Chapter 12

Describing Distributions with Numbers

We look at important numerical summaries of distributions. Two measures of central tendency are average (or, equivalently, mean) and median. *Mean*, $\bar{x}$, is sum of numbers divided by n. *Median*, M, is middle number in list of numbers arranged smallest to largest and is *located* with the $\frac{n+1}{2}$ rule. Measures of variability are standard deviation (as well as previously discussed and closely related variance) and first and third quartiles. *Standard deviation*, s, is "average distance from mean". *First and third quartiles*, Q_1 and Q_3, are medians of upper half and lower half, respectively, of list of numbers arranged smallest to largest. We look at five-number summary

$$\{\min, \quad Q_1, \quad M, \quad Q_3, \quad \max\},$$

and related boxplot. We also calculate mean, median and SD for grouped data.

Exercise 12.1 (Describing Distributions with Numbers)

1. *Mean and median: temperatures.* Consider small set of temperatures at $n = 9$ different locations in Westville during a day in January:

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

   (a) *Mean* is
   $$\bar{x} = \frac{0+0+0+0+1+1+2+2+3}{9} = \frac{9}{9} =$$
   (choose one) **0** / **1** / **1.5** / **2**.

   (b) *Median* is *middle* of 9 ordered temperatures: $\frac{n+1}{2} = \frac{9+1}{2} = 5$th observation. Since first temperature is 0, second is 0, third is 0, fourth is 0, fifth temperature is $M =$ **0** / **1** / **1.5** / **2**.

(c) If sample $n = 5$ temperatures $\{0, 1, 2, 2, 3\}$,
mean $\bar{x} = \frac{0+1+2+2+3}{5} = \frac{8}{5} =$ (choose one) **0 / 1 / 1.6 / 2**,
median M is middle, $\frac{n+1}{2} = \frac{5+1}{2} = 3$rd, of 5: (choose one) **0 / 1 / 1.6 / 2**,

(d) If sample $n = 5$ temperatures $\{0, 0, 1, 2, 3\}$,
mean $\bar{x} = \frac{0+0+1+2+3}{5} = \frac{6}{5} =$ **0 / 1 / 1.2 / 1.4**,
median M is $\frac{n+1}{2} = \frac{5+1}{2} = 3$rd observation: **0 / 1 / 1.2 / 1.4**,

2. *Mean and median: tablets.*
Consider number of tablets given to $n = 9$ patients:

$$0,\ 1,\ 1,\ 2,\ 2,\ 2,\ 3,\ 3,\ 4.$$

(a) Mean is $\bar{x} = \frac{0+1+1+2+2+2+3+3+4}{9} = \frac{18}{9} =$ **0 / 1 / 1.5 / 2**,
Median is $\frac{n+1}{2} = \frac{9+1}{2} = 5$rd observation: **0 / 1 / 1.5 / 2**,

(b) If sample $n = 5$ patients taking $\{0, 1, 2, 3, 4\}$ tablets,
mean $\bar{x} = \frac{0+1+2+3+4}{5} = \frac{10}{5} =$ **1 / 1.6 / 1.8 / 2**,
median M is $\frac{n+1}{2} = \frac{5+1}{2} = 3$rd observation: **1 / 1.6 / 1.8 / 2**,

(c) If sample $n = 6$ patients taking $\{0, 0, 1, 2, 3, 4\}$ tablets,
mean $\bar{x} = \frac{0+0+1+2+3+4}{6} = \frac{10}{6} \approx$ **0 / 1.5 / 1.7 / 2**,
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**,

(d) If 2 added to $n = 6$ tablets $\{0, 0, 1, 2, 3, 4\} + 2 = \{2, 2, 3, 4, 5, 6\}$,
mean $\bar{x} = \frac{2+2+3+4+5+6}{6} = \frac{22}{6} =$ **0 / 1 / 1.7 / 3.7**,
median M is $\frac{n+1}{2} = \frac{6+1}{2} = 3.5$rd observation; in other words,
average of 3rd and 4th observations $\frac{3+4}{2} =$ **0 / 1.5 / 3.5 / 4**,
where both mean and median increase by 2 tablets.

3. *Average sensitive to outliers; median resistant (robust) to outliers.*
Set of temperatures at $n = 9$ different locations in Westville,

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

is mistakenly recorded as

$$0, 0, 0, 0, 1, 1, 2, 2, 30.$$

(a) **True** / **False**. Typing mistake, "30", is an outlier.

(b) With*out* outlier,
mean is $\bar{x} = \frac{0+0+0+0+1+1+2+2+3}{9} = \frac{9}{9} =$ **0 / 1 / 1.5 / 2**,
median is $\frac{n+1}{2} = \frac{9+1}{2} = 5$rd observation: $M =$ **0 / 1 / 1.5 / 2**.

(c) *With* outlier,
mean is $\bar{x} = \frac{0+0+0+0+1+1+2+2+30}{9} = \frac{36}{9} =$ **0 / 1 / 1.5 / 4**,
median is $\frac{n+1}{2} = \frac{9+1}{2} = 5$rd observation: $M =$ **0 / 1 / 1.5 / 4**.

(d) Mean with*out* outlier, 1, is
(circle one) **much larger than / equal to / much smaller than**
mean *with* outlier, 4, so mean is
(circle one) **sensitive / resistant (robust)** to outlier.

(e) Median with*out* outlier, 1, is
(circle one) **much larger than / equal to / much smaller than**
median *with* outlier, 1, so median is
(circle one) **sensitive / resistant (robust)** to outlier.

(f) *Mean* with outlier, 4, is
(circle one) **much larger than / equal to / much smaller than**
median with outlier, 1, so mean is "pulled" ("drawn")
from median towards outlier.
from outlier towards median.

(g) Most appropriate measure of center of data with*out* outliers is
mean / median; whereas most appropriate measure of center of data *with* outliers is **mean / median**.

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

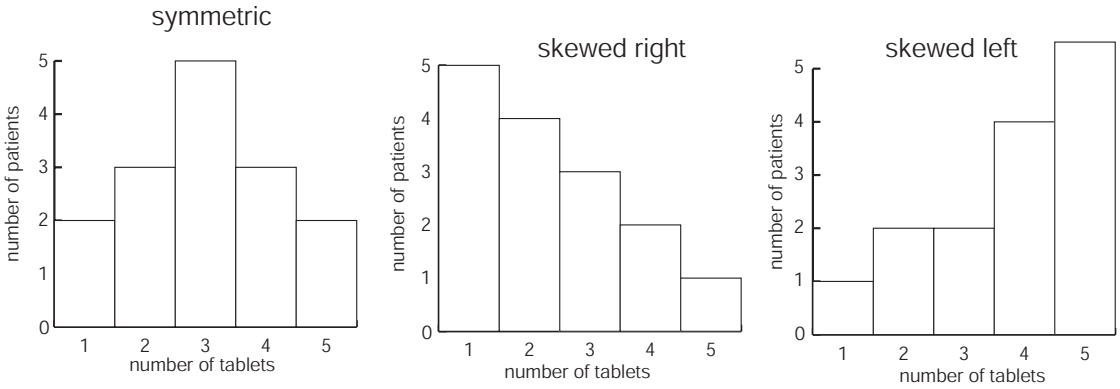

Figure 12.1 (Three histograms for number of tablets data)

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,
mean is $\bar{x} = \frac{2(1)+3(2)+5(3)+3(4)+2(5)}{15} = \frac{45}{15} =$ **2 / 2.3 / 3 / 3.8**,
median is $\frac{n+1}{2} = \frac{15+1}{2} = 8$th observation: $M =$ **2 / 2.5 / 3 / 3.5**,
in other words, (choose one) $\bar{x} < M$ / $\bar{x} = M$ / $\bar{x} > M$,
so mean and median are roughly same for roughly symmetric data.

(c) For skewed right data,
mean is $\bar{x} = \frac{5(1)+4(2)+3(3)+2(4)+1(5)}{15} = \frac{35}{15} \approx$ **2 / 2.3 / 3 / 3.8**,
median is $\frac{n+1}{2} = \frac{15+1}{2} = 8$th observation: $M =$ **2 / 2.5 / 3 / 3.5**,
in other words, (choose one) $\bar{x} < M$ / $\bar{x} = M$ / $\bar{x} > M$,
so mean greater than median for right skewed data.

(d) For skewed left data,
mean is $\bar{x} = \frac{1(1)+2(2)+3(2)+4(4)+5(6)}{15} = \frac{57}{15} =$ **2 / 2.3 / 3 / 3.8**,
median is $\frac{n+1}{2} = \frac{15+1}{2} = 8$th observation: $M =$ **2 / 2.5 / 3 / 4**,
in other words, (choose one) $\bar{x} < M$ / $\bar{x} = M$ / $\bar{x} > M$,
so mean less than median for left skewed data.

(e) Most appropriate measure symmetric data is **mean** / **median**; whereas most appropriate measure of skewed data is **mean** / **median**.

5. *Standard deviation and variance: tire weights.* An entire shipment of 18 tire weights (pounds) are broken into three lots (samples) of six tires each:

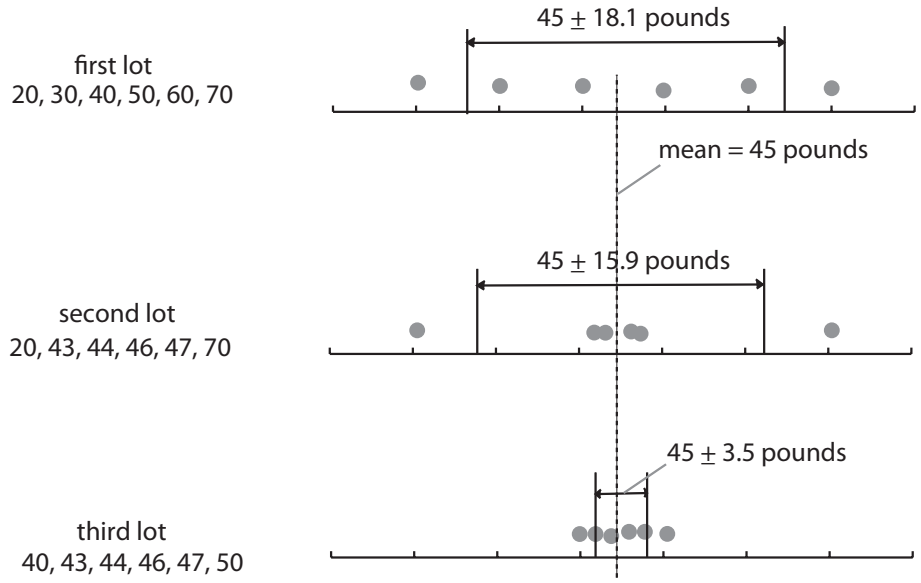

Figure 12.2 (Tire weights: three lots of tires)

(a) *How "spread out" ("dispersed") are tire weights in each lot?*
Lot 1 "spread" **more than** / **less than** / **equal to** lot 2 "spread".
Lot 1 "spread" **more than** / **less than** / **equal to** lot 3 "spread".
Lot 2 "spread" **more than** / **less than** / **equal to** lot 3 "spread".

(b) *Measuring spread in each lot (sample): standard deviation (SD).*
Lot 1, since mean $\bar{x} = \frac{20+30+40+50+60+70}{6} = 45$,

observation	squared distance from mean
20	$(20-45)^2 = (-25)^2 = 625$
30	$(30-45)^2 = (-15)^2 = 225$
40	$(40-45)^2 = (-5)^2 = 25$
50	$(50-45)^2 = (5)^2 = 25$
60	$(60-45)^2 = (15)^2 = 225$
70	$(70-45)^2 = (25)^2 = 625$
	sum = 1750

standard deviation (SD) $s = \sqrt{\frac{1750}{6-1}} \approx$ **17.2** / **18.7** / **19.2** pounds
Lot 2, $s = \sqrt{\frac{1260}{6-1}} \approx$ (circle one) **13.2** / **14.5** / **15.9** pounds
Lot 3, $s = \sqrt{\frac{60}{6-1}} \approx$ (circle one) **3.5** / **4.5** / **5.6** pounds
SD clearly indicates lot 1 ($s \approx 18.7$) is more spread out than lot 2 ($s \approx 15.9$) which, in turn, is more spread out than lot 3 ($s \approx 3.5$).
Use standard deviation button on calculator.

(c) Standard deviation measures average variability around **mean** / **median**.
So $\bar{x}$ and s used for symmetric distributions free of outliers.

(d) *Measuring spread (dispersion) in each lot: variance.*
Lot 1, $s^2 \approx 18.7^2 =$ (circle one) **12.25** / **252.81** / **349.69** pounds^2
Lot 2, $s^2 \approx 15.9^2 =$ (circle one) **12.25** / **252.81** / **349.69** pounds^2
Lot 3, $s^2 \approx 3.5^2 =$ (circle one) **12.25** / **252.81** / **349.69** pounds^2
Variance not used as often as SD because units are "squared"; in this case, "pounds^2" are given for variance but "pounds" for SD, s.

6. *Standard deviation and variance: tablets.*
Consider number of tablets given to $n = 9$ patients:

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

(a) *Measuring variability (spread) in entire set of tablets.*
since mean $\bar{x} = \frac{0+1+1+2+2+2+3+3+4}{9} = 2$,

observation	squared distance from mean
0	$(0-2)^2 = (-2)^2 = 4$
1	$(1-2)^2 = (-1)^2 = 1$
1	$(1-2)^2 = (-1)^2 = 1$
2	$(2-2)^2 = (0)^2 = 0$
2	$(2-2)^2 = (0)^2 = 0$
2	$(2-2)^2 = (0)^2 = 0$
3	$(3-2)^2 = (1)^2 = 1$
3	$(3-2)^2 = (1)^2 = 1$
4	$(4-2)^2 = (2)^2 = 4$
	sum = 12

standard deviation (SD) $s = \sqrt{\frac{12}{9-1}} \approx$ **1.05** / **1.15** / **1.22** tablets
variance $1.22^2 \approx$ **1.11** / **1.26** / **1.49** tablets^2

Use standard deviation button on calculator: On TI-84+, for example,

type 9 temperatures into L_1, then STAT CALC 1–Var Stats L_1 to find SD is $s_X \approx 1.22$.

(b) If sample $n = 5$ patients take $\{0, 1, 2, 3, 4\}$ tablets,
since mean $\bar{x} =$ **1.0** / **1.5** / **2.0** tablets

observation	squared distance from mean
_____	_____ = _____
_____	_____ = _____
_____	_____ = _____
_____	_____ = _____
_____	_____ = _____
	sum = _____

standard deviation (SD) $s \approx$ **1.05** / **1.41** / **1.58** tablets
variance $1.58^2 \approx$ **1** / **1.5** / **2.5** tablets^2

Type 5 numbers into L_2. Then, STAT CALC 1–Var Stats L_2 to find sample SD is $s_X \approx 1.58$.

(c) If 2 added to $n = 5$ patients $\{0, 1, 2, 3, 4\} + 2 = \{2, 3, 4, 5, 6\}$,
since mean $\bar{x} =$ **2** / **3** / **4** tablets

observation	squared distance from mean
_____	_____ = _____
_____	_____ = _____
_____	_____ = _____
_____	_____ = _____
_____	_____ = _____
	sum = _____

standard deviation (SD) $s \approx$ **1.05** / **1.41** / **1.58** tablets
variance $1.58^2 \approx$ **1** / **1.5** / **2.5** tablets^2
where SD (and variance) unchanged because *variability* unchanged

(d) If $n = 5$ patients take $\{1, 1, 1, 1, 1\}$ tablets,
standard deviation (SD) $s \approx$ **0** / **1.41** / **1.58** tablets
variance $0^2 \approx$ **0** / **1.5** / **2.5** tablets^2
because no variability if all patients take same number of tablets

7. *Grouped data: mean, SD, and median.*
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 mean.* 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**

(b) *Exact grouped standard deviation*

observation	squared distance from mean
_____	_____ = _____
_____	_____ = _____
_____	_____ = _____
_____	_____ = _____
	sum = _____

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

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

For TI-84+, type number of tablets into L_1 and number of patients into L_2 and then type STAT CALC 1–Var Stats L_1, L_2.

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

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

8. *More grouped data: mean, SD and median.*
Consider distribution table for sample of patient ages used in high blood pressure experiment.

class interval	midpoint	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 mean.*
Since actual values of ages *not* given in 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 mean 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**.

In TI-84+, type midpoints into L_1 and number of patients into L_2 and then type STAT CALC 1–Var Stats L_1, L_2.

(b) *Approximate standard deviation.*

observation	squared distance from mean
_____	_____ = _____
_____	_____ = _____
_____	_____ = _____
_____	_____ = _____
_____	_____ = _____
	sum = _____

$$s = \sqrt{\frac{(32.5-44.25)^2(1)+(37.5-44.25)^2(2)+(42.5-44.25)^2(8)+(47.5-44.25)^2(7)+(52.5-44.25)^2(2)}{20-1}} \approx$$

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

(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**.

9. *Median, quartiles, 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) Since

$$0, 1, 1, 2, \underbrace{2, 2}_{M}, 3, 3, 5, 7,$$

median temperature located $\frac{10+1}{2} = 5.5$th position and so $M =$ **1** / **2** / **3**

(b) Since *lower half* of ordered data set

$$0, 1, \underbrace{1}_{Q_1}, 2, 2,$$

first quartile located $\frac{5+1}{2} = 3$rd position, $Q_1 =$ **1** / **2** / **3**

(c) Since *upper half* of ordered data set

$$2, 3, \underbrace{3}_{Q_3}, 5, 7,$$

third quartile located $\frac{5+1}{2} = 3$rd position, $Q_3 =$ **3** / **5** / **7**

(d) *Five–number summary* for temperatures, {min, Q_1, M, Q_3, max} =
(i) {0, 1, 1.5, 3, 4}
(ii) {0, 0, 1.5, 3, 6}
(iii) {0, 1, 2, 3, 7}

(e) *Boxplot* for temperatures.

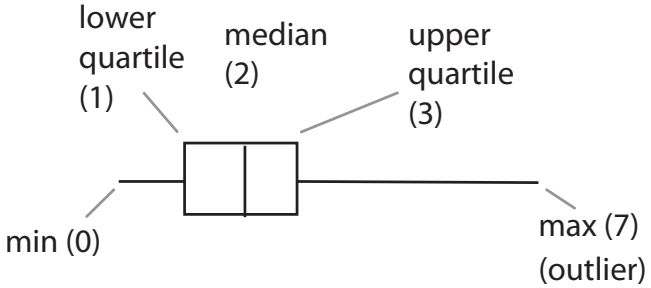

Figure 12.3 (Boxplot for temperatures)

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

10. *Median, quartiles, five number summary and boxplot: more temperatures.* Another *sample* of $n = 9$ temperatures, set B

$$0, \underbrace{0, 0}_{Q_1}, 0, \underbrace{1}_{M}, 1, \underbrace{2, 2}_{Q_3}, 3$$

is compared to the first set of temperatures.

(a) Five–number summary, {min, Q_1, M, Q_3, max} =
(i) {0, 1, 1.5, 3, 4}
(ii) {0, 0, 1, 2, 3}
(iii) {0, 1, 2, 3, 7}

(b) Consider side-by-side boxplots for temperature sets A and B.

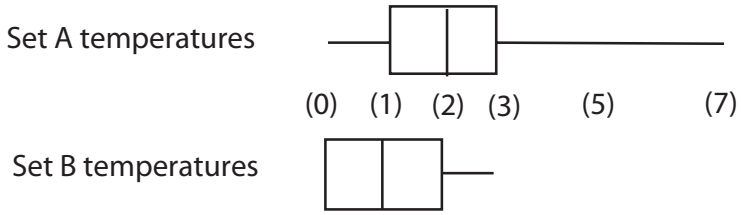

Figure 12.4 (Side-by-side boxplot for two sets of temperatures)

Set A has **warmer** / **same** / **colder** median temperature than set B.
Set A **less** / **same** / **more** right skewed in temperatures than set B.

(c) First and third quartiles measures variability around **mean** / **median**. So median and quartiles used for skewed distributions, or distributions with outliers.