Summary Statistics - Chapter 5

5.1 **mode** - measure of central tendency

5.1.1 **median** - center of a data set
- if there are odd # in set => middle data point
- if there are even # in set => midway between these 2 pts.

ex. 1, 2, 10, 11, 13, 19 - median $= \frac{10+11}{2} = \underline{\underline{10.5}}$

more formally: position $= (0.5)\cdot(n+1)$
$n =$ # of data points

ex. 5.1 median of

1	2	3	4	5	6	7	8	9
1,	3,	4,	6,	7,	9,	11,	15,	19

position $= (0.5)(9+1) = 5$

5.1.2 **mean (average)**

$$\bar{y}_n = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1 + y_2 + y_3 + \dots + y_{n-1} + y_n}{n}$$

ex 5.2 mean of $\frac{3+9+4+19+11+6+7+1+15}{9} = \frac{75}{9} = \underline{\underline{8.3}}$

What is Sigma?

$$\sum_{n=1}^{4} n = \underline{1+2+3+4 = 10}$$

what to sum

start at this value

go to this value (4)

ex 5.3

1) a) $\sum_{k=1}^{3} k = 1+2+3 = 6$

b) $\sum_{k=1}^{3} [k-1] = (1-1)+(2-1)+(3-1) = \underline{\underline{3}}$

c) $\sum_{k=1}^{3} k^2 = 1^2+2^2+3^2 = 1+4+9 = \underline{\underline{14}}$

d) $\sum_{k=1}^{3} \frac{1}{k} = \frac{1}{1}+\frac{1}{2}+\frac{1}{3} \quad \frac{6+3+2}{6} = \frac{11}{6}$

$= 1,8_3$

2) set y_k w/ $n=5$; $y_1=4$, $y_2=2$, $y_3=3$, $y_4=2$
$y_5=4$

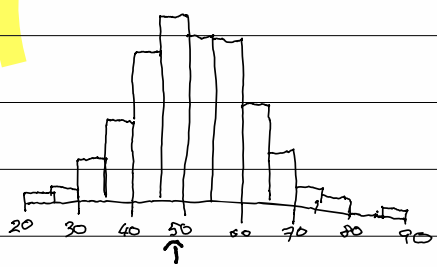
a) $\frac{1}{5}\sum_{k=1}^{5} y_k = \frac{1}{5}(4+2+3+2+4) = \frac{15}{5} = \underline{\underline{3}}$

b) $\sum_{k=1}^{5}[y_k-3] = (4-3)+(2-3)+(3-3)+(2-3)+(4-3)$
$= \cancel{1+(-1)}+0+\cancel{(-1)+1} = \underline{\underline{0}}$

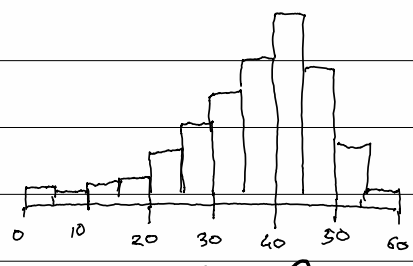
5.1.3 median vs. mean

- median divides data into two halves
- mean is more of a point of a balance
- median is more robust to change compared to mean

ex. 5.5



mean
median
mode



mean med. mode

- for a symmetric histogram and frequency distribution

mean = median = mode

- for right-skewed hist. and freq. dist.

mode < median < mean

- for left-skewed hist. and freq. dist

mean < median < mode

5.2 Quartiles and box plot

5.2.1 Quartiles

- divide data into quartiles
  - lowest 25% of data
  - second 25%-50% data

Q = quartiles Q_1 = 25% lowest, Q_2 = rest 25%

Q_2 = median of whole data set

Q_1 and Q_3 medians of lower half and upper half of the data set

5.2.2 Interquartile range (IQR)

- difference between the Q_1 and Q_3

$$IQR = Q_3 - Q_1$$

5.2.3 Box plot

- useful for visualization of data distribution
- first: create # line
  - mark min
  - Q_1
  - median = Q_2
  - Q_3
  - max

5.2.4 A visual summary

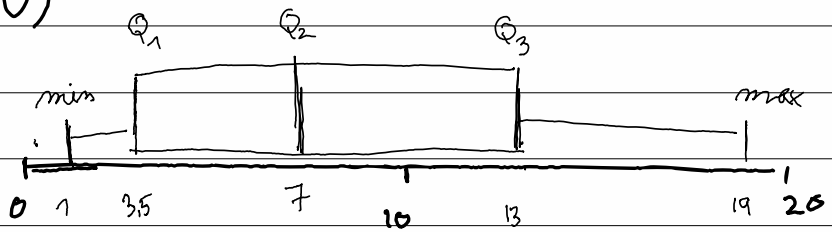
ex. 5.6 data = {3, 9, 4, 19, 11, 6, 7, 1, 15}

sorted = {1, 3, 4, 6, 7, 9, 11, 15, 19}

1) $n = 9$
2)
3) 7
4) min = 1
5) max = 19
6) range = max − min = 19 − 1 = 18
7) 1, 3, 4, 6 $\quad Q_1 = \frac{3+4}{2} = 3{,}5$
8) 9, 11, 15, 19 $\quad Q_3 = \frac{11+15}{2} = 13$

9) $IQR = Q_3 - Q_1 = 13 - 3{,}5 = \underline{\underline{9{,}5}}$

10)



5.7 excercise d)?

5.2.5 Outliers

- data point that lies outside the rest of the data's distribution

- lower fence $= Q_1 - 1.5 \cdot IQR$
- upper fence $= Q_3 + 1.5 \cdot IQR$

ex. 5.8 a) River data $\{6.3, 5.9, 7.0, 6.9, 5.9\}$

sorted $= \{5.9, 5.9, 6.3, 6.9, 7.0\}$

1) $n = 5$

2) min $= 5.9$, max $= 7.0$

3) $\text{mean} = \dfrac{5.9 + 5.9 + 6.3 + 6.9 + 7.0}{5} = 6.4$

4) $Q_2 = 6.3$

5) $Q_1 = 5.9$ $\quad Q_3 = 6.95$

6) $IQR = Q_3 - Q_1 = 6.95 - 5.9 = \underline{1.05}$

7) • lower fence $= Q_1 - 1.5 \cdot IQR$

$= 5.9 - 1.5 \cdot 1.05$

$= 5.9 - 1.575 \Rightarrow 4.325$

– they would either need to be $< \underline{4.325}$

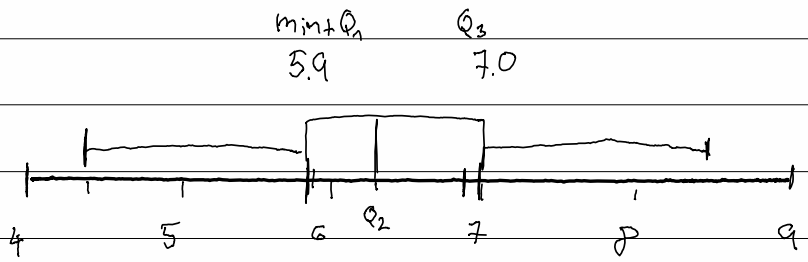
• upper fence $Q_3 + 1.5 \cdot IQR$

$6.95 + 1.5 \cdot 1.05$

$6.95 + 1.575 \Rightarrow 8.525$

– they would need to be $> \underline{8.525}$

8)



9) a) variance

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

n = size of sample

$\bar{x}$ = sample mean

data	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
5.9	-0.5	+0.25
5.9	-0.5	+0.25
6.3	-0.1	0.01
6.9	0.5	0.25
7.0	0.6	0.36

$$s^2 = \frac{1.12}{5-1} = \underline{\underline{0.25}}$$

b) standard deviation

$$S = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{1.12}{5-1}}$$

$S = 0{,}5$

c) coefficient of variation $= \frac{S}{\bar{y}} = \frac{0.5}{6.4} = 0{,}078125$

$= 7.9\%$

2) b) Virus resistant bacteria = {14, 15, 13, 21, 15, 14, 26, 16, 20, 13}

- sorted = {13, 13, 14, 14, 15, 15, 16, 20, 21, 26}

1) $n = 10$ median $= \frac{n+1}{2}$

2) min = 13 ; max = 26

3) mean $\bar{y} = \frac{13 + 13 + 14 + 14 + 15 + 15 + 16 + 20 + 21 + 26}{10}$

$\bar{y} = \frac{167}{10} = 16.7$

4) $Q_2 = 15$

5) $Q_1 = 14$ $Q_3 = 20$

6) $IQR = Q_3 - Q_1 = 20 - 14 = \underline{\underline{6}}$

7) upper fence $= Q_3 + 1.5 \cdot IQR$

$= 20 + 1.5 \cdot 6$

$= 29$

the observations would need to be > 29 to be an outliar

8)

9) a)

data	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
13	3.7	13.69
13	3.7	13.69
14	2.7	7.29
14	2.7	7.29
15	1.7	2.89
15	1.7	2.89
16	0.7	10.89
20	-3.3	
21	-4.3	18.49
26	-9.3	86.49

$$s^2 = \frac{163{,}61}{10-1} = \underline{\underline{18{,}17}}$$

b)

Standard dev = $\underline{\underline{S = 4.26}}$

c) $\frac{S}{\bar{y}} = \frac{4.26}{16.7} = 0.255 = 25.5\%$

5.3.1 Deviation from the mean

$y_i - \bar{y}$

5.3.2 Standard deviation

$$S = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}$$

5.3.3 Variance

$$S^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$

5.3.4 Coefficient variation $\frac{S}{\bar{y}}$