2.4

- a measure of the center is not enough to describe a distribution well.
- range : difference between the largest and the smallest observations.
- introduce some notation. as before, we have the observations denoted

  $$x_1, x_2, \ldots, x_n$$

  these observations may not be ordered. the index usually refers to the order in which the observations were collected. when we are looking at an ordered sample, we denote the observations

  $$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

- range is calculated as $x_{(n)} - x_{(1)}$.
- example from the text : music teacher salaries Denmark vs USA. see figure 2.10 in the text.

  $$\begin{aligned} \text{Denmark} &= 45 - 35 = 10 \\ \text{USA} &= 60 - 20 = 40 \end{aligned}$$

- the range is simple to compute and simple to understand. uses only the extreme values and ignores the other values.
- the range shares the worst properties of the mean and the worst property of the median. it is both affected very badly by outliers and ignores the numerical values of nearly all the data.

  $$\begin{aligned} \text{EN} &= 199 - 1.34 = 197.66 \\ \text{SB} &= 136 - 0 = 136 \end{aligned}$$

- a better measure of spread is based on deviations from the mean. we use the notation

  $$x_i - \bar{x}, \quad i = 1, 2, \ldots, n$$

  for the deviation.

  - a deviation is positive when the observation falls above the mean and negative when the the observation falls below the mean
  - the interpretation of the mean as the balance point implies that the positive deviations balance out the negative ones. this implies

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

a simple example might help

$$\begin{aligned}\boldsymbol{x} &= (1,2,3,4,5)\\ \bar{x} &= 3\\ \boldsymbol{x}-\bar{x} &= (-2,-1,0,1,2)\end{aligned}$$

obviously they are not always balanced like this, another example

$$\begin{aligned}\boldsymbol{x} &= (8,1,5,6,4)\\ \bar{x} &= 4.8\\ \boldsymbol{x}-\bar{x} &= (3.2,-3.8,0.2,1.2,-0.8)\end{aligned}$$

- the average of the squared deviations is called the variance. the variance uses the square of the units of measurement of the original data, the standard deviation is easier to interpret.

$$\text{Sum of Squares} = \sum_{i=1}^{n}(x_i-\bar{x})^2$$

the equation for the standard deviation is given as

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

- the standard deviation is the typical distance of an observation from the mean
  - the greater the spread of the data the larger s
  - can s be negative?
  - can s be zero?
  - s can be influenced by outliers since it uses both the mean and all of the data (including outliers)
- why do we divide by $n-1$ instead of n?

  the deviations have only $n-1$ pieces of information about variability. the deviations must sum to 0, so that if we knew $n-1$ of them, the last would be completely determined.
- example 13 in the text
- the empirical rule : if the distribution is unimodal and symmetric (bell shaped), the data will have the following properties
  - 68% of the observations will fall within one standard deviation of the mean ($\bar{x} \pm s$)

- 95% of the observations fall within 2 standard deviations ($\bar{x} \pm 2s$)
- all or nearly all of the data will fall within 3 standard deviations ($\bar{x} \pm 3s$)

- problem 2.47
- example 15

2.5

- percentiles : the p^{th} percentile is the value of the observation such that p percent of the observations fall at or below that value
- the 50th percentile is referred to as the median
- quartiles the first quartile (Q1) has $p = 25$, the second quartile is $p = 50$ and the third $p = 75$.
- example 16 to find the quartiles
- a good measure of spread is the interquartile range (IQR) given as

$$IQR = Q3 - Q1$$

- example : college data

EN				
0%	25%	50%	75%	100%
1.34	30.00	43.49	58.00	199.00

SB				
0%	25%	50%	75%	100%
0	2.4	6.34	18.6675	136

Then we have

$$IQR_{EN} = 28$$

$$IQR_{SB} = 16.2675$$

- detecting outliers using the $1.5 \times IQR$ criteria : an observation is a potential outlier if it falls more than $1.5 \times IQR$ below Q1 or $1.5 \times IQR$ above Q3
- constructing a box plot

1. a box is drawn from Q1 to Q3
2. a line goes inside the box at Q2 (median)
3. a line goes from the lower end of the box to the smallest observation that is not a potential outlier (using the $1.5 \times IQR$ criteria). these lines are called whiskers.

4. the potential outliers are shown separately as dots.

- example 17
- boxplots are good at showing the difference in distribution between two or more samples.
- outliers can be the most interesting part of the sample
- z-score : using the empirical rule we can determine if an observation is an outlier using the z-score.

$$z = \frac{x_i - \bar{x}}{s}$$

- an observation is a potential outlier if $z \geq 3$.
- problem 2.71

3.1

- association : exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable.
- response variable : is the outcome variable on which comparisons are made.
- explanatory variable : defines the groups to be compared with respect to values on the response variable
- example : number of marriages from previous chapter. the example says nothing about divorce. if we wanted to examine the affect of gender on divorce, then we would say that the response variable is divorce and the explanatory variable is gender.
- each row/column pair is called a cell.
- grade school goals : students were surveyed (1992) in Michigan for a study. students indicated which of grades, popularity and sports was most important, students were in grades 4 - 6.

	Grades	Popular	Sports	Total
Boy	117	50	60	227
Girl	130	91	30	251
Total	247	141	90	478

given gender, how likely is each of the goal categories? in this case, goal is the response variable and gender the explanatory variable.

	Grades	Popular	Sports
Boy	0.515	0.22	0.264
Girl	0.518	0.362	0.119

given the goal category, how likely is each gender? in this case, gender is the response variable and goal is the explanatory variable.

	Grades	Popular	Sports
Boy	0.473	0.354	0.667
Girl	0.526	0.645	0.333

what would the table look like if there was no association?

- are we restricted in the number of response variables? explanatory variables?

		Grades	Popular	Sports
Boy	4	29	13	12
	5	43	19	24
	6	45	18	24
Girl	4	34	18	13
	5	45	36	9
	6	51	37	8

- contingency table : a display for two categorical variables. its rows list the categories of one variable and its columns list the categories of the other variable. each entry in the table is the frequency of cases in the sample with certain outcomes on the two variables.

- problems 3.1 and 3.2