Sections 2.3 and 2.4

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 205: Elementary Statistics for the Biological and Life Sciences

Descriptive statistics

- For continuous data, a histogram (or dotplot) provides a "snapshot" of the data.
- This snapshot can be augmented with a few numbers to give a brief quantitative description of the data.
- These numbers (mean, median, mode, standard deviation, interquartile range, etc.) are called **sample statistics.**

Sample median

- The median is a number that splits the data into two groups.
- Half the observations are *smaller* than the median, and half are *larger*.
- Need to order the data first, then find “middle” observation.
- This is unique if n is odd. Take average of middle *two* if n even.

Example 2.3.1: weight gain in lambs

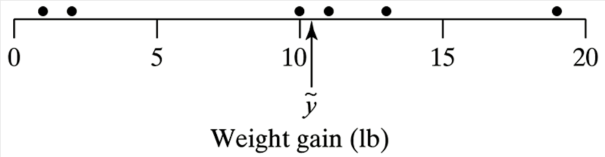
- $n = 6$ lambs weight gain (lbs) recorded over two weeks. The ordered values are:

$$1, 2, 10, 11, 13, 19.$$

- The sample median is

$$\tilde{y} = \frac{10 + 11}{2} = 10.5 \text{ lbs.}$$

- 3 obs. larger than median & 3 smaller:



Sample mean

- The **sample mean** is

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

  where the y_i's are the observations in the sample and n is the sample size.
- The sample mean is the average of the n data values.
- Has interpretation as "point of balance."
- If every observation has the same weight, then $\bar{y}$ is fulcrum of balance.

Example 2.3.1: weight gain in lambs

- Sample mean is

$$\bar{y} = \frac{1+2+10+11+13+19}{6} = \frac{56}{6} = 9.33 \text{ lbs.}$$

- Median causes see-saw to tip

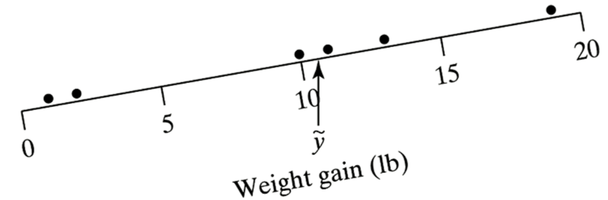


Figure 2.3.2 Plot of the lamb weight-gain data with the sample median as the fulcrum of a balance

Example 2.3.1: weight gain in lambs

- Sample mean is

$$\bar{y} = \frac{1 + 2 + 10 + 11 + 13 + 19}{6} = \frac{56}{6} = 9.33 \text{ lbs.}$$

- Mean balances see-saw

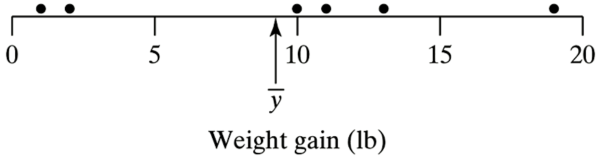


Figure 2.3.3 Plot of the lamb weight-gain data with the sample mean as the fulcrum of a balance

Mean versus median

- Median is robust to outliers, mean is not.
- What happens with lamb weight gain when we replace the largest value 19 by 100?
- Original data: $\tilde{y} = 10.5$ and $\bar{y} = 9.33$ lbs. New data:

$$1, 2, 10, 11, 13, 100,$$

  $\tilde{y} = 10.5$ and $\bar{y} = 22.83$ lbs.
- Mean is also pulled in direction of skew further than median.

Example 2.3.1: Cricket singing times

Male Mormon crickets sing to attract mates. The song duration from $n = 51$ crickets was measured in minutes.

Table 2.3.1 Fifty-one cricket singing times (min)

4.3	3.9	17.4	2.3	0.8	1.5	0.7	3.7
24.1	9.4	5.6	3.7	5.2	3.9	4.2	3.5
6.6	6.2	2.0	0.8	2.0	3.7	4.7	
7.3	1.6	3.8	0.5	0.7	4.5	2.2	
4.0	6.5	1.2	4.5	1.7	1.8	1.4	
2.6	0.2	0.7	11.5	5.0	1.2	14.1	
4.0	2.7	1.6	3.5	2.8	0.7	8.6	

R code: cricket music

R code to get mean and median:

```
times=c(4.3,3.9,17.4,2.3,0.8,1.5,0.7,3.7,24.1,9.4,5.6,3.7,5.2,3.9,4.2,
3.5,6.6,6.2,2.0,0.8,2.0,3.7,4.7,7.3,1.6,3.8,0.5,0.7,4.5,2.2,4.0,6.5,
1.2,4.5,1.7,1.8,1.4,2.6,0.2,0.7,11.5,5.0,1.2,14.1,4.0,2.7,1.6,3.5,2.8,
0.7,8.6)
mean(times)
median(times)
```

Output:

```
> mean(times)
[1] 4.335294
> median(times)
[1] 3.7
```

Example 2.3.1: Cricket singing times

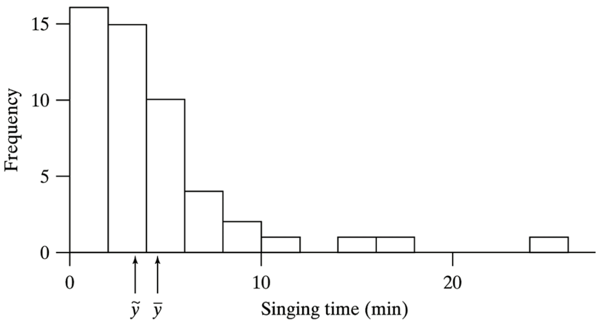


Figure 2.3.4 Histogram of cricket singing times

Figure: $n = 51$ cricket singing times; mean pulled toward right tail – the direction of skew – more than median.

Mean versus median

- Median may make more sense for skewed data, i.e. may be more typical.
- Mean annual U.S. household income in 2004 is \$60,500. Median is \$43,300. The millionaires pull the mean higher than the median.
- Also the median can be computed in some situations where the mean cannot.
- Example: survival times. The median can be computed as soon as half the experimental units are dead. The mean needs all units dead.

Quartiles

- The median cuts the data in half; half the observations are smaller and half larger.
- If we look at the lower half of the data, the **first quartile** Q_1 cuts the lower half in two.
- The **third quartile** Q_3 cuts the upper half in two.
- Q_1, the median, and Q_3 cut the data into four parts with roughly equal numbers of observations.
- Q_1 is the median of the lower half; Q_3 is the median of the upper half.

Example 2.4.2 Pulses

$n = 12$ college student pulses were measured (beats per minute)

$$62\ 64\ 68\ 70\ 70\ 74\ 74\ 76\ 76\ 78\ 78\ 80$$

- Since n is even, the median is given by median $= \frac{74+74}{2} = 74$.
- Q_1 is the median of the lower half

$$62\ 64\ 68\ 70\ 70\ 74$$

$Q_1 = \frac{68+70}{2} = 69$.

- Q_3 is the median of the upper half

$$74\ 76\ 76\ 78\ 78\ 80$$

$Q_3 = \frac{76+78}{2} = 77$.

The interquartile range

- The **interquartile range**, IQR, is $\text{IQR} = Q_3 - Q_1$.
- For the pulse data, $\text{IQR} = 77 - 69 = 8$ bpm.
- The IQR gives the length of an interval containing the middle 50% of the data. It measures how "spread out" the data are.
- Half of the 12 students pulses lie in an interval of length 8 bpm.

Minimum, maximum, and five number summary

- The **maximum** of the sample, max, is the largest value.
- The **minimum** of the sample, min, is the smallest value.
- The **five number summary** is min, Q_1, median, Q_3, max.
- The **range** of the data is max-min.
- For the pulses, the five number summary is

$$\text{min} = 62, Q_1 = 69, \text{median} = 74, Q_3 = 77, \text{max} = 80.$$

- The range of the pulses is $80 - 62 = 18$ bpm.

Boxplots

- The five number summary can be placed on an x-axis to give a “snapshot” of the data.
- A boxplot simply places a box around Q_1 to Q_3 and draws lines or “whiskers” from Q_1 to the min, and Q_3 to the max.
- Gives a visual representation of a typical value (the median), the spread of the middle 50% (the box) and the spread of the whole data set (the whiskers) all at once.

Example 2.4.3: Radish growth

The length (mm) of $n = 14$ radish shoots grown in total darkness over three days from seeds is

Table 2.4.1 Radish growth, in mm, after three days in total darkness

15	20	11	30	33
20	29	35	8	10
22	37	15	25	

The five number summary (p. 48) is

$$\text{min} = 8, Q_1 = 15, \text{median} = 21, Q_3 = 30, \text{max} = 37.$$

Example 2.4.3: Radish growth

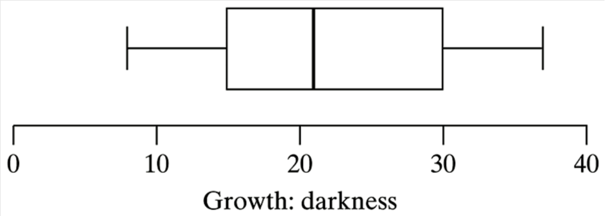


Figure: Boxplot of radishes grown in darkness

Outliers & modified boxplots

- Outliers are observations that are really small or really large and far away from the bulk of the data.
- Many data sets do not have outliers; many do.
- We formally define an outlier to be any observation that is
  - *Smaller* than $Q_1 - 1.5 \times \text{IQR}$, or
  - *larger* than $Q_3 + 1.5 \times \text{IQR}$.
- These numbers are called the **lower** and **upper fences.**
- A **modified boxplot** plots outliers separately and only extends the whiskers as far out as the largest and smallest *non-outlying* observations.
- The default boxplot in R is a modified boxplot.

Example 2.4.5: Radish growth, constant light

$n = 14$ radishes were also grown in *constant* light over three days. Their lengths are

$$\begin{array}{cccccccccc} \text{min} & & Q_1 & & \tilde{y} = 9.5 & & Q_3 & & \text{max} \\ \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ 3 & 5\ 5 & 7 & 7\ 8\ 9 & & 10\ 10\ 10 & 10 & 14\ 20 & 21 \end{array}$$

- Compute IQR $= 10 - 7 = 3$.
- The lower fence is $Q_1 - 1.5 \times \text{IQR} = 7 - 1.5(3) = 2.5$.
- The upper fence is $Q_3 + 1.5 \times \text{IQR} = 10 + 1.5(3) = 14.5$.
- There are no observations smaller than 2.5 but there are two larger than 14.5: 20 and 21.
- 20 and 21 are *outliers*

Modified boxplot for radishes grown in constant light

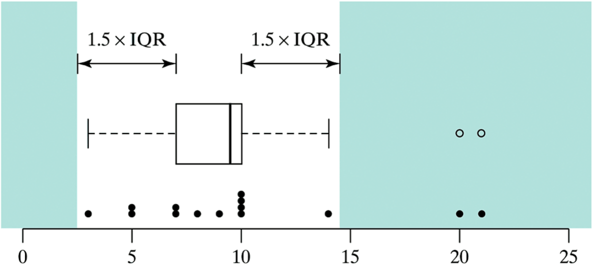


Figure: Dotplot & boxplot of radishes grown in constant light

Example 2.4.1: Radish growth

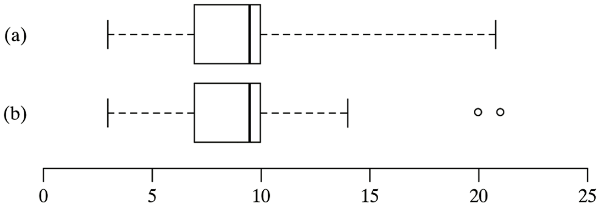
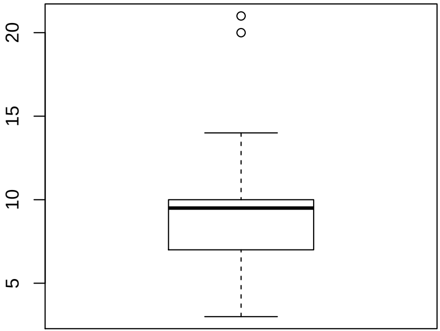


Figure: Radishes grown in constant light; boxplot and modified boxplot.

Various R functions

```
r=c(3,5,5,7,7,8,9,10,10,10,10,14,20,21)
boxplot(r)
mean(r)
median(r)
quantile(r,0.25) # 1st quartile is 25th percentile
quantile(r,0.75) # 3rd quartile is 75th percentile
min(r)
max(r)
```

R's boxplot



Review questions

- What is difference between bar chart and histogram?
- Can a distribution be both skewed and symmetric?
- Can a bimodal distribution be symmetric?
- What do outliers do to the mean relative to the median?
- What is the five number summary? How do these numbers relate to a boxplot?
- What is the definition of an outlier?
- A distribution is skewed to the left; which tail is longer?