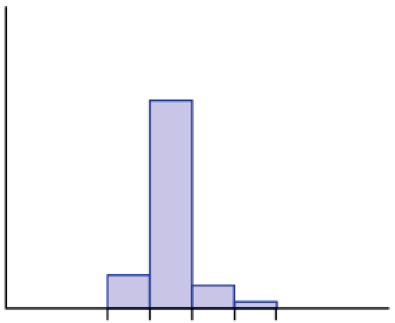
Symmetric



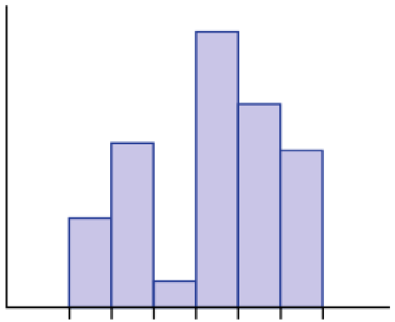
Not symmetric



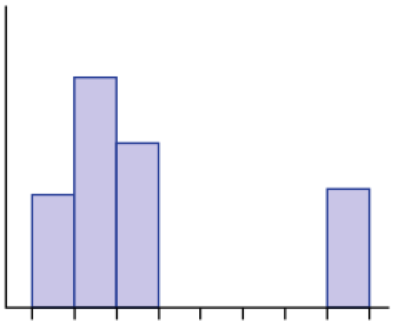
Spread out



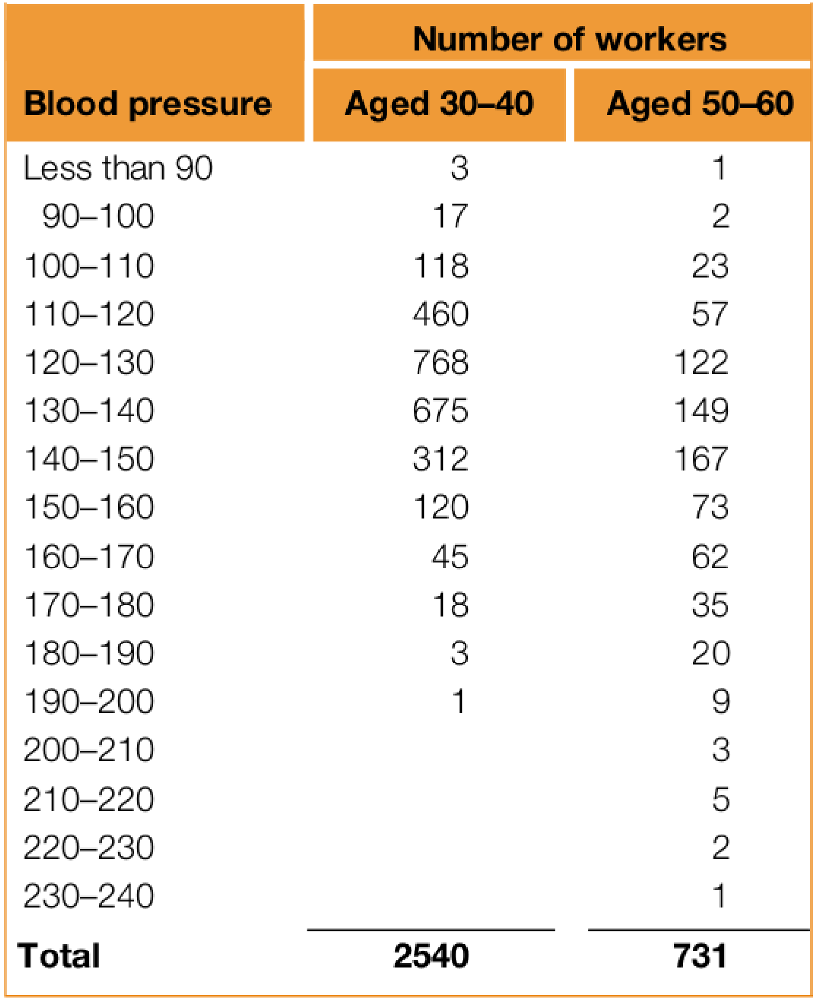
Concentrated



Gap



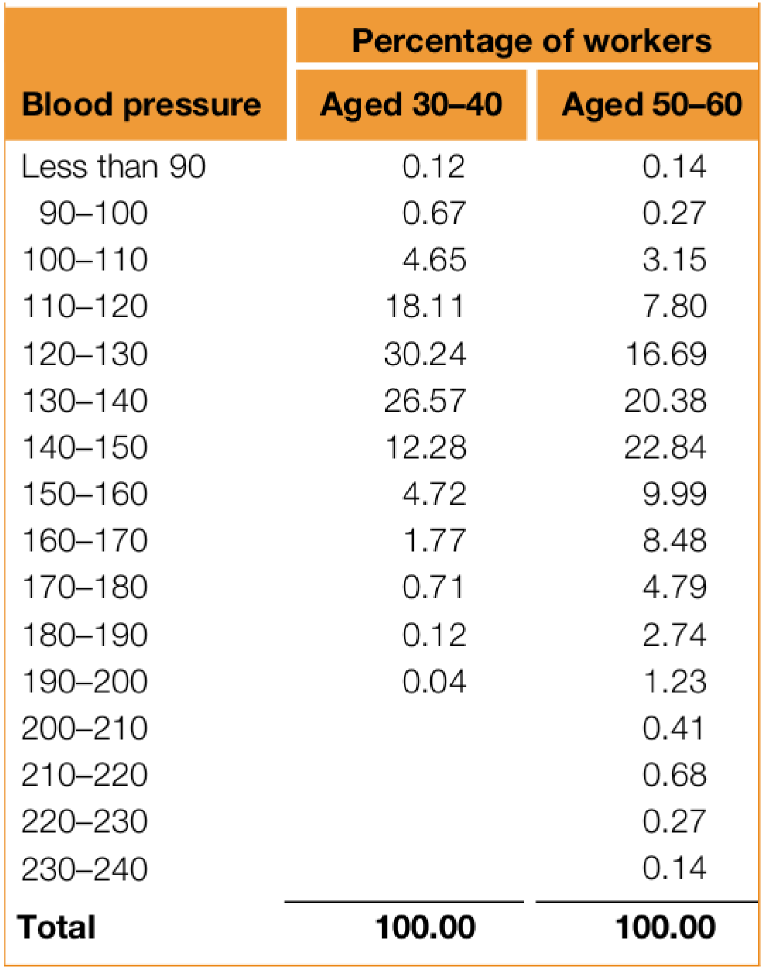
Data far from others



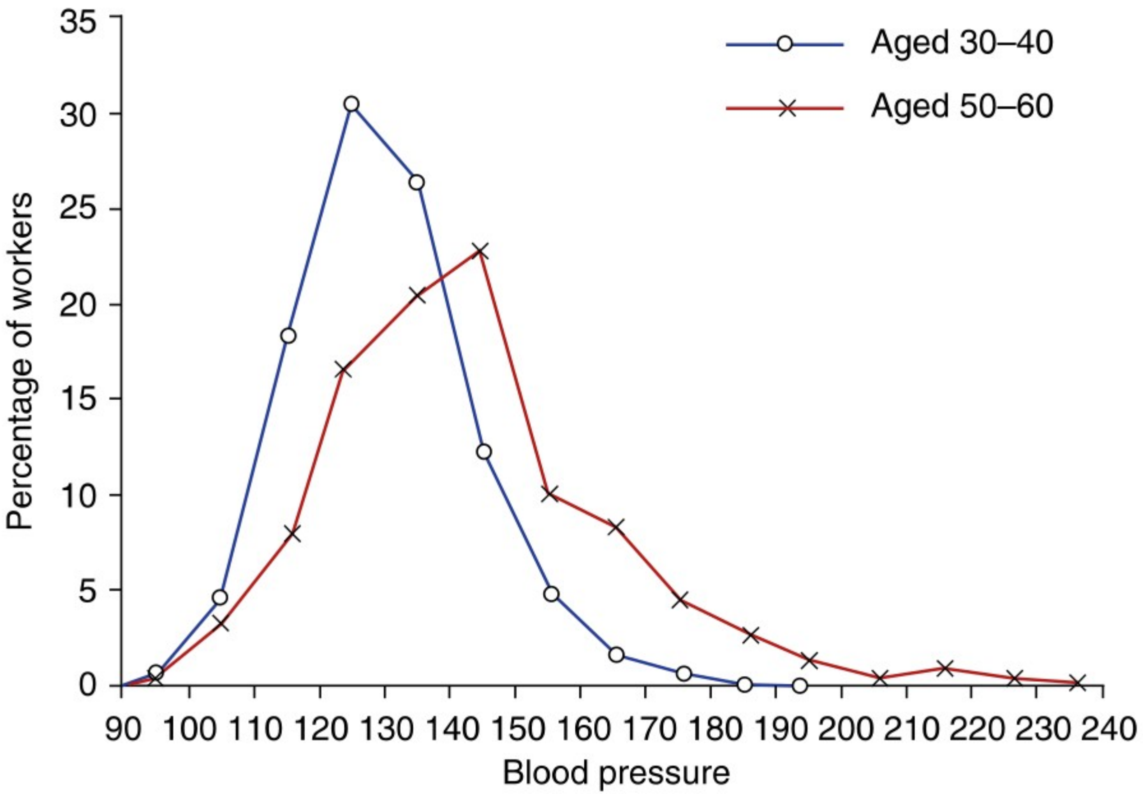
Blood pressure	Number of workers	
	Aged 30–40	Aged 50–60
Less than 90	3	1
90–100	17	2
100–110	118	23
110–120	460	57
120–130	768	122
130–140	675	149
140–150	312	167
150–160	120	73
160–170	45	62
170–180	18	35
180–190	3	20
190–200	1	9
200–210		3
210–220		5
220–230		2
230–240		1
Total	**2540**	**731**

Data: blood pressue values for two groups of workers

Number of samples are unequal

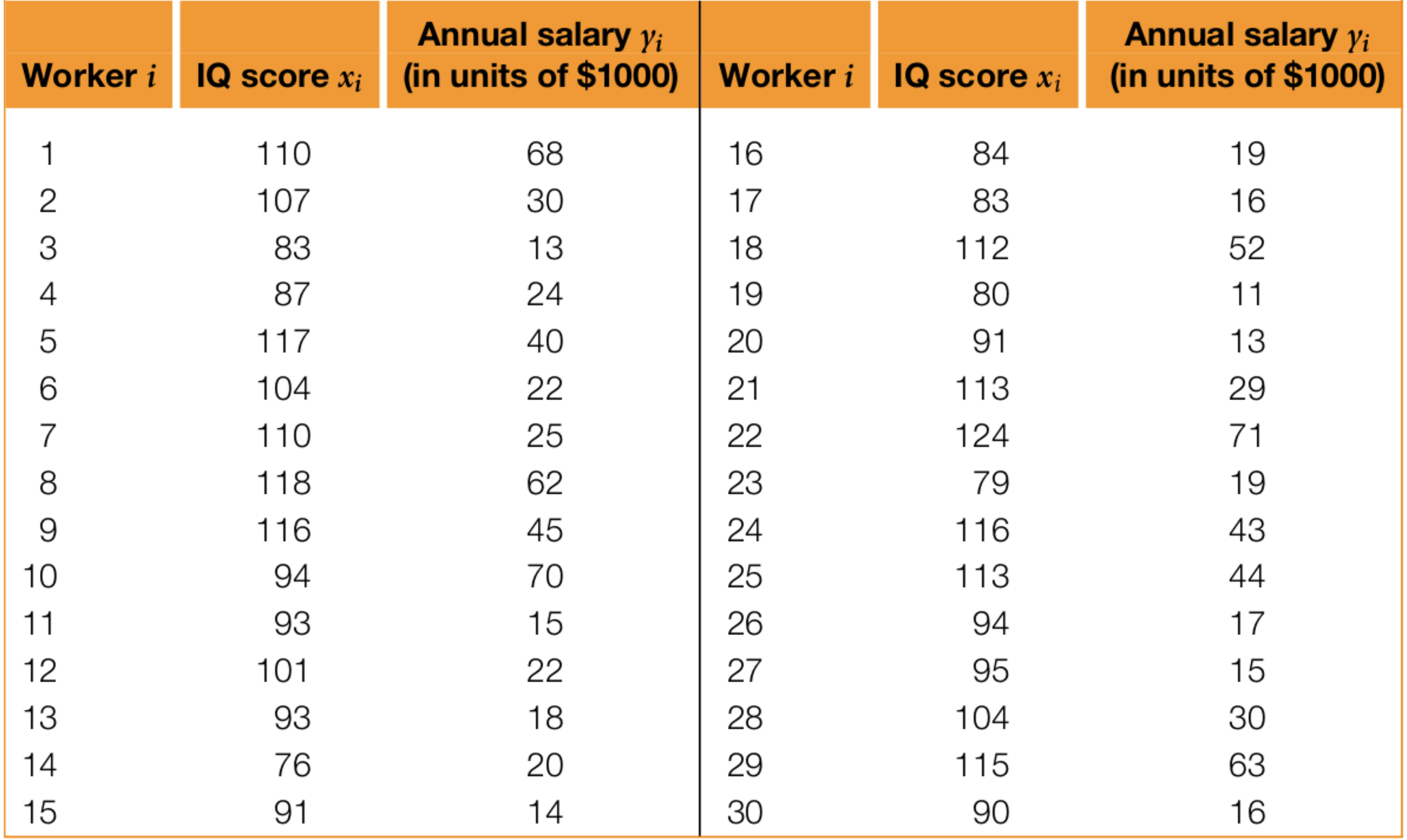


Blood pressure	Percentage of workers	
	Aged 30–40	Aged 50–60
Less than 90	0.12	0.14
90–100	0.67	0.27
100–110	4.65	3.15
110–120	18.11	7.80
120–130	30.24	16.69
130–140	26.57	20.38
140–150	12.28	22.84
150–160	4.72	9.99
160–170	1.77	8.48
170–180	0.71	4.79
180–190	0.12	2.74
190–200	0.04	1.23
200–210		0.41
210–220		0.68
220–230		0.27
230–240		0.14
Total	**100.00**	**100.00**



Relative frequency polygons

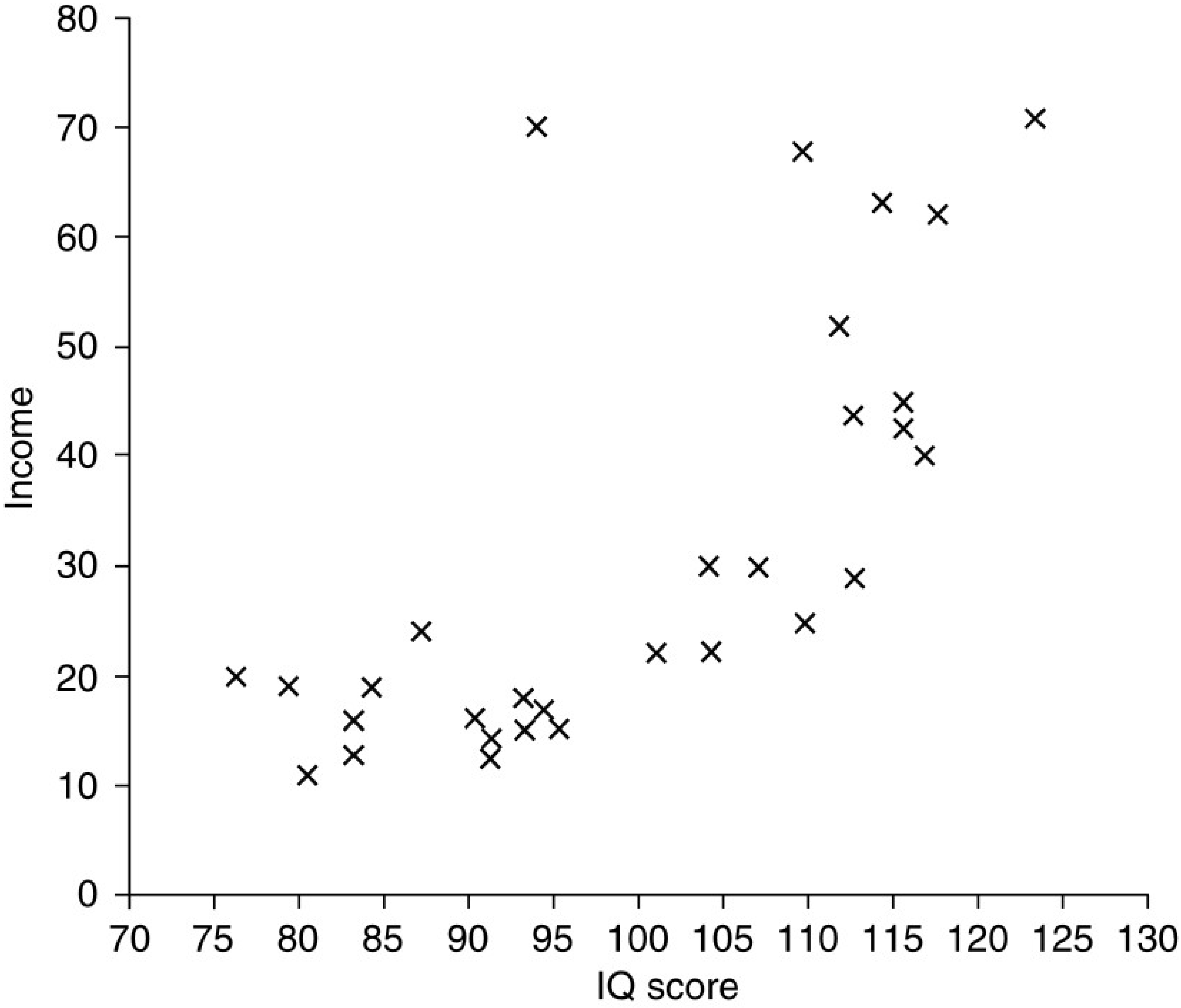
BP of older workers seem to be more spread out



Worker i	IQ score x_i	Annual salary y_i (in units of $1000)	Worker i	IQ score x_i	Annual salary y_i (in units of $1000)
1	110	68	16	84	19
2	107	30	17	83	16
3	83	13	18	112	52
4	87	24	19	80	11
5	117	40	20	91	13
6	104	22	21	113	29
7	110	25	22	124	71
8	118	62	23	79	19
9	116	45	24	116	43
10	94	70	25	113	44
11	93	15	26	94	17
12	101	22	27	95	15
13	93	18	28	104	30
14	76	20	29	115	63
15	91	14	30	90	16

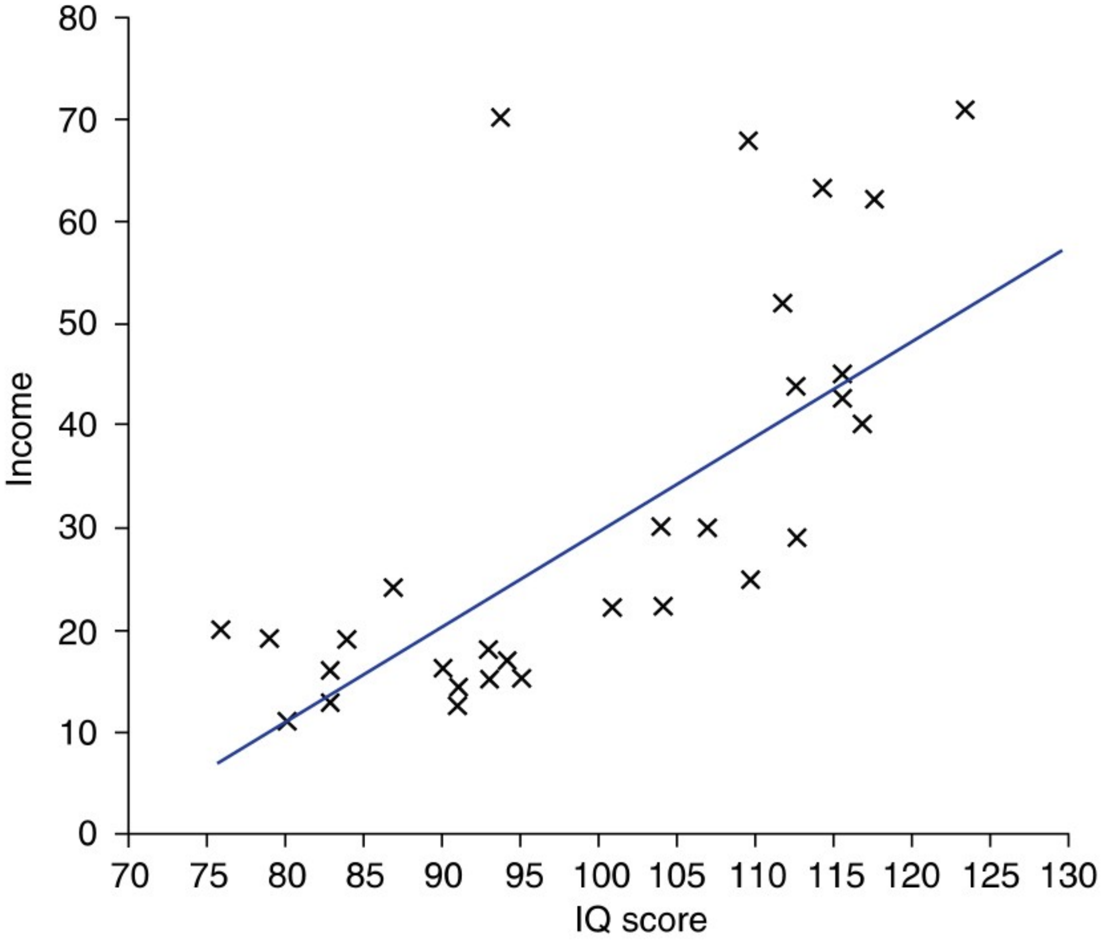
Data: salary vs IQ for employees

Is there a relationship between salary and IQ?



Scatter plot: IQ versus salary

Higher IQ seems to indicate higher salary, in general



Can even predict values

Can also see if there are outliers

Data error?

Can only be used for 2D data