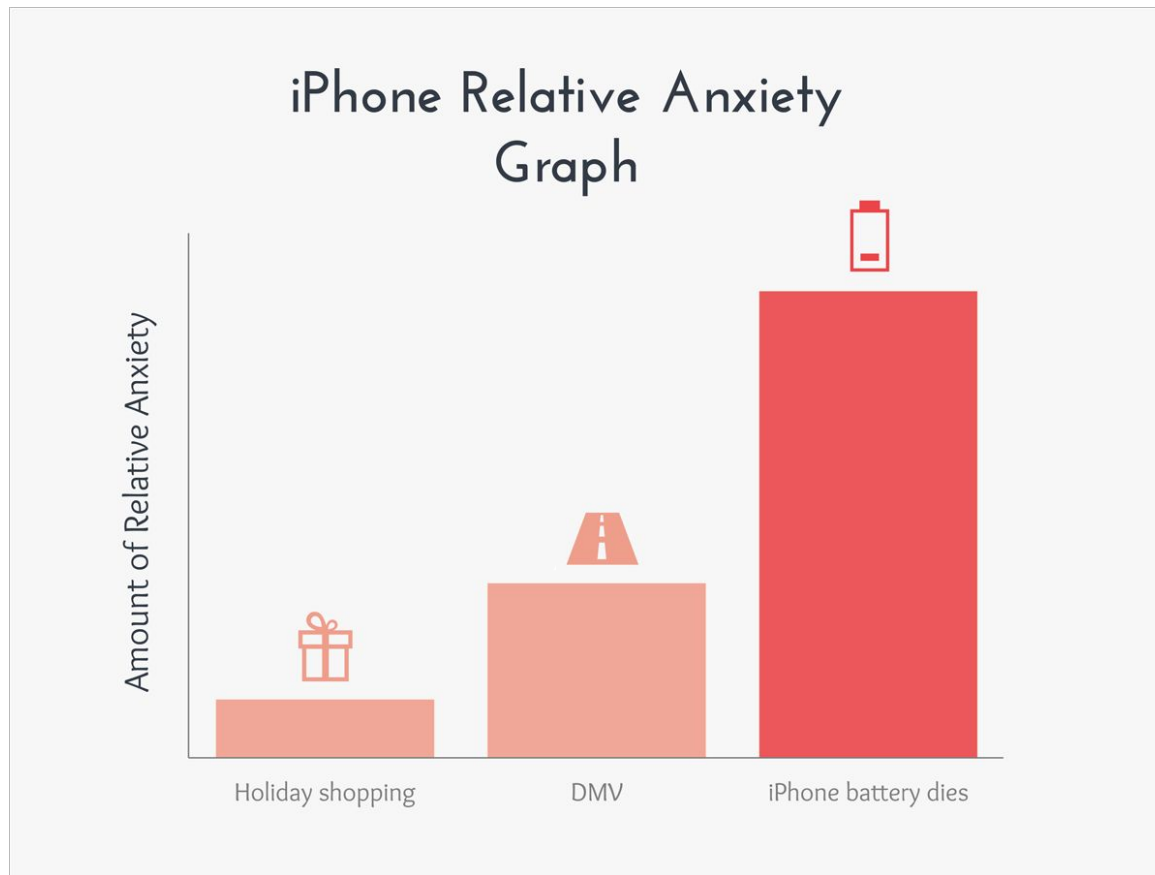


# Descriptive Statistics


## Graphing Techniques



# Points and grades from examination

---

No.	Points	Grade	No.	Points	Grade	No.	Points	Grade
1	15	1	12	12	3	23	15	2
2	17	1	13	16	2	24	9	4
3	19	1	14	13	1	25	17	1
4	10	2	15	7	3	26	16	1
5	2	2	16	15	1	27	13	1
6	14	2	17	20	2	28	6	2
7	5	4	18	16	2	29	16	3
8	17	2	19	14	3	30	18	1
9	11	1	20	3	2			
10	16	2	21	15	1			
11	10	3	22	12	1			

- 
- 
- Sample size  $n=30$
  - Data sorting → **Frequency table**
    - **both for quantitative and qualitative data**

# Exam grade

---

<b>Exam grade</b>				
	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<b>1</b>	12	40,0	12,0	40,0
<b>2</b>	11	36,7	23,0	76,7
<b>3</b>	5	16,7	28,0	93,3
<b>4</b>	2	6,7	30,0	100,0
Total	30	100,0		

## Notation

- Frequency ...  $n_i$
- Relative frequency ...  $f_i$
- Cumulative Frequency ...  $N_i$

$$N_i = \sum_{j \leq i} n_j$$

$$f_i = \frac{n_i}{n}$$

- Cumulative Percent ...  $F_i$

$$F_i = \sum_{j \leq i} f_j$$

# Points from class test

---

Points from class test					
Points	Frequency	Percent	Points	Frequency	Percent
2	1	3,33	13	2	6,67
3	1	3,33	14	2	6,67
5	1	3,33	15	4	13,33
6	1	3,33	16	5	16,67
7	1	3,33	17	3	10,00
9	1	3,33	18	1	3,33
10	2	6,67	19	1	3,33
11	1	3,33	20	1	3,33
12	2	6,67	<b>Total</b>	<b>30</b>	<b>100,00</b>



---

**Quantitative variables**



**Grouping into class intervals**

# How to select the intervals

---

- Number of intervals → in order to describe the characteristics of the data
- Simple recommendation
  - intervals of the same width

$$k = \sqrt{n}$$

k ... number of intervals

n ... sample size



...then

---

$$h = \frac{R}{k}$$

h ... width of interval

R ... Range =  $x_{\max} - x_{\min}$

k ... number of intervals

Our example:

$$n = 30$$

$$R = 20 - 2 = 18$$

$$k = \sqrt{30} = 5,48 \cong 6$$

$$h = \frac{18}{6} = 3$$

# Points from class test

Points from class test				
Interval	Frequency	Percent	Cumulative Frequency	Cumulative Percent
<b>5 and less</b>	3	10,0	3	10,0
<b>6-9</b>	3	10,0	6	20,0
<b>10-13</b>	7	23,3	13	43,3
<b>14-17</b>	14	46,7	27	90,0
<b>18 and more</b>	3	10,0	30	100,0
<b>Total</b>	30	100,0		



# Measures of Central Tendency

---

- Measures that represent with a proper value the tendency of most data to gather around this value
- Number of different measures of central tendency
  - *the arithmetic mean*
  - *the median*
  - *the mode*

# The arithmetic mean

---

## Notation

arithmetic mean .....  $\bar{X}$

- the sum of the values of a variable divided by the number of scores (by the sample size)

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n} = \frac{\sum_{i=1}^n x_i}{n}$$

# Properties of the arithmetic mean

---

1. it is expressed in the same unit of measure as the observed variable
2. it is the point in a distribution of measurements about which the sum of deviations are equal to zero

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

**Note:** deviation explains the distance and direction from a reference point – here *the arithmetic mean*, it is positive when the value is greater than the mean and negative when lower than the mean

3. the mean is **very sensitive** to extreme values

# Personal income (thousands CZK)

No.	$X_i$	$X_i - \bar{X}$	No.	$X_i$	$X_i - \bar{X}$
1	13,2	-12,62	9	16,4	-9,42
2	13,5	-12,32	10	17,2	-8,62
3	14,0	-11,82	11	19,0	-6,82
4	14,5	-11,32	12	25,8	-0,02
5	14,5	-11,32	13	27,0	1,18
6	15,2	-10,62	14	35,0	9,18
7	15,6	-10,22	15	35,5	9,68
8	16,2	-9,62	16	120,5	94,68
			$\Sigma$	413,1	0,00

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$

$$\bar{X} = \frac{13,2 + \dots + 120,5}{16} = \frac{413,1}{16} = 25,82 \text{ thousands CZK}$$

- 
- 12 of 16 values are below the arithmetic mean, because of the highest value  $x_{16}=120,5$  (*directors income*)
  - personal income is a commonly studied variable in which other measure of central tendency is preferred

# Other measures of central tendency

---

## ○ **The median....** $\tilde{x}$

The value above and below which one-half of the frequencies fall

- n...odd number

→ median case number =  $(n+1)/2$

- n...even number

→ the arithmetic mean of the two middle values

**Properties: Insensitive to extreme values**



# Other measures of central tendency

---

- **The mode....**  $\hat{x}$

The value that occurs with greatest frequency

- for qualitative (nominal and ordinal) and quantitative discrete data
- from a statistical perspective it is also the most probable value

# Personal income (thousands CZK)

---

n=16... even number

<b>No.</b>	$X_i$	<b>No.</b>	$X_i$
<b>1</b>	<b>13,2</b>	<b>9</b>	<b>16,4</b>
<b>2</b>	<b>13,5</b>	<b>10</b>	<b>17,2</b>
<b>3</b>	<b>14,0</b>	<b>11</b>	<b>19,0</b>
<b>4</b>	<b>14,5</b>	<b>12</b>	<b>25,8</b>
<b>5</b>	<b>14,5</b>	<b>13</b>	<b>27,0</b>
<b>6</b>	<b>15,2</b>	<b>14</b>	<b>35,0</b>
<b>7</b>	<b>15,6</b>	<b>15</b>	<b>35,5</b>
<b>8</b>	<b>16,2</b>	<b>16</b>	<b>120,5</b>

**the median**

**the mode**

# Personal income (thousands CZK)

n=16... even number

No.	$x_i$	No.	$x_i$
1	13,2	9	16,4
2	13,5	10	17,2
3	14,0	11	19,0
4	14,5	12	25,8
5	14,5	13	27,0
6	15,2	14	35,0
7	15,6	15	35,5
8	16,2	16	120,5

**the median**

**the mode**

$$\tilde{x} = \frac{x_8 + x_9}{2} = \frac{16,2 + 16,4}{2} = 16,3$$

$$\hat{x} = 14,5$$



# Use of mean, median and mode

---

## The arithmetic mean

- member of mathematical system in advanced statistical analysis
- preferred measure of central tendency if the distribution is not skewed

## The median

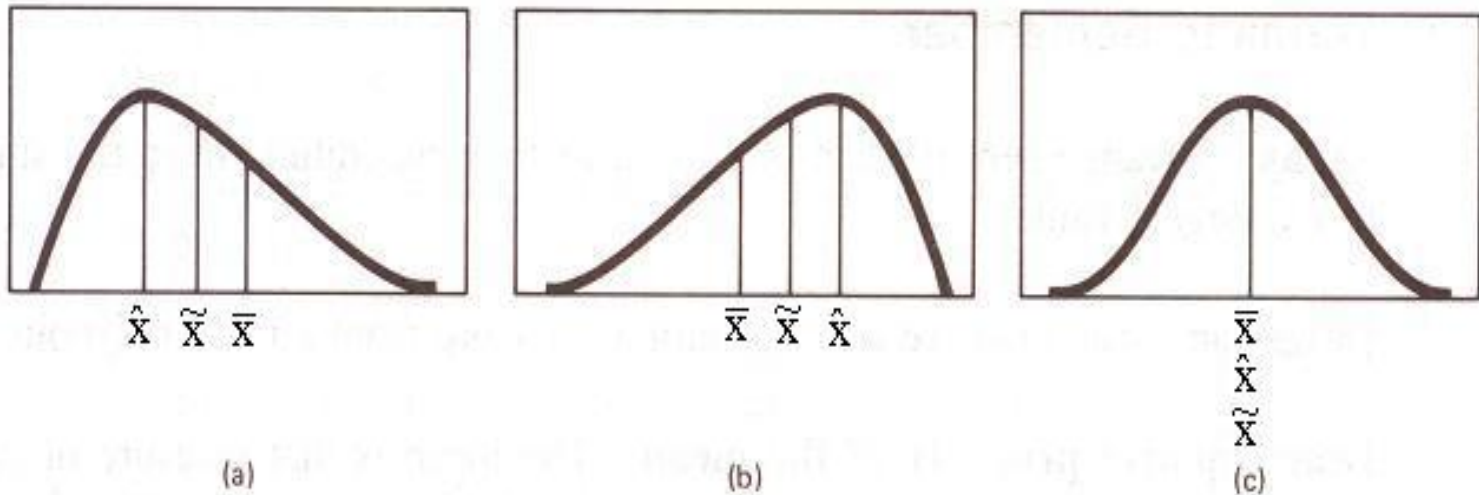
- when the distribution is skewed

## The mode

- whenever a quick, rough estimate of central tendency is desired

# The mean, median, mode and skewness

---



---

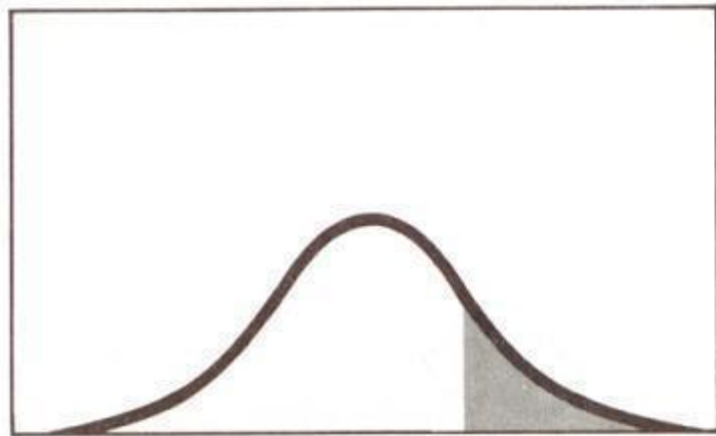
The relationship among the mean, median, and mode in (a) positively skewed, (b) negatively skewed, and (c) symmetrical distributions.



# Measures of Dispersion

---

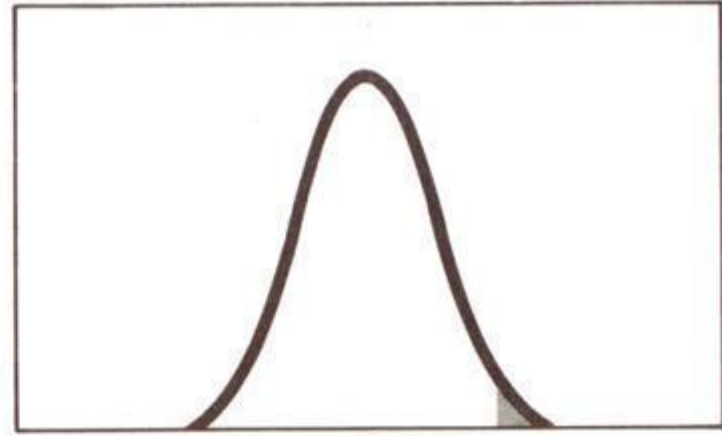
- to describe the spread of the data, its variation around a central value
- we want to express the distance along the scale of values



$\bar{X}$  110  
128

SCALE OF VALUES

(a)



$\bar{X}$  110  
128

SCALE OF VALUES

(b)

---

Two frequency curves with identical means but differing in dispersion or variability.

# The Range....R

---

- it is the distance between the largest and the smallest value

$$R = x_{\max} - x_{\min}$$

- it does not explain the variability inside the range !
- very simple and straightforward measure of dispersion



# The Variance... $s^2$

---

- it is an average squared deviation of each value from the mean
  - ▶ it is the sum of the squared deviations from the mean divided by  $n$
- when computing the variation based on **sample** we correct the calculation

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

# Working formulas

---


- For easier computation

*Formula 1*

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

*Formula 2*

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

- 
- 
- **the variance** explains both
    - the variability of the values around the arithmetic mean
    - the variability among the values
  - difficult interpretation  
(it is expressed in the squares of the unit of measure)

# The Standard Deviation...s

---

- it is the square root of variance
  - when computing the variation based on **sample**

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



# Properties of the standard deviation

---

- it is expressed in the same unit of measure as the observed variable
- the size of the standard deviation is related to the variability in the values
  - the more homogeneous values, the smaller SD
  - the heterogeneous values, the larger SD
- member of mathematical system in advanced statistical analysis (like the arithmetic mean)

# Two data sets with the same arithmetic mean and different SD

Array A

$x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
4	0	0
4	0	0
4	0	0
4	0	0
4	0	0

$$\sum_{i=1}^n x_i = 20 \quad \sum_{i=1}^n (x_i - \bar{x}) = 0 \quad \sum_{i=1}^n (x_i - \bar{x})^2 = 0$$

$$\bar{x} = 4 \quad n = 5$$

$$s = \sqrt{\frac{0}{5}} = 0$$

Array B

$x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2	-2	4
2	-2	4
3	-1	1
4	0	0
9	+5	25

$$\sum_{i=1}^n x_i = 20 \quad \sum_{i=1}^n (x_i - \bar{x}) = 0 \quad \sum_{i=1}^n (x_i - \bar{x})^2 = 34$$

$$\bar{x} = 4 \quad n = 5$$

$$s = \sqrt{\frac{34}{5}} = 2,6$$

# Example – Personal income (thousands CZK)

No.	$x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
<b>1</b>	<b>13,2</b>	<b>-12,62</b>	<b>159,2644</b>
<b>2</b>	<b>13,5</b>	<b>-12,32</b>	<b>151,7824</b>
...	...	...	...
<b>16</b>	<b>120,5</b>	<b>94,68</b>	<b>8 964,3024</b>
		<b><math>\Sigma</math></b>	<b>10 370,04</b>

$$s^2 = \frac{10370,04}{16 - 1} = 691,3363$$

$$s = \sqrt{s^2} = \sqrt{691,3363} = 26,2938 \quad \text{thousands CZK}$$

# Coefficient of Variation...V


---

- the ratio of the standard deviation to the mean

$$V = \frac{S}{\bar{X}}$$

- often reported as a percentage (%) by multiplying by 100



- 
- 
- it is a relative measure of dispersion
  - used when comparing two data sets with different units or widely different means
  - values higher than 50% indicate large variability

# Example – Personal income (thousands CZK)

No.	$x_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
<b>1</b>	<b>13,2</b>	<b>-12,62</b>	<b>-159,2644</b>
<b>2</b>	<b>13,5</b>	<b>-12,32</b>	<b>-151,7824</b>
...	...	...	...
<b>16</b>	<b>120,5</b>	<b>94,68</b>	<b>8 964,3024</b>
		<b><math>\Sigma</math></b>	<b>10 370,04</b>

$$s = 26,2938 \quad \bar{x} = 25,82$$

$$V = \frac{s}{\bar{x}} = \frac{26,2938}{25,82} = 1,01835$$

$$V = 1,01835 * 100 = 101,835\%$$



# Percentiles (Centiles)

---

- value below which a certain percent of observations fall
- scale of percentile ranks is comprised of 100 units
- insensitive to extreme values

# Deciles

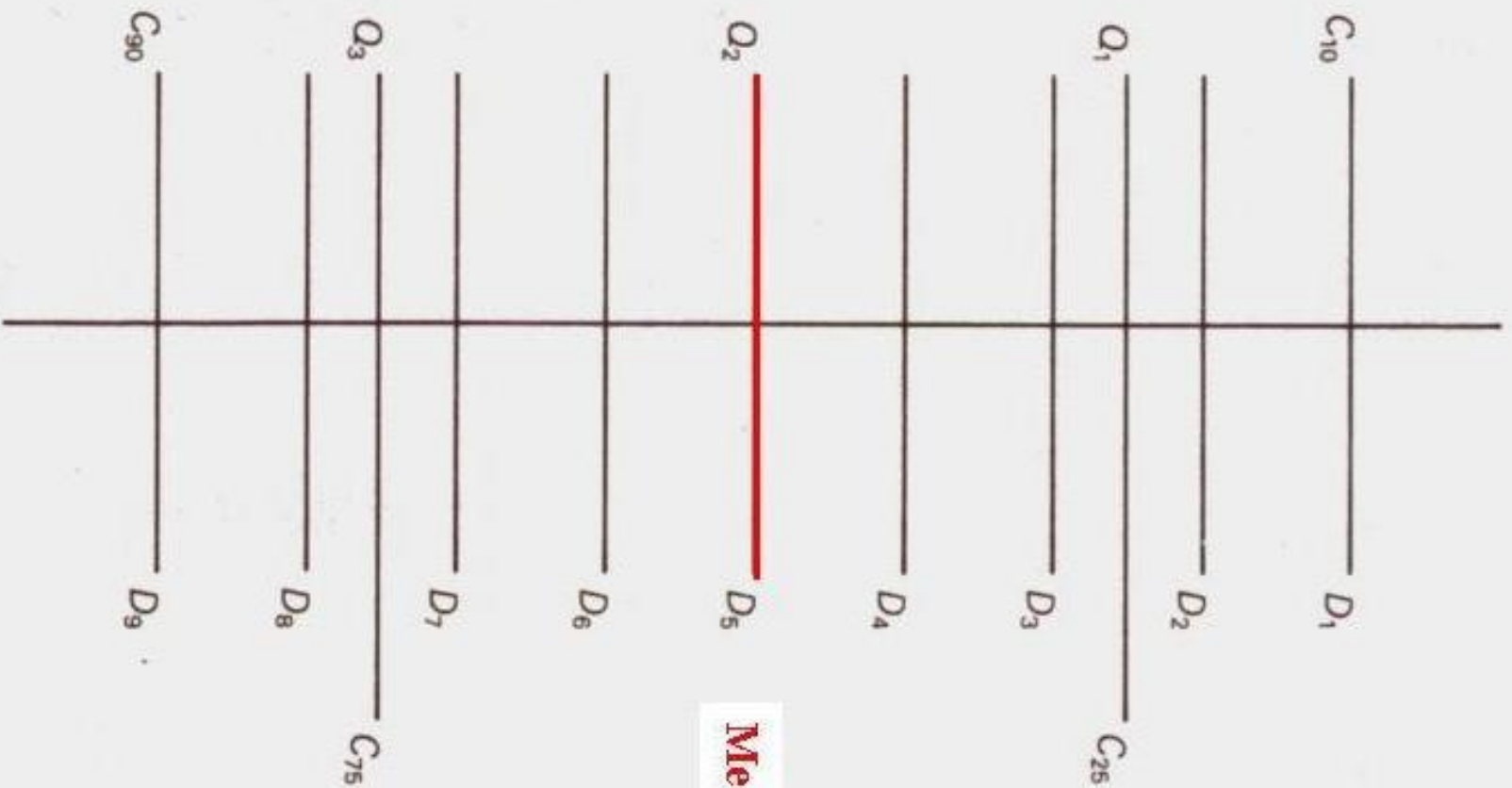
---

- divides a distribution into 10 equal parts
- there are 9 deciles
- $D_1$  - 1st decile
  - 10 percent of values fall below it
- $D_9$  - 9th decile
  - 90 percent of values fall below it

# Quartiles

---

- divides a distribution into 4 equal parts
  - $Q_1$  - 25 percent of values fall below it
    - 25th centile
  - $Q_2$  - 50 percent of values fall below it
    - 50th centile
  - $Q_3$  - 75 percent fall below it
    - 75th centile



**Median**



---

# **Graphing Techniques**



# Constructing graphs – Bar graph

---

- x – axis: labels of categories
- y – axis: frequency (relative frequency)

*The height of each rectangle is the category`s frequency or relative frequency.*



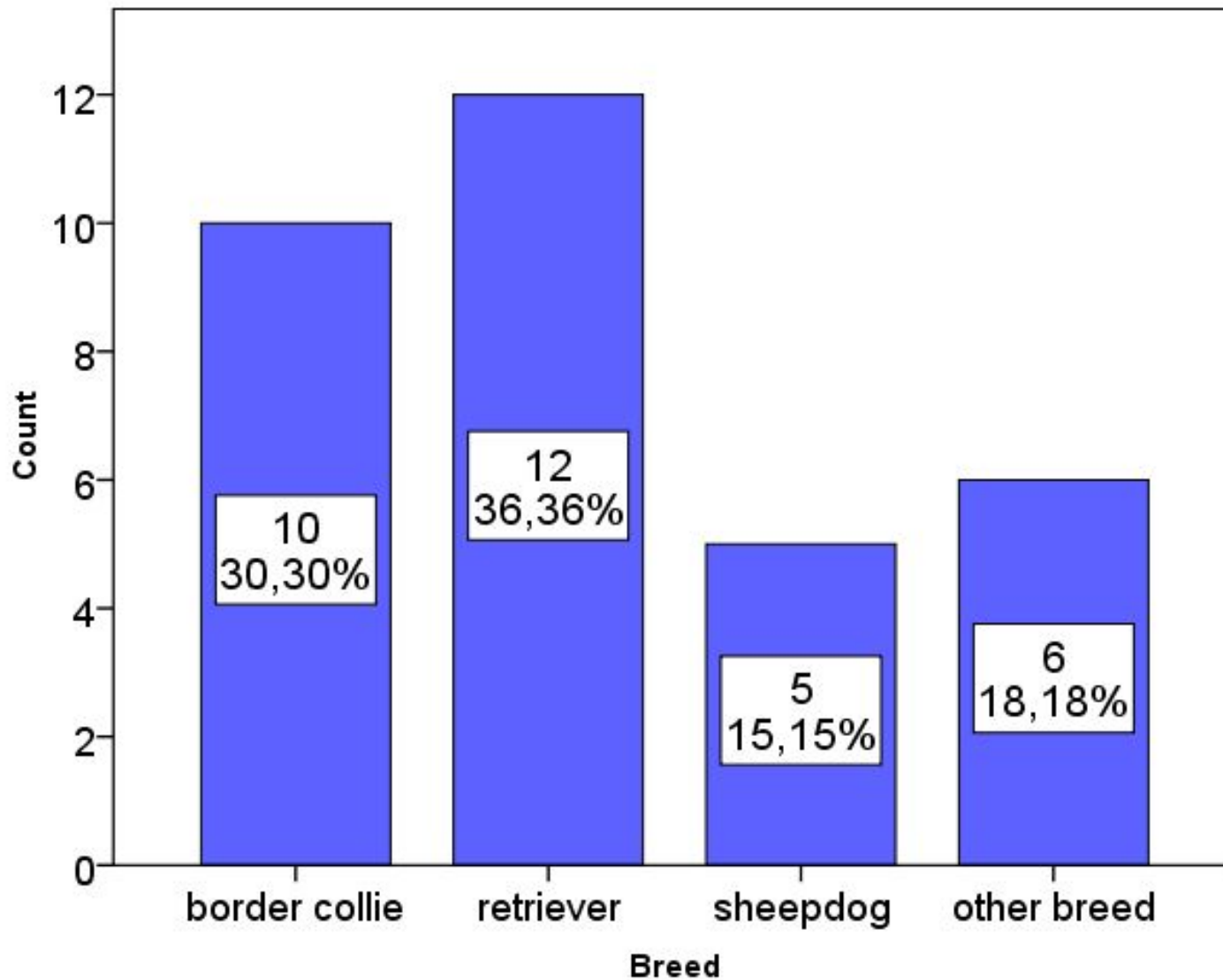


# Arranging the graph

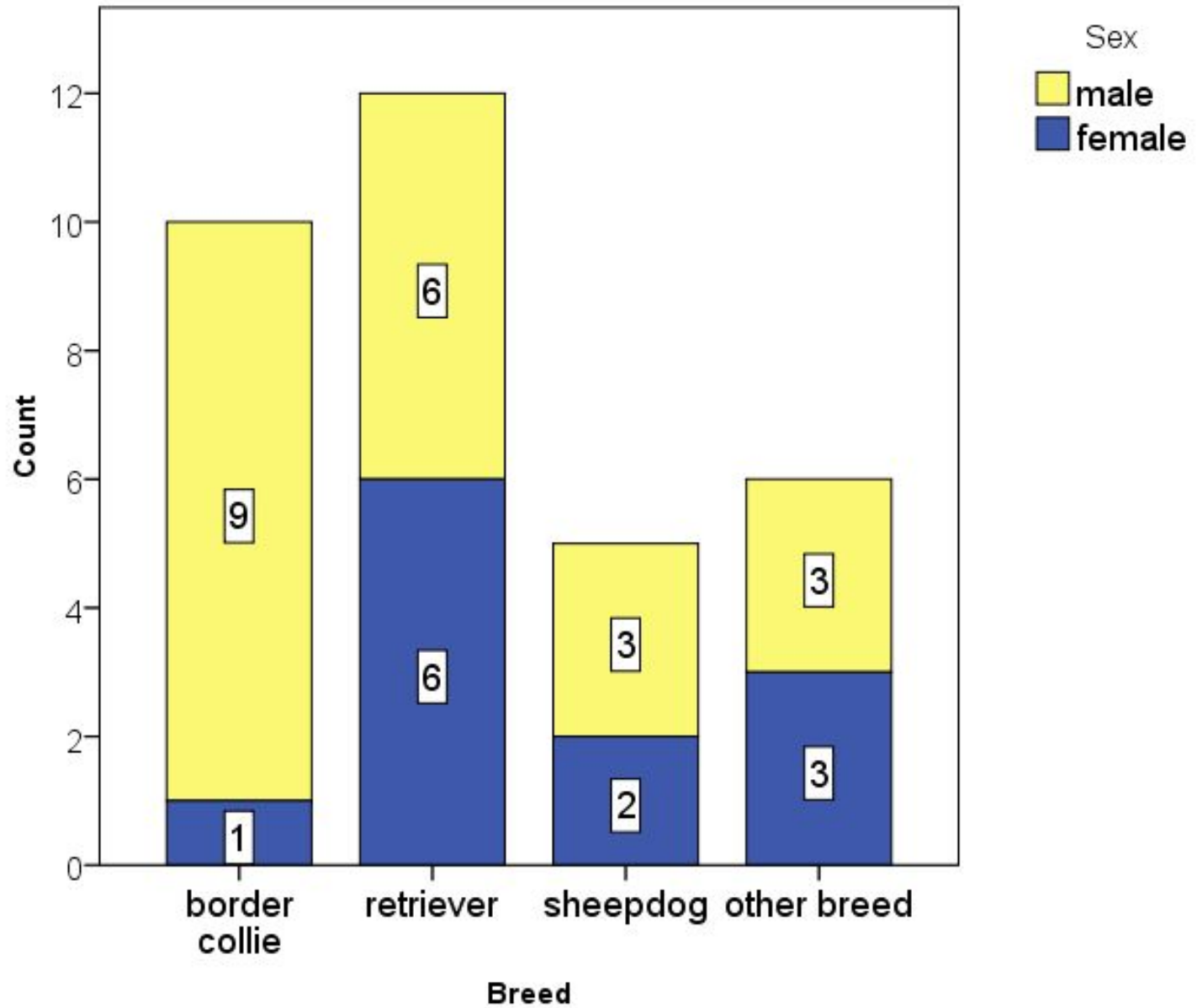
---

- **nominal variables** – we can arrange the categories in any order: alphabetically, decreasing/increasing order of frequency
- **ordinal variables** – the categories should be placed in their naturally occurring order

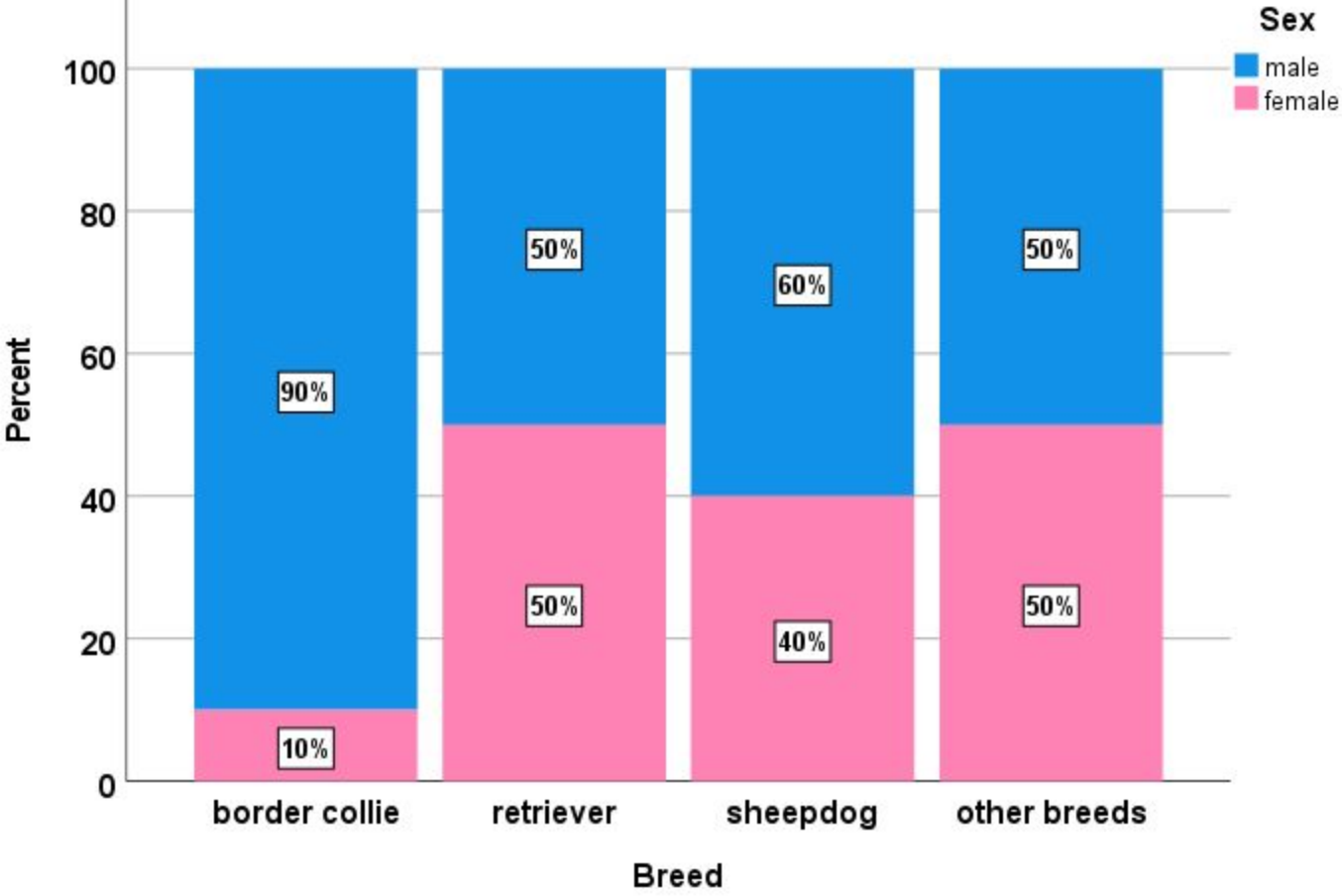
### Guide Dogs - frequency of breeds



Guide Dogs - frequency of breeds



Stacked Bar Percent of Breed by Sex

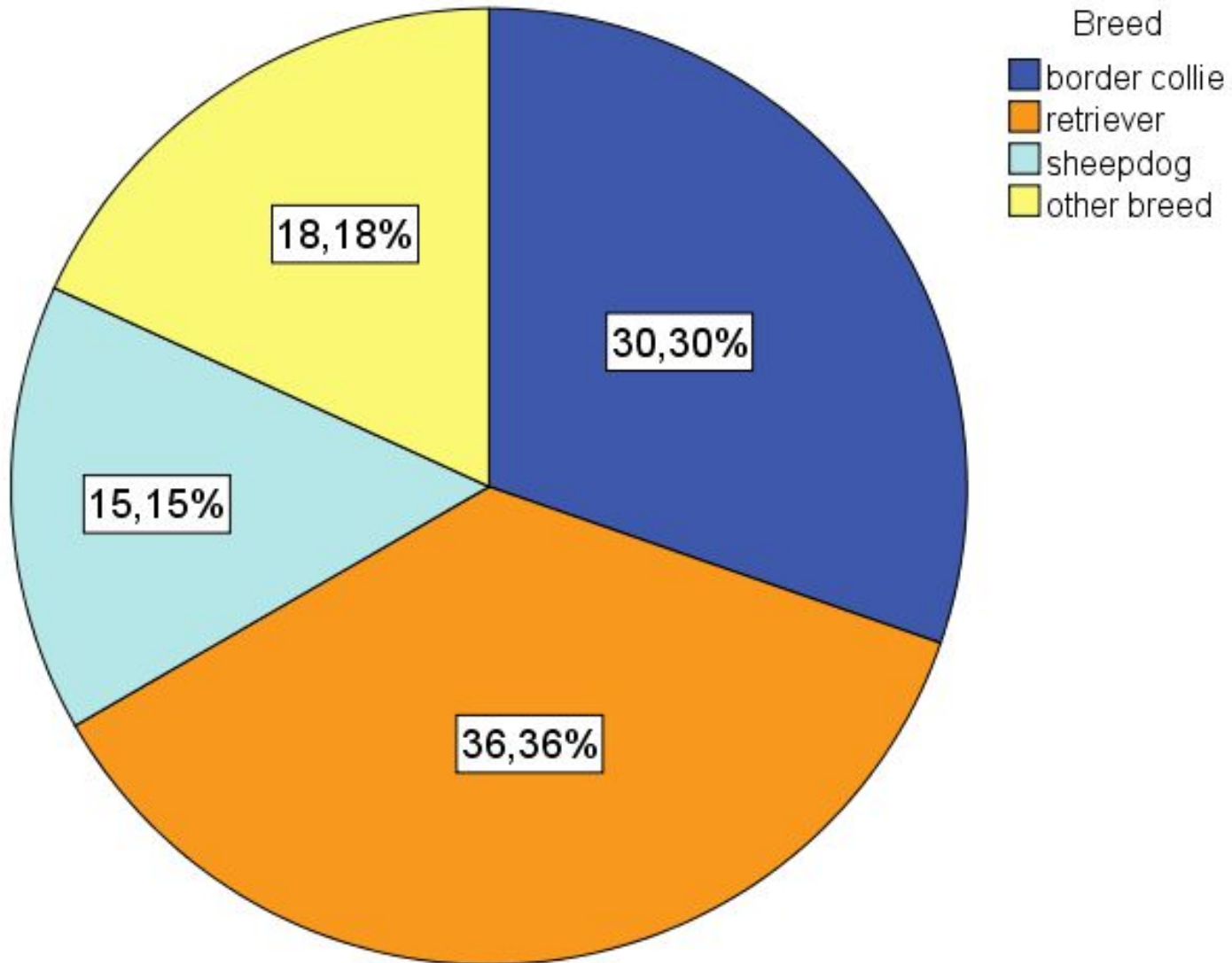


# Constructing graphs – Pie graph

---

- **Pie chart** – a circle divided into sectors
  - each sector represents a category of data
  - the area of each sector is proportional to the frequency of the category

## Guide Dogs - frequency of breeds



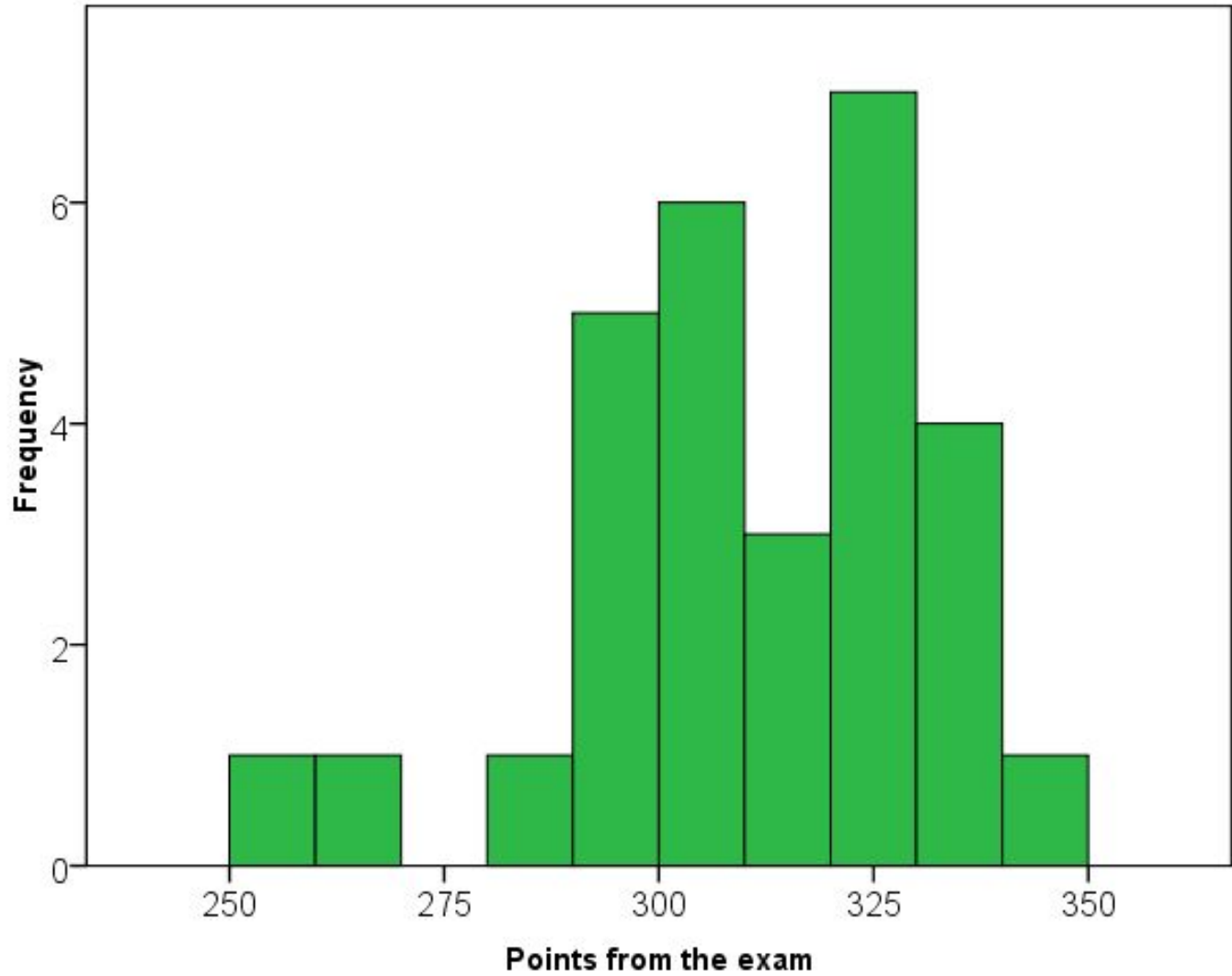


# Constructing graphs – Histogram

---

- bar graph for **quantitative** data
- values are grouped into intervals (classes)
- constructed by drawing rectangles for each class of data
- the height of each rectangle is the frequency of the class
- the width of each rectangle is the same

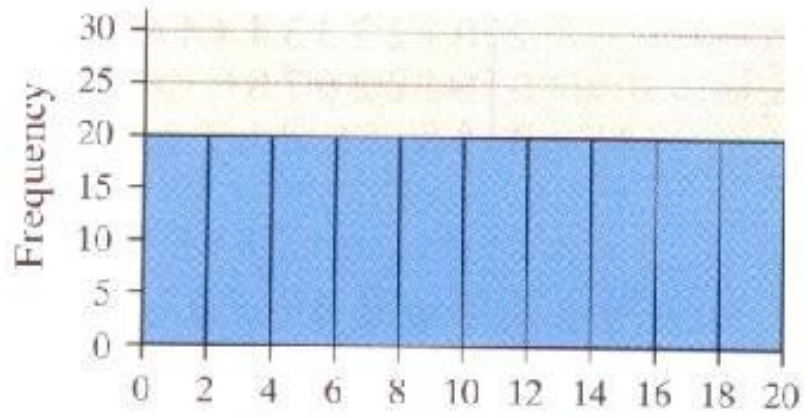
**Points from the exam - histogram**



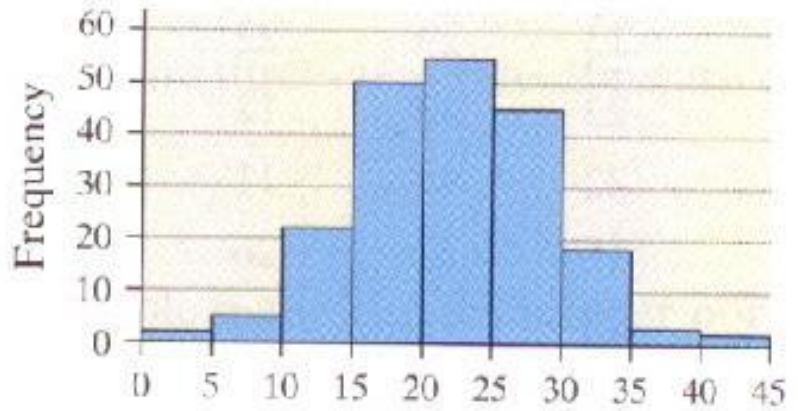


# Histogram

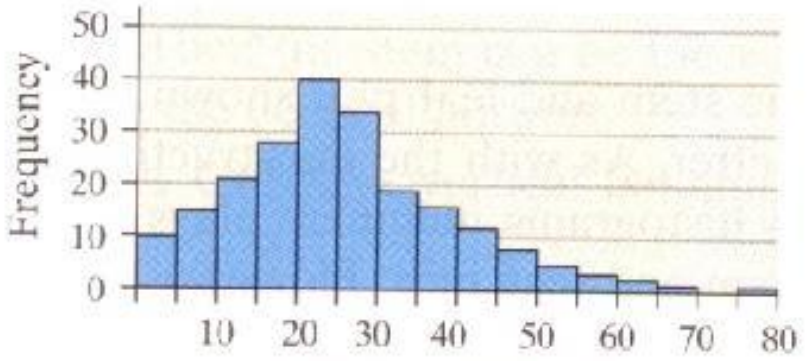
Figure 15



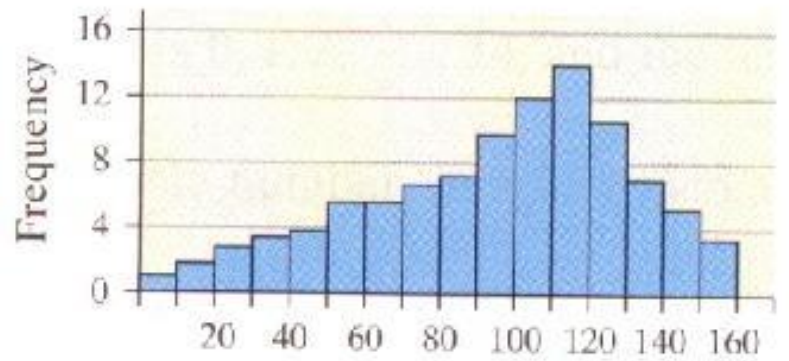
(a) Uniform (symmetric)



(b) Bell-shaped (symmetric)

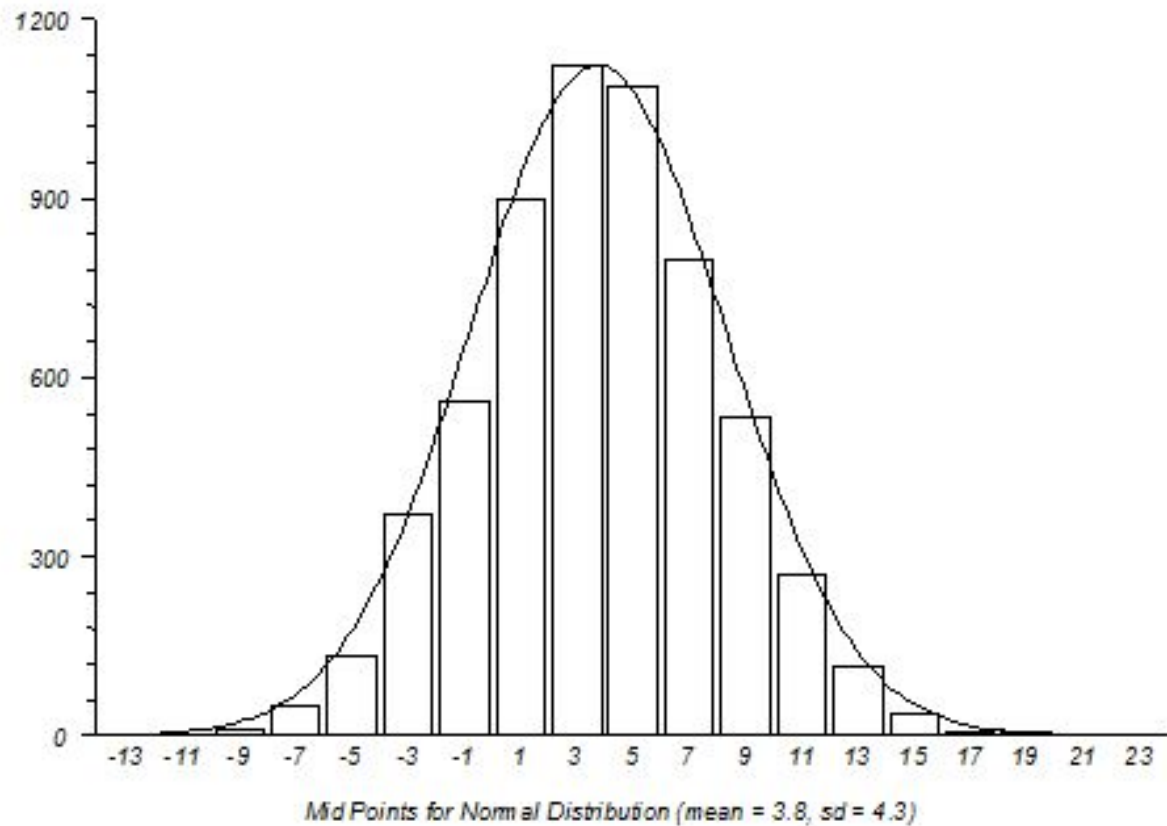


(c) Skewed Right



(d) Skewed Left

Histogram for Normal Distribution (mean = 3.8, sd = 4.3)



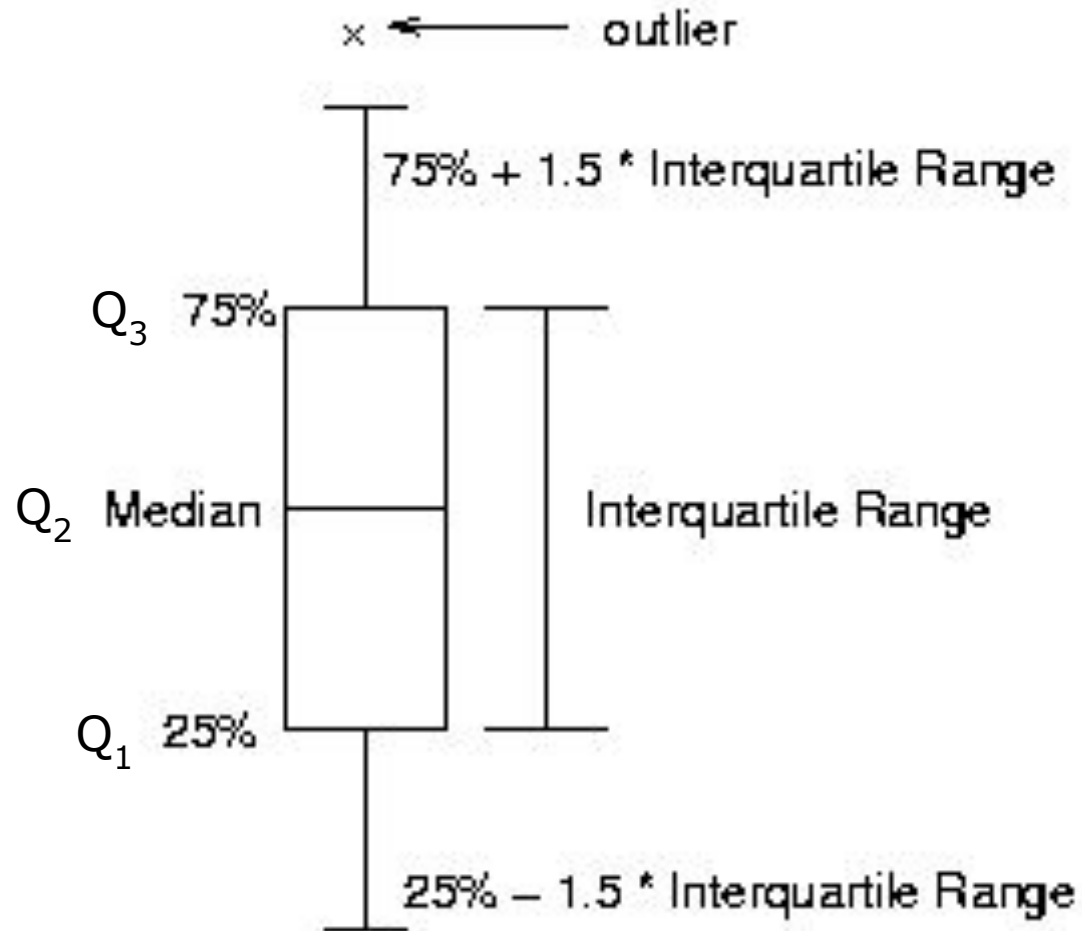


# Constructing graphs – **Boxplot**

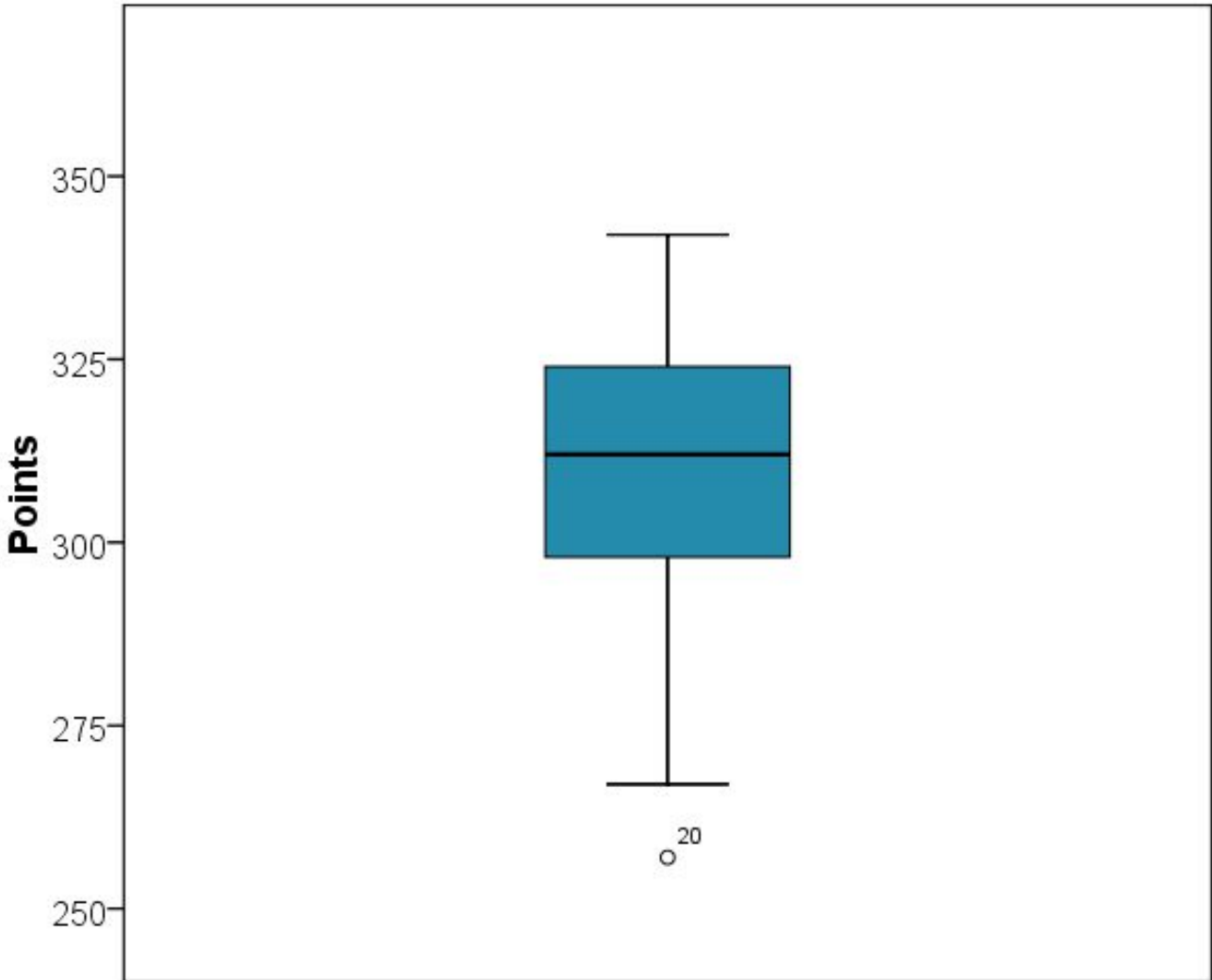
---

- box-and-whisker diagram
- five number summary

# Boxplot



**Guide Dogs - points from the exam**



**Guide Dogs - points from the exam**

