Chapter 3

Numerically Summarizing Data

In this chapter, we look at important *summaries* of data.

3.1 Measures of Central Tendency

Three measures of central tendency are average (or, equivalently, mean[1]), median and mode. These measures are either statistics for samples or parameters for populations.

measure	statistic (sample, n members)	parameter (population, N members)
average (mean)	$\bar{x} = \frac{x_1+x_2+\cdots+x_n}{n} = \frac{\sum x_i}{n}$	$\mu = \frac{x_1+x_2+\cdots+x_N}{N} = \frac{\sum x_i}{N}$
median	M: middle, $\frac{n+1}{2}$, of ordered sample	middle, $\frac{N+1}{2}$, of ordered population
mode	most frequent observation in sample	most frequent observation in population

Exercise 3.1 (Measures of Central Tendency)

1. *Average, Median, Mode: Molting Crayfish.* Consider small *population* of number of molts for $N = 9$ Crayfish in 3 month experiment:

   0, 0, 0, 0, 1, 1, 2, 2, 3.

   (a) *Population average*[2] is

   $$\mu = \frac{0+0+0+0+1+1+2+2+3}{9} = \frac{9}{9} =$$

   (choose one) **0** / **1** / **1.5** / **2**.

[1] Mean (or average) is sometimes also called *arithmetic mean* to distinguish it from *geometric mean* which is nth root of *product* of a list of numbers.

[2] Symbol "μ" for population average is pronounced "mew" and is a parameter.

(b) *Population median* is *middle* of 9 ordered molt numbers: $\frac{N+1}{2} = \frac{9+1}{2} = 5$th observation. Since first molt number is $x_1 = 0$, second is $x_2 = 0$, third is $x_3 = 0$, fourth is $x_4 = 0$, fifth molt number is **0 / 1 / 1.5 / 2**.

(c) *Population mode* is most frequent observation of 9 molt numbers: (choose one) **0 / 1 / 1.5 / 2**.

(d) If *sample* $n = 5$ molt numbers $\{0, 1, 2, 2, 3\}$,
sample average $\bar{x} = \frac{0+1+2+2+3}{5} = \frac{8}{5} =$ (choose one) **0 / 1 / 1.6 / 2**,
sample median M is middle, 3rd, of 5: (choose one) **0 / 1 / 1.6 / 2**,
sample mode is most frequent of 5: (choose one) **0 / 1 / 1.6 / 2**.

(e) If *sample* $n = 5$ molt numbers $\{0, 0, 1, 2, 3\}$,
sample average $\bar{x} = \frac{0+0+1+2+3}{5} = \frac{6}{5} =$ **0 / 1 / 1.2 / 1.4**,
sample median M is $\frac{n+1}{2} = \frac{5+1}{2} = 3$rd observation: **0 / 1 / 1.2 / 1.4**,
sample mode is most frequent of 5: **0 / 1 / 1.2 / 1.4**.

2. *Average, Median, Mode: Goals Scored.*
Consider small *population* of number of goals scored in $N = 9$ soccer games:

0, 1, 1, 2, 2, 2, 3, 3, 4.

(a) Population average is $\mu = \frac{0+1+1+2+2+2+3+3+4}{9} = \frac{18}{9} =$ **0 / 1 / 1.5 / 2**,
Population median is $\frac{N+1}{2} = \frac{9+1}{2} = 5$th observation: **0 / 1 / 1.5 / 2**,
Population mode is most frequent of 9: **0 / 1 / 1.5 / 2**.
StatCrunch: Stat, Summary Stats, Columns, select goals, select mean, median, mode, then Compute

(b) If *sample* $n = 5$ of $\{0, 1, 2, 3, 4\}$ goals scored,
sample average $\bar{x} = \frac{0+1+2+3+4}{5} = \frac{10}{5} =$ **1 / 1.6 / 1.8 / 2**,
sample median M is $\frac{n+1}{2} = \frac{5+1}{2} = 3$rd observation: **1 / 1.6 / 1.8 / 2**,
sample mode is most frequent of 5: **none /1 / 1.6 / 1.8 / 2**.
StatCrunch: Stat, Summary Stats, Columns, select sample 1, select mean, median, mode, Compute

(c) If *sample* $n = 6$ of $\{0, 0, 1, 2, 3, 4\}$ goals scored,
sample average $\bar{x} = \frac{0+0+1+2+3+4}{6} = \frac{10}{6} \approx$ **0 / 1.5 / 1.7 / 2**,
sample median M is $\frac{n+1}{2} = \frac{6+1}{2} = 3.5$rd observation; in other words,
average of 3rd and 4th observations $\frac{1+2}{2} =$ **0 / 1.5 / 1.7 / 2**,
sample mode is most frequent of 6: **0 / 1.5 / 1.7 / 2**.
StatCrunch: Stat, Summary Stats, Columns, select sample 2, select mean, median, mode, Compute

3.2 Measures of Dispersion

Three measures of dispersion are standard deviation, variance and range. These measures are either statistics for samples or for parameters for populations. Empirical Rule and Chebyshev's Inequality are also discussed.

measure	statistic (sample, n members)	parameter (population, N members)
standard deviation	$s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}}$ (definition) $= \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}{n-1}}$ (computational)	$\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{N}}$ (definition) $= \sqrt{\frac{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{N}}{N}}$ (computational)
variance	s^2	σ^2
range	$R =$ maximum $-$ minimum	maximum $-$ minimum

Exercise 3.2 (Measures of Dispersion)

1. *Range, Standard Deviation and Variance: Goals Scored.*
   Consider small *population* of number of goals scored in $N = 9$ soccer games:

   0, 1, 1, 2, 2, 2, 3, 3, 4.

   (a) *Measuring dispersion (spread) in entire set of goals scored.*
   range = Max $-$ Min $= 4 - 0 =$ (circle one) **1** / **2** / **4** goals scored
   standard deviation (SD) $\sigma = \sqrt{\frac{\sum(x_i-\mu)^2}{N}} \approx$ **1.05** / **1.15** / **1.23** goals
   variance $\sigma^2 \approx 1.15^2 \approx$ **1.11** / **1.26** / **1.33** goals^2
   StatCrunch: Stat, Summary Stats, Columns, select goals, select Unadj. Std Dev, then Compute

   (b) If *sample* $n = 5$ patients taking $\{0, 1, 2, 3, 4\}$ goals scored,
   $R =$ Max $-$ Min $= 4 - 0 =$ (circle one) **1** / **2** / **4** goals scored
   standard deviation (SD) $s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} \approx$ **1.05** / **1.41** / **1.58** goals scored
   variance $s^2 \approx 1.58^2 \approx$ **1** / **1.5** / **2.5** goals scored^2
   StatCrunch: Stat, Summary Stats, Columns, select sample 1 (from goals), select Std Dev, Compute

   (c) If *sample* $n = 6$ with $\{0, 0, 1, 2, 3, 4\}$ goals scored,
   $R =$ Max $-$ Min $= 4 - 0 =$ (circle one) **1** / **2** / **4** goals scored
   standard deviation (SD) $s = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n-1}} \approx$ **1.05** / **1.41** / **1.63** goals scored
   variance $s^2 \approx 1.63^2 \approx$ **1.56** / **1.87** / **2.67** goals scored^2
   StatCrunch: Stat, Summary Stats, Columns, select sample 2 (from goals), select Std Dev, Compute

2. *Empirical Rule: Goals Scored.*
   Empirical Rule states, *if* data is bell-shaped (mound-shaped),

   - approximately 68% of data falls within one SD of average,
   - approximately 95% of data falls within *two* SDs of average,
   - approximately 99.7% of data falls within *three* SDs of average.

   Consider small *population* of number of goals scored in $N = 9$ soccer games:

0, 1, 1, 2, 2, 2, 3, 3, 4.

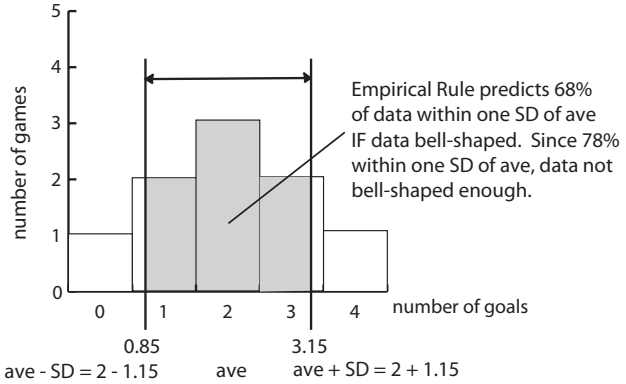

Figure 3.1 (Histogram for number of games versus of goals)

(a) Empirical Rule should apply to this data because histogram *appears* (choose one) **bell-shaped** / **skewed right** / **skewed left**.

(b) population average $\mu =$ (choose one) **1** / **2** / **3** goals
StatCrunch: Stat, Summary Stats, Columns, select goals, select mean, then Compute

(c) population SD: $\sigma \approx$ (choose one) **1.15** / **2.15** / **3.15** goals
StatCrunch: Stat, Summary Stats, Columns, select goals, select Unadj. Std Dev, then Compute

(d) $\mu - \sigma = 2 - 1.15 =$ (choose one) **0.85** / **1.05** / **2.05** goals

(e) $\mu + \sigma = 2 + 1.15 =$ (choose one) **0.85** / **1.05** / **3.15** goals

(f) $\mu \pm \sigma =$ (choose one) $\mathbf{(0.05, 2.05)}$ / $\mathbf{(0.85, 3.15)}$ / $\mathbf{(1.05, 4.25)}$

(g) so 0 is (choose one) **inside** / **outside** interval (0.85, 3.15)
and 1 is (choose one) **inside** / **outside** interval (0.85, 3.15)
and 2 is (choose one) **inside** / **outside** interval (0.85, 3.15)
and 3 is (choose one) **inside** / **outside** interval (0.85, 3.15)
so 7 goal scores, {1, 1, 2, 2, 2, 3, 3}, of all 9 goal scores,
or $\frac{7}{9} \approx 78\%$ are within one SD of average, are inside interval (0.85, 3.15).

(h) **True** / **False**. Empirical Rule predicts 68% should fall within one SD of average, inside interval (0.85, 3.15), and yet we see 78% of data inside interval (0.85, 3.15). This indicates data is not bell-shaped enough to apply this rule.

3. *More Empirical Rule: Heights.*
Heights of 14-year old students are assumed normally distributed with mean $\mu = 5'$ and standard deviation $\sigma = 0.5'$.

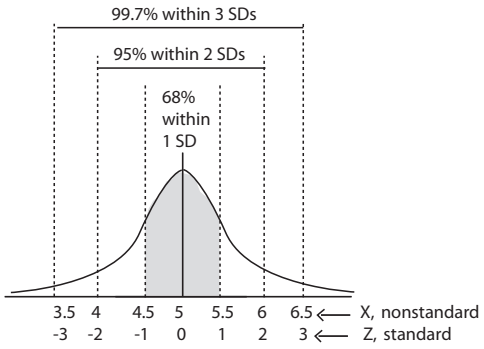

Figure 3.2 (Height of 14-year olds.)

Percentage of heights:

(a) between 4.5′ and 5.5′: **68%** / **95%** / **99.7%**

(b) between 4′ and 6′: **68%** / **95%** / **99.7%**

(c) between 3.5′ and 6.5′: **68%** / **95%** / **99.7%**

(d) between 5′ and 5.5′: $\frac{68}{2}$ = **34%** / **47.5%** / **68%**

(e) between 4′ and 5.5′: $\frac{95}{2} + \frac{68}{2}$ = **34%** / **47.5%** / **81.5%**

(f) less than 4′: $\frac{100-95}{2}$ = **2.5%** / **5%** / **7.5%**

4. *Chebyshev's Rule: pH levels.*
Chebyshev's Rule states, for *any* data set, at least $1 - \frac{1}{k^2}$ proportion of data falls within k standard deviations of average.

Consider sample of $n = 28$ Ph levels taken at Sand Dunes Park.

4.3	5	5.9	6.5	7.6	7.7	7.7	8.2	8.3	9.5
10.4	10.4	10.5	10.8	11.5	12	12	12.3	12.6	12.6
13	13.1	13.2	13.5	13.6	14.1	14.1	15.1		

(a) sample average $\bar{x}$ = (choose one) **10.55** / **11.55** / **12.55**
sample standard deviation s = (choose one) **2.45** / **3.01** / **4.55**
StatCrunch: Stat, Summary Stats, Columns, select pH, select Mean, Std Dev, then Compute

(b) pH level one SD above average is $\bar{x} + s = 10.55 + 3.01 = 13.56$.
pH level two SDs below average is $\bar{x} - 2s = 10.55 - 2(3.01) = 4.53$.
Fill in blanks in figure below.

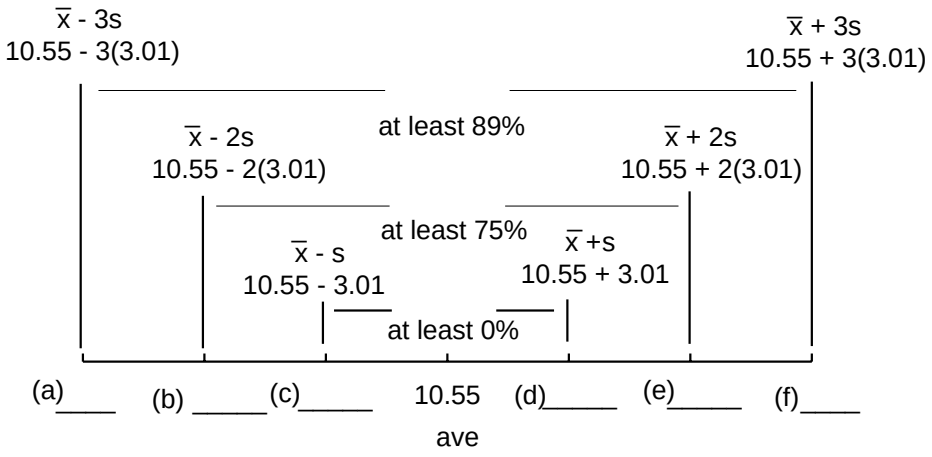

Figure 3.3 (pH levels 1, 2 and 3 SDs from average)

(c) Smallest pH level, 4.3, is **inside** / **outside** interval (7.54, 13.56).
pH level, 10.5, is **inside** / **outside** interval (7.54, 13.56).

(d) pH levels *within one SD of average*, refer to pH levels **inside** / **outside** interval (7.54, 13.56). pH levels within *two* SDs of average, refers to pH levels **inside** / **outside** interval (4.53, 16.57).

(e) Instead of saying "pH levels within *one* SD of mean", it is also possible to say "pH levels within k SDs of mean", where $k = 1$.
If pH levels are within *two* SDs of average, $k =$ **1** / **2** / **3**
If pH levels are within two and a half SDs of average, $k =$ **1** / **1.5** / **2.5**

(f) If $k = 1.5$, then $1 - \frac{1}{k^2} = 1 - \frac{1}{1.5^2} \approx 0.56$ or 56%.
If $k = 2$, then $1 - \frac{1}{k^2} =$ $\mathbf{\frac{1}{4}}$ / $\mathbf{\frac{2}{4}}$ / $\mathbf{\frac{3}{4}}$ or **25%** / **50%** / **75%**.

(g) Chebyshev's inequality says *at least* a $1 - \frac{1}{k^2} = 0.75$ proportion or 75% of 28 pH levels *should be* within *two* ($k = 2$) SDs of average.
In fact, 27 of 28 pH levels (look at data above and see for yourself), or $\frac{27}{28} = 0.964$ or 96.4%, are in interval (4.53, 16.57).
Chebyshev's inequality (circle one) **has** / **has not** be violated here.

(h) What proportion *should* fall within $k = 3$ SDs of average?
According to Chebyshev, at least $1 - \frac{1}{3^2} =$ (circle one) $\mathbf{\frac{3}{4}}$ / $\mathbf{\frac{6}{7}}$ / $\mathbf{\frac{8}{9}}$ or 89%.
In fact, actual proportion in (1.52,19.58) is $\mathbf{\frac{26}{28}}$ / $\mathbf{\frac{27}{28}}$ / $\mathbf{\frac{28}{28}}$ or 100%.
Chebyshev's inequality (circle one) **has** / **has not** be violated here.

(i) What proportion *should* fall within $k = 2.5$ SDs of average?
According to Chebyshev, at least $1 - \frac{1}{2.5^2} =$ $\mathbf{\frac{20}{25}}$ / $\mathbf{\frac{21}{25}}$ / $\mathbf{\frac{22}{25}}$ or 84%.
In fact, actual proportion in (3.03,18.08) is **90%** / **95%** / **100%**.
Chebyshev's inequality (circle one) **has** / **has not** be violated here.

(j) Since *at least* 75% or 21 pH levels are *inside* interval (4.534, 16.574), then *at most* (circle one) **25%** / **35%** / **45%** are *outside* interval (4.534, 16.574).

3.3 Measures of Central Tendency and Dispersion from Grouped Data

Often, data is not presented as a simple list of numbers. Data is "grouped", it is given in the form of a distribution table. We will look at how to calculate the average, median and standard deviation in this case. If frequency f_i is replaced weight w_i in formulas below, grouped data formulas become weighted formulas.

measure	statistic (sample, n members)	parameter (population, N members)
average (mean)	$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_n f_n}{f_1 + f_2 + \cdots + f_n} = \frac{\sum x_i f_i}{\sum f_i}$	$\mu = \frac{x_1 f_1 + x_2 f_2 + \cdots + x_N f_N}{f_1 + f_2 + \cdots + f_N} = \frac{\sum x_i f_i}{\sum f_i}$
standard deviation	$s = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{(\sum f_i) - 1}}$ (definition)	$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2 f_i}{\sum f_i}}$ (definition) $= \sqrt{\frac{\sum x_i^2 f_i - \frac{\left(\sum x_i f_i\right)^2}{\sum f_i}}{\sum f_i}}$ (computational)
variance	s^2	σ^2
median	M: middle, $\frac{n+1}{2}$, of ordered sample	middle, $\frac{N+1}{2}$, of ordered population

Exercise 3.3(Measures of Central Tendency and Dispersion Grouped Data)

1. *Grouped Data, Discrete Case: Average, Median and SD.*
   Consider discrete distribution table for sample of number of tablets used in high blood pressure experiment.

number of tablets	number of patients
1	5
2	10
3	4
4	1
total	20

   (a) *Exact Grouped Average.* Raw data from table is

$$1, 1, 1, 1, 1, \underbrace{2, 2, \ldots, 2}_{10}, 3, 3, 3, 3, 4,$$

   so exact grouped average is

$$\begin{aligned} \bar{x} &= \frac{1 + 1 + 1 + 1 + 1 + 2 + 2 + \cdots + 2 + 3 + 3 + 3 + 3 + 4}{20} \\ &= \frac{1(5) + 2(10) + 3(4) + 4(1)}{20} \end{aligned}$$

(circle one) **2.05** / **2.35** / **2.75**

StatCrunch: Stat, Summary Stats, Grouped/Binned data, Bins in: tablets, Counts in: patients, Statistics: Mean, Std. dev., Median, Compute.

(b) *Exact Grouped Standard Deviation.*

$$s = \sqrt{\frac{(1-2.05)^2(5) + (2-2.05)^2(10) + (3-2.05)^2(4) + (4-2.05)^2(1)}{20-1}} \approx$$

(circle one) **0.83** / **0.89** / **0.92**

Use Std. dev. from StatCrunch.

Median for table is *located* at $\frac{n+1}{2} = \frac{20+1}{2} = 10.5$th location; in other words, average of 10th and 11th observations, $M = \frac{2+2}{2} =$ (circle one) **2** / **3** / **4**

Use Median from StatCrunch.

(c) **True** / **False**. Grouped average, SD and median are exactly equal to raw data average, SD and median because same data used in both cases.

2. *Symmetry and Skewness For Average and Median: Number of Tablets.*
Number of tablets given to $n = 15$ patients varied in three samples.

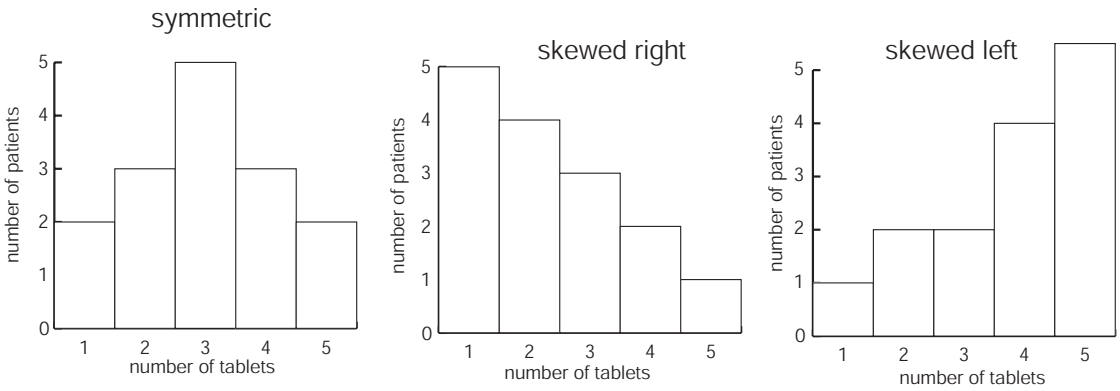

Figure 3.4 (Three histograms for number of tablets data)

symmetric		skewed right		skewed left	
number tablets	number patients	number tablets	number patients	number tablets	number patients
1	2	1	5	1	1
2	3	2	4	2	2
3	5	3	3	3	2
4	3	4	2	4	4
5	2	5	1	5	6
total	15	total	15	total	15

(a) For symmetric data, 2 patients given 1 tablet, 3 patients given 2 tablets and so on and so tablets given to 15 patients are (choose one)
(i) 1, 1, 1, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5.
(ii) 1, 2, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5.
(iii) 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5.

(b) For symmetric data,
average is $\bar{x} = \frac{2(1)+3(2)+5(3)+3(4)+2(5)}{15} = \frac{45}{15} = \mathbf{2}$ / $\mathbf{2.3}$ / $\mathbf{3}$ / $\mathbf{3.8}$,
median is $\frac{n+1}{2} = \frac{15+1}{2} = 8$th observation: $M =$ $\mathbf{2}$ / $\mathbf{2.5}$ / $\mathbf{3}$ / $\mathbf{3.5}$,
in other words, (choose one) $\boldsymbol{\bar{x} < M}$ / $\boldsymbol{\bar{x} = M}$ / $\boldsymbol{\bar{x} > M}$.
StatCrunch: Stat, Summary Stats, Grouped/Binned, tablets 1, patients 1, Mean, Median, Compute.

(c) For skewed right data,
average is $\bar{x} = \frac{5(1)+4(2)+3(3)+2(4)+1(5)}{15} = \frac{35}{15} \approx \mathbf{2}$ / $\mathbf{2.3}$ / $\mathbf{3}$ / $\mathbf{3.8}$,
median is $\frac{n+1}{2} = \frac{15+1}{2} = 8$th observation: $M =$ $\mathbf{2}$ / $\mathbf{2.5}$ / $\mathbf{3}$ / $\mathbf{3.5}$,
in other words, (choose one) $\boldsymbol{\bar{x} < M}$ / $\boldsymbol{\bar{x} = M}$ / $\boldsymbol{\bar{x} > M}$.
StatCrunch: Stat, Summary Stats, Grouped/Binned, tablets 2, patients 2, Mean, Median, Compute.

(d) For skewed left data,
average is $\bar{x} = \frac{1(1)+2(2)+3(2)+4(4)+5(6)}{15} = \frac{57}{15} = \mathbf{2}$ / $\mathbf{2.3}$ / $\mathbf{3}$ / $\mathbf{3.8}$,
median is $\frac{n+1}{2} = \frac{15+1}{2} = 8$th observation: $M =$ $\mathbf{2}$ / $\mathbf{2.5}$ / $\mathbf{3}$ / $\mathbf{4}$,
in other words, (choose one) $\boldsymbol{\bar{x} < M}$ / $\boldsymbol{\bar{x} = M}$ / $\boldsymbol{\bar{x} > M}$.
StatCrunch: Stat, Summary Stats, Grouped/Binned, tablets 3, patients 3, Mean, Median, Compute.

3. *Grouped Data, Continuous Case: Average, SD and Median.*
Consider continuous distribution table for sample of patient ages used in high blood pressure experiment.

class interval	age midpoints	number
30-34	$\frac{30+35}{2} = 32.5$	1
35-39	37.5	2
40-44	42.5	8
45-49	47.5	7
50-54	52.5	2
total		20

(a) *Approximate Grouped Average.*
Since actual values of ages *not* given in continuous distribution table, each age *approximated* by *midpoint* of each class,

$$x_1 = 32.5, x_2 = 37.5, x_3 = 37.5, \underbrace{x_4 = 42.5, x_5 = 42.5, \ldots, x_{11} = 42.5}_{8}$$

$$\underbrace{x_{12} = 47.5, x_{13} = 47.5, \ldots, x_{18} = 47.5}_{7}, x_{19} = 52.5, x_{20} = 52.5,$$

so approximate average is

$$\bar{x} = \frac{32.5(1) + 37.5(2) + 42.5(8) + 47.5(7) + 52.5(2)}{20} =$$

(circle one) **42.5** / **44.25** / **46.0**.

StatCrunch: Stat, Summary Stats, Grouped/Binned, age, number, Mean, Std. Dev, Median, Compute.

(b) *Approximate Standard Deviation.*

$$s = \sqrt{\frac{(32.5 - 44.25)^2(1) + \cdots + (52.5 - 44.25)^2(2)}{20 - 1}} \approx$$

(circle one) **2.83** / **3.89** / **4.94**

Use Std. dev. from StatCrunch.

(c) *Approximate median.*
Median for table is *located* at $\frac{n+1}{2} = \frac{20+1}{2} = 10.5$th location; in other words, average of 10th and 11th ages, $M = \frac{42.5+42.5}{2} =$ **42.5** / **44.25** / **46.0**.

Use Median from StatCrunch.

(d) **True** / **False**. *Exact* mean, SD and median are calculated for grouped *discrete* data, but *approximate* mean, SD and median are calculated for grouped *continuous* data.

4. *Weighted Average: Eye blinks*
Consider table of number of eye blinks per minute in a psychological study:

eye blinks	chance of eye blinks (weights)
0	10%
1	40%
2	10%
3	10%
4	30%
total	100%

StatCrunch: Stat, Summary Stats, Grouped/Binned, eye blinks, weights, Mean, Std. Dev, Median, Compute.
Weighted average is $\bar{x} = \frac{0(10)+1(40)+2(10)+3(10)+4(30)}{100} =$ **1.5** / **2.1** / **2.2**.
Weights are more general than frequencies: frequencies are a count of items, weights may or may not be a count of items.

3.4 Measures of Position and Outliers

We look at the measure of position, the z-score:

$$z = \frac{x - \bar{x}}{s}, \text{ (sample)} \quad z = \frac{x - \mu}{\sigma} \text{ (population)}$$

Exercise 3.4 (Measures of Position and Outliers)

1. *z-scores: IQ scores.* IQ scores differ for different ages. Mean, SD for 16 year olds are $\mu = 100$ and $\sigma = 16$; mean, SD for 20 year olds are $\mu = 120$ and $\sigma = 20$.

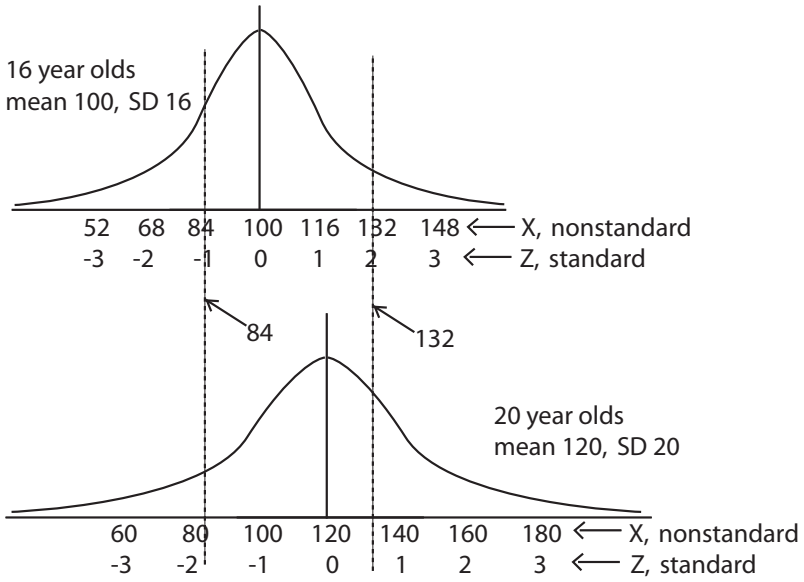

Figure 3.5 (Comparing IQ Scores with Z-Scores)

   (a) A 16 year old with IQ 132 has z-score $z = \frac{132-100}{16} =$ **0** / **1** / **2**.
   This IQ is two SDs above average.

   (b) A 16 year old with IQ 84 has z-score $z = \frac{84-100}{16} =$ **−2** / **−1** / **0**.
   This IQ is one SD *below* average.

   (c) A 20 year old with IQ 132 has z-score $z = \frac{132-120}{20} =$ **0** / **0.6** / **2**.
   This IQ is 0.6 of a SD above average.

   (d) A 20 year old with IQ 84 has z-score $z = \frac{84-120}{20} =$ **−2** / **−1.8** / **−1**.
   This IQ is 1.8 SDs below average.

   (e) **True** / **False**. z-scores allow comparison of *position* of data points in different data sets, data sets with different averages and SDs.

   (f) If $z = \frac{x-\mu}{\sigma}$, then $x = z\sigma + \mu$, so a 16 year old with IQ three SDs above average has IQ $x = 3(16) + 100 =$ (choose one) **116** / **132** / **148**.

   (g) A 20 year old with IQ two SDs below average has IQ
   $x = -2(20) + 120 =$ (choose one) **60** / **80** / **100**.

2. *Using z-scores to find outliers: Temperatures.*
   Consider small *sample* of $n = 10$ temperatures, set A:

$$0,\ 1,\ 1,\ 2,\ 2,\ 2,\ 3,\ 3,\ 5,\ 7.$$

(a) average temperature, $\bar{x}$ = (choose one) **0** / **1.6** / **2.6** degrees.
SD in temperature, $s \approx$ (choose one) **1.15** / **1.23** / **2.07** degrees.
(StatCrunch: Stat, Summary Stats, temp A, Mean, Std. Dev., Compute.)

(b) Temperature 0^o has z-score $z \approx \frac{0-2.6}{2.07} \approx$ **−1.98** / **−1.27** / **−0.56**.
This temperature is roughly 1.3 SDs below average.

(c) Temperature 7^o has z-score $z \approx \frac{7-2.6}{2.07} \approx$ **1.68** / **1.97** / **2.13**.
This temperature is roughly 2.1 SDs above average.

(d) z-scores less than $z = -2$ or greater than $z = 2$ are considered *outliers*.
So, 7^o **is** / **is not** an outlier because it is more than 2 SDs above average.

(e) Temperature 1.5 SDs above average is
$x \approx 1.5(2.07) + 2.6 =$ (choose one) **3.335** / **3.745** / **5.705**.

3.5 The Five-Number Summary and Boxplots

We look at five-number summary

$$\{\min, \quad P_{25} = Q_1, \quad M = P_{50} = Q_2, \quad P_{75} = Q_3, \quad \max\},$$

and related boxplots.

Exercise 3.5 (Percentiles, the Five-Number Summary and Boxplots)

1. *Percentiles and Quartiles: Temperatures.*
Consider small *sample* of $n = 10$ temperatures, set A:

$$0, 1, 1, 2, 2, 2, 3, 3, 5, 7.$$

(a) 50th percentile is temperature where 50% of data is *below* this temperature.
If 50% is below, then 50% must be above this temperature. 50th percentile is **median** / **average**.

(b) *Locate, then identify percentile.*
Position of median is $\frac{n+1}{2} = \frac{10+1}{2} =$ (choose one) **5** / **5.5** / **6**;
in other words, median is average of 5th, 6th ordered temperatures,
$M = P_{50} = \frac{2+2}{2} =$ **0** / **1.5** / **2**.

(c) 25th percentile is temperature with 25% of temperatures below this, 75% above, so 25th percentile is **above** / **equal to** / **below** median.
Position of 25th is median of data below median, $\{0, 1, 1, 2, 2\}$:
$\frac{n+1}{2} = \frac{5+1}{2} =$ **2** / **3** / **4**;
in other words, 25th is 3rd ordered temperature $P_{25} =$ **0** / **1** / **2.5**.

(d) 25th percentile has special name, the lower (or first) *quartile*, $P_{25} = Q_1$.
50th percentile is **second** / **upper** / **third** quartile, $M = P_{50} = Q_2$.
75th percentile is **first** / **second** / **third** quartile, $P_{75} = Q_3$.

(e) Position of 75th is median of data above median, $\{2, 3, 3, 5, 7\}$:
$\frac{n+1}{2} = \frac{5+1}{2} =$ **2 / 3 / 4**;
in other words, 75th is 3rd ordered temperature $P_{75} =$ **2 / 3 / 4**.

2. *Quartiles, Fences and IQR to find Outliers: Temperature.*
Consider small *sample* of $n = 10$ temperatures, set A:

0, 1, 1, 2, 2, 2, 3, 3, 5, 7.

(a) Since $Q_1 = 1$ and $Q_3 = 3$,
interquartile range, $IQR = Q_3 - Q_1 = 3 - 1 =$ **1 / 2 / 3**,
is, like SD, a measure of spread.

(b) *Lower fence* $= Q_1 - 1.5IQR = 1 - 1.5(2) =$ **−3 / −2 / −1**,
is an unusually small value.

(c) *Upper fence* $= Q_3 + 1.5IQR = 3 + 1.5(2) =$ **5 / 6 / 7**,
is an unusually large value.

(d) Data points less lower fence or greater than upper fence considered *outliers*.
So, 7^o **is / is not** an outlier since it is larger than upper fence, 6^o.

(e) If 7 is mistyped as 70,
IQR changes **not at all / a lot**, but SD changes **a little / a lot**.
So IQR, is resistant to outliers whereas SD is sensitive to outliers.

3. *Five Number Summary and Boxplot: Temperatures.*
Consider small *sample* of $n = 10$ temperatures, set A:

0, 1, 1, 2, 2, 2, 3, 3, 5, 7.

(a) Five–number summary for temperatures, is (choose one)
(i) {0, 1, 1.5, 3, 4}
(ii) {0, 0, 1.5, 3, 6}
(iii) {0, 1, 2, 3, 7}
(StatCrunch: Choose Stat, Summary Stats, temp A, Min, Q1, Median, Q3, Max.)

(b) Consider boxplot for temperatures.

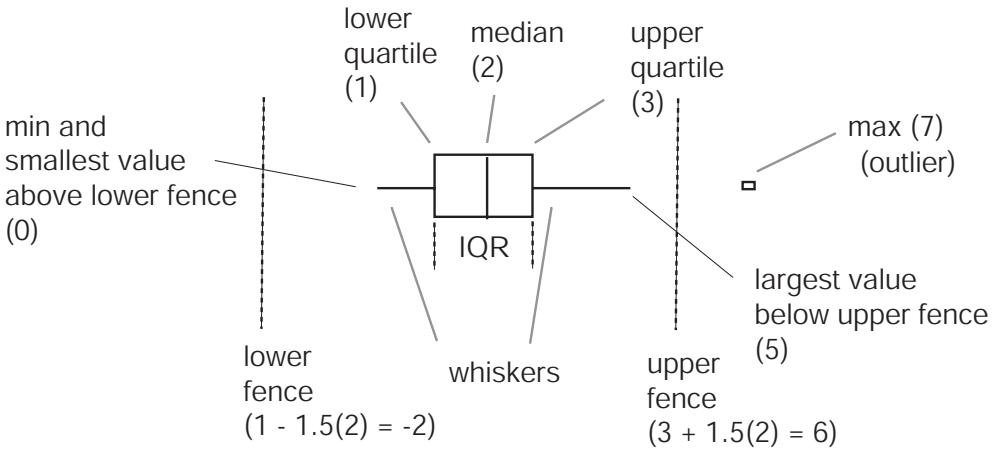

Figure 3.6 (Boxplot for temperatures)

Boxplot indicates data **symmetric** / **skewed right** / **skewed left**.

StatCrunch: Choose Graphics, Boxplot, temp A check Use fences to identify outliers, check Draw boxes horizontally, Compute

4. *Five Number Summary and Boxplot: More Temperatures.*
Another *sample* of $n = 9$ temperatures, set B

$$0,\ 0,\ 0,\ 0,\ 1,\ 1,\ 2,\ 2,\ 3$$

is compared to the first set of temperatures.

(a) Five–number summary for this set of temperatures, is (choose one)
(i) {0, 1, 1.5, 3, 4}
(ii) {0, 0, 1, 2, 3}
(iii) {0, 1, 2, 3, 7}

StatCrunch: Stat, Summary Stats, temp B, temp A, Min, Q1, Median, Q3, Max.

(b) Consider side-by-side boxplot for temperatures.

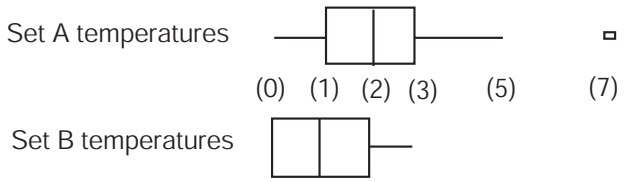

Figure 3.7 (Side-by-side boxplot for two sets of temperatures)

Set A has **warmer** / **same** / **colder** median temperature than set B.
Set A has **smaller** / **same** / **larger** IQR in temperature than set B.