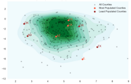
Chapter 2 Notes

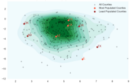
Describing Quantitative Distributions with Numbers
STP-231

Arizona State University



Introduction
Key Terms

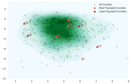
- A **Statistic** is a numerical measure calculated from sample data
- **Descriptive Statistics** define characteristics of data that describe it. Observational unit inference is made (if possible)
- We will look at the mean, median, and the mode



Sample Median

Split the ordered data into two equal halves

- Half of observations in the sample are above, half below
- To find the median, called $\tilde{y}$, rearrange the values of the data set in ascending order.
- For an odd number of observations, the median is the middle value, located at the $n/2$ element in the list
- If even, the median is the average of the 2 middle values.



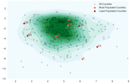
Sample Median
Examples

2.3.3 from Statistics for the Life Sciences by Myra Samuels 2016

- A researcher applied the carcinogenic compound benzo(a)pyrene to the skin of five mice, and measured the concentration in the liver tissue after 48 hours. The results (nmol/gm) were as follows:

  6.3 5.9 7.0 6.9 5.9

  Find the median



Sample Mean

Aka the average

- The sum of all the values divided by the total number of observations (where y_i is an observation)

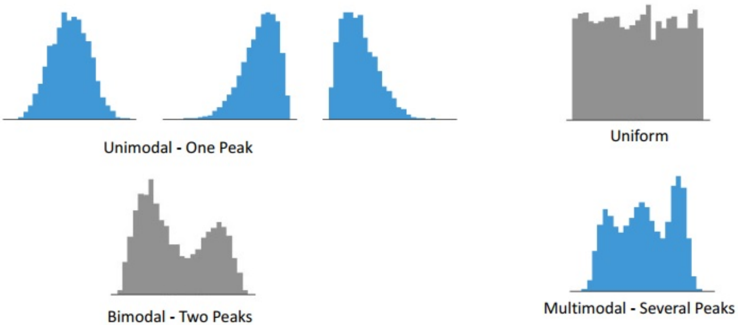
$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

- For example, from the previous example

$$\overline{y} = \frac{1}{5}(6.3 + 5.9 + 7.0 + 6.9 + 5.9) = 6.4$$

Sample Mode

The most frequent value(s)



Example

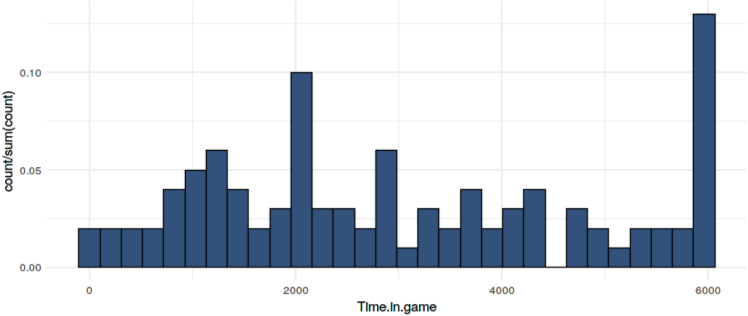
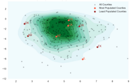
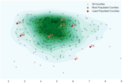


Figure: Lacrosse penalty data over time. Where would the mean, median, mode be? Which is most useful?



Resistant Statistics

- Resistant if extreme values have little to no influence on its outcome
- Median is better than mode in this regard
- However, if skewed, both will be pulled towards the longer tail
- Mean is usually pulled more than the median



Resistant Statistics
Example

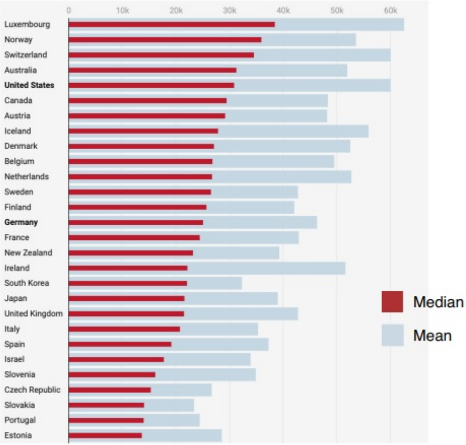
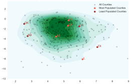


Figure: Median and mean income by countries, 2012/2014 (PPP).
https://blog.datawrapper.de/weekly-chart-income/



Deviation

- Difference between a sample data point and the mean of the sample

$$y_i - \overline{y}$$

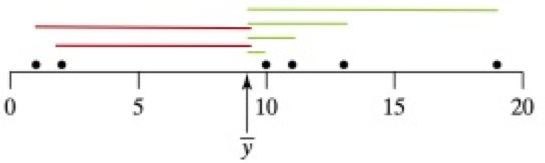
- The mean is uniquely defined as the value that “balances” the deviations.

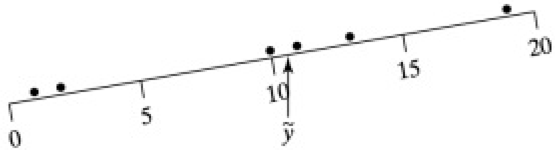
$$\sum_{i=1}^{n}(y_i - \overline{y}) = 0$$

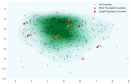
$$\begin{aligned}
&y_1 - \frac{1}{n}(y_1 + y_2 + \ldots + y_n) + \ldots + y_n - \frac{1}{n}(y_1 + y_2 + \ldots + y_n) \\
&= \left(\frac{ny_1}{n} - \frac{1}{n}(y_1 + y_2 + \ldots y_n)\right) + \ldots + \left(\frac{ny_n}{n} - \frac{1}{n}(y_1 + y_2 + \ldots y_n)\right) \\
&= \frac{y_1(n-1)}{n} - \left(\frac{y_2}{n} + \ldots \frac{y_n}{n}\right) + \ldots + \frac{y_n(n-1)}{n} - \left(\frac{y_1}{n} + \ldots \frac{y_{n-1}}{n}\right) \\
&= \frac{(n-1)y_1-(n-1)y_1}{n} + \frac{(n-1)y_1-(n-1)y_2}{n} + \ldots + \frac{(n-1)y_n-(n-1)y_n}{n} = 0
\end{aligned}$$

Deviation
Continued

- Consider the lamb data



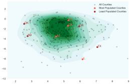




Example

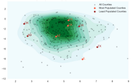
- An experiment of a weight loss drug finds that the group receiving the drug lost on average 10 pounds more than the placebo group, but a median value of about 7 pounds lost.
- Can you explain whats going on?
- How would you find the mode in this example?
- Individual cases vary, but on **average** a participant on the drug would expect to lose 10 pounds

Measures of Spread



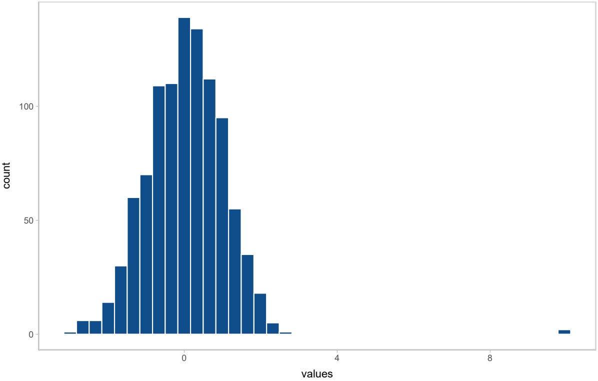
Measures of Spread

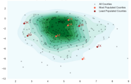
- We are still discussing statistics rather than parameters.
- A histogram is defined by not only its center, but its spread as well
- We can look at quartiles and standard deviation (as well as variation)



The Range

- The simplest way to define the spread, the distance from minimum to maximum…great unless we have outliers





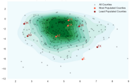
Quartiles
Examples

Note, if odd number of observations, do not include median in quartile calculations. If even, include the left side of median in Q_1 calculation, right median for Q_3 calculation.

- Q_1: median of all data to the left of overall median Q_2 (25th percentile)
- Q_3: median of the data to the right of the overall median Q_2 (75th percentile)
- Interquartile range (IQR):

  $$\text{IQR} = Q_3 - Q_1$$

  Contains 50% of data.
- The five number summary
- Minimum, first quartile, median, third quartile, maximum



Quartiles
Example

Find the quartiles and IQR

- In a study of milk production in sheep (for use in making cheese), a researcher measured the 3-month milk yield for each of 11 ewes. The yields (in litres) were as follows[1]:

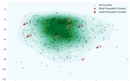
  56.5 89.8 110.1 65.6 63.7 82.6 75.1 91.5 102.9 44.4 108.1

  Order them as:

  $$44.4, 56.5, 63.7, 65.6, 75.1, 82.6, 89.8, 91.5, 102.9, 108.1, 110.1$$

  The median is 82.6, the 1st quartile is 63.7, and the last is 102.9. We find the median to the left of the median and the median to the right of the median, excluding the median!

[1] Statistics for Life Sciences, Myra Samuels 2016



Quartiles
Example

Find the quartiles and IQR with an even number of data points

- The following is a list of 12 wait times at the grocery store (in minutes)

$$1, 2, 4, 4, 5, 8, 9, 10, 12, 13, 14, 15$$

- The first quartile is the median of

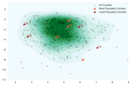
$$1, 2, 4, 4, 5, 8 = 4$$

- The third quartile is the median of

$$9, 10, 12, 13, 14, 15 = (12 + 13)/2 = 12.5$$

- The median is

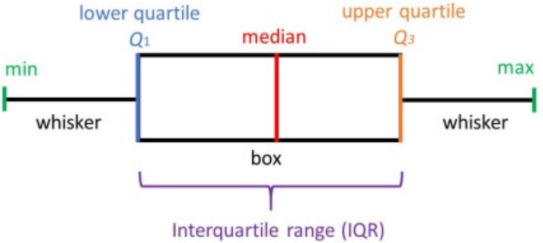
$$(8 + 9)/2 = 8.5$$

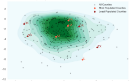


Box Plot

Aka box and whisker plot

- Used as a visual representation of the five-number summary





Outliers

- Observations that are far outside the overall pattern
- Worthy of investigation. Data entry error, strange phenomena? Even if we can explain them, could be an issue when analyzing data
- How do we decide what is classified as an outlier?
- Define lower and upper fences

Five Number Summary

0.4 1.6 4.1 6.0 17.1

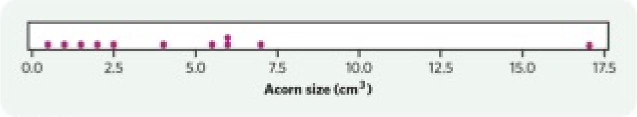
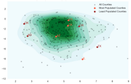
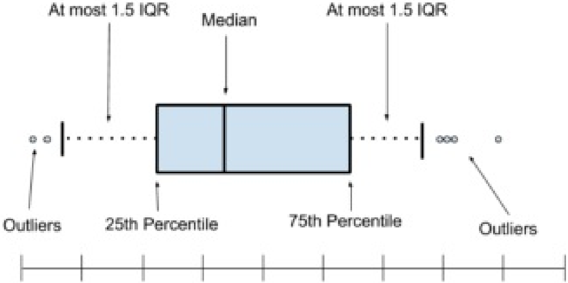


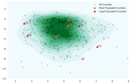
Figure 2.6
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018 W. H. Freeman and Company



Modified Box Plot

- Box remains in its normal position
- Whiskers are updated, by drawing lower whiskers to the lowest data that remains above the lower fence
- Draw upper whiskers to highest data point that remains below upper fence

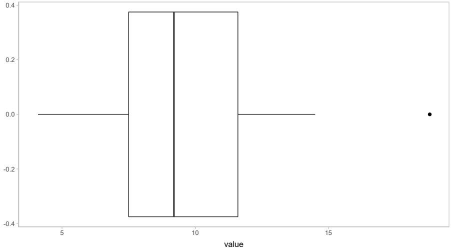


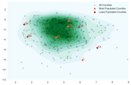


Example

MAO (Monoamine Oxidase) enzyme levels of 18 people were measured. The results (expressed as number of moles of benzaldehyde product per 108 platelets):

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
y_i	4.1	5.2	6.8	7.3	7.4	7.8	7.8	8.4	8.7	9.7	9.9	10.6	10.7	11.9	12.7	14.2	14.5	18.8

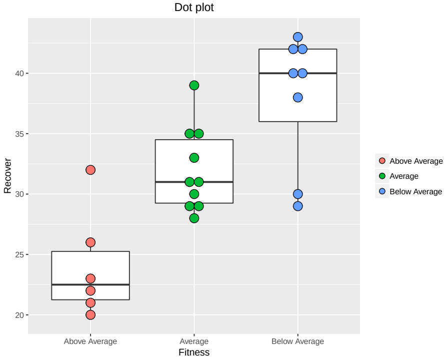


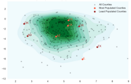


Example
Using Box plot with dot plot

Below is a table of recovery time of people after exercise based on their self-described fitness level

Recovery time	fitness level
29.0	below average
42.0	below average
⋮	⋮
33.0	average
⋮	⋮
22.0	above average





Histogram vs box plot

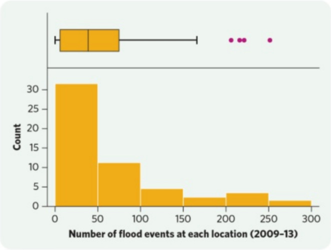
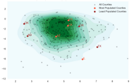


Figure 2.7
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 2018
W. H. Freeman and Company

- Histograms show modality, skewness, range, and spread
- Boxplots show exact median, exaxt 50% spread, and easier to see (suspected) outliers



Sample Standard Deviation (s)

- Roughly defined as the average distance between a point and the sample mean The standard deviation is the square root of the variance.

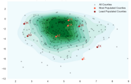
$$s = \sqrt{s^2} = \sqrt{\text{variance}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}}$$

- From a frequency distribution:

$$s = \sqrt{\frac{\sum_{i=1}^{n} f_i \times (y_i - \overline{y})^2}{n-1}}$$

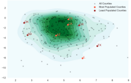
- And the **Coefficient of Variation**

$$\text{Coefficient of Variation} = \frac{s}{\overline{y}}$$



Properties of standard deviation

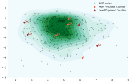
- Does not exist for a sample with a single point
- Represented in the units of the variable
- Less resistant than IQR more resistant than the range, but still pretty affected by outliers and skewness
- The square makes the variance more sensitive because of the square operation



Calculate some Values

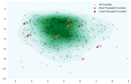
Plaza Gallery in Chicago sells home furniture online and is concerned some of their shipments to customers arriving late. Five days were randomly selected and the number of late shipment complaints were recorded. The observations were 3,5,5,6,6,8. Find the range, sample variance, and sample standard deviation for these data. R-code to calculate quickly:

```
data<-c(3,5,5,6,6,8)
var_ex<-var(data); print(var_ex)
sd_ex<-sd(data); print(sd_ex)
range_ex<-range(data); print(range_ex)
```



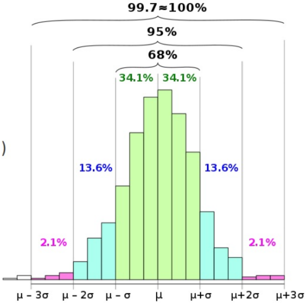
Degrees of Freedom

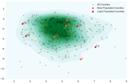
- Maybe you are wondering why we divide by $n-1$ not n for the standard deviation estimate
- The sum of deviations is always 0
- Then $n-1$ observations can vary freely but one is constrained since they must sum to zero, so only $n-1$ terms contriute to information about deviation
- Therefore, dividing by $n-1$ estimates the true population variance from the sample



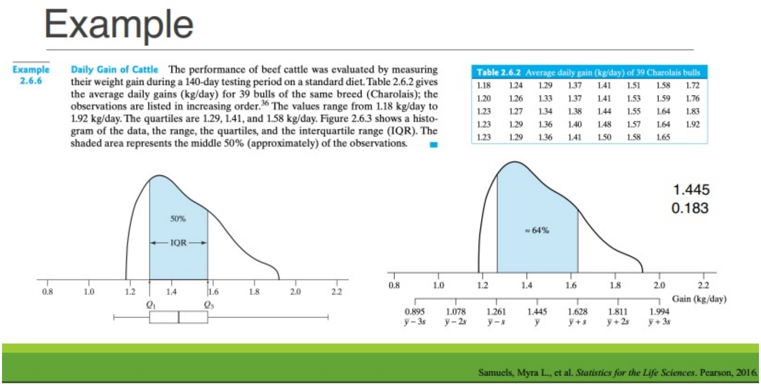
Empirical Rule for a Sample

- For any sample of observations with a symmetric and unimodal distribution (and big enough n), we expect to find
  - 68% of all values fall within $(\overline{y} - s, \overline{y} + s)$
  - 95% of all values fall within $(\overline{y} - 2s, \overline{y} + 2s)$
  - 99.7% of all values fall within $(\overline{y} - 3s, \overline{y} + 3s)$





Example



Example

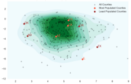
Example 2.6.6 **Daily Gain of Cattle** The performance of beef cattle was evaluated by measuring their weight gain during a 140-day testing period on a standard diet. Table 2.6.2 gives the average daily gains (kg/day) for 39 bulls of the same breed (Charolais); the observations are listed in increasing order.[36] The values range from 1.18 kg/day to 1.92 kg/day. The quartiles are 1.29, 1.41, and 1.58 kg/day. Figure 2.6.3 shows a histogram of the data, the range, the quartiles, and the interquartile range (IQR). The shaded area represents the middle 50% (approximately) of the observations. ■

Table 2.6.2 Average daily gain (kg/day) of 39 Charolais bulls

1.18	1.24	1.29	1.37	1.41	1.51	1.58	1.72
1.20	1.26	1.33	1.37	1.41	1.53	1.59	1.76
1.23	1.27	1.34	1.38	1.44	1.55	1.64	1.83
1.23	1.29	1.36	1.40	1.48	1.57	1.64	1.92
1.23	1.29	1.36	1.41	1.50	1.58	1.65	

1.445
0.183

Samuels, Myra L., et al. *Statistics for the Life Sciences*. Pearson, 2016.



Example II

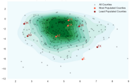
2.6.11 from the book

- Listed in increasing order are the serum creatine phophokinase (CK) levels (U/I) of 36 healthy men (these are the data from example 2.2.6):

(***)25	62	82	95	110	139
42	64	83	95	113	145
48	67	84	100	118	151
*57	68	92	101	119	163
58	70	93	104	121	201
60	78	94	110	123	203

The sample mean CK level is 98.3 U/I and the SD is 40.4 U/I. What %-age of the observations are within

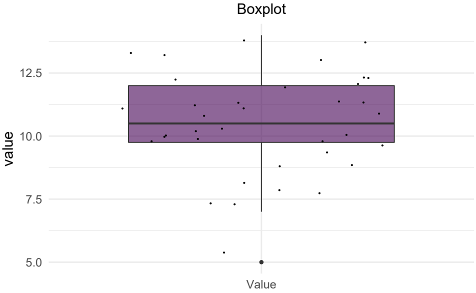
- 1 SD of the mean? 26/36
- 2 SDs of the mean? 34/36
- 3 SDs of the mean? 36/36



Example

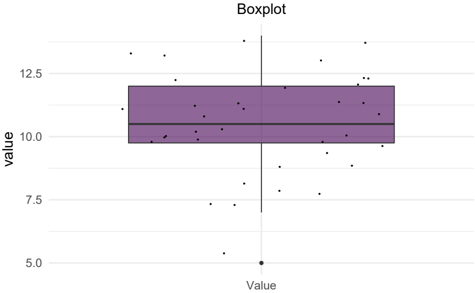
Find the mean, standard deviation, and construct a boxplot:

Number of piglets	Frequency (number of sows)	$f_i * y_i$
5	1	5
6	0	0
7	2	14
8	3	24
9	3	27
10	9	90
11	8	88
12	5	60
13	3	39
14	2	28
	$n = 36$	$\sum y_i = 375$



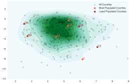
Example

Number of piglets	Frequency (number of sows)
5	1
6	0
7	2
8	3
9	3
10	9
11	8
12	5
13	3
14	2



(a) $s = 1.99$ and $\overline{y} = 10.42$

Transformations



Data Transformations

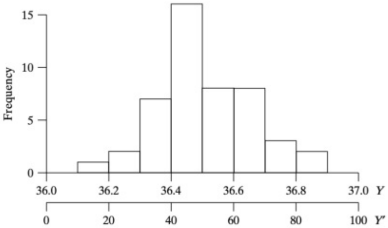
- Big part of any statistical analysis
- Why? Maybe you need to change the scale, maybe you wanna transform the shape (i.e. to go from exponential to linear), or maybe you wanna change your units

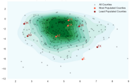
Linear Transformations

- Generally, of the classical linear form y=mx+b
- Does not change the shape of the distribution
- Scale data by multiplication/division or shift data by adding/subtracting (or both)

Figure 2.7.1 Distribution of 47 temperature measurements showing original and linearly transformed scales

$$Y' = (Y - 36) \times 100$$





Linear Transformations
Continued

- We can add or subtract a constant to the original variable

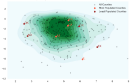
$$\text{If } Y' = Y \pm C \implies \overline{y}' = \overline{y} \pm C \text{ and } s' = s$$

- Multiply by a constant to the original variable

$$\text{If } Y' = mY \implies \overline{y}' = m \cdot \overline{y} \text{ and } s' = m \cdot s$$

- Multiply by a constant to the original variable and then add/subtract a constant

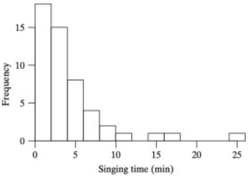
$$\text{If } Y' = mY \pm C \implies \overline{y}' = m \cdot \overline{y} \pm C \text{ and } s' = m \cdot s$$

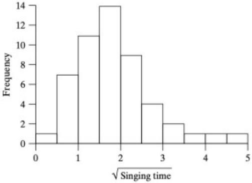


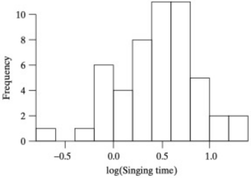
Other Transformations

$$Y' = \sqrt{Y}$$
$$Y' = \ln(Y)$$
$$Y' = \frac{1}{Y}$$
$$Y' = Y^2$$

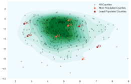
Square root and log transformations pull right tail inward and push out left tail





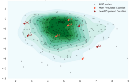


Statistical Inference



Inference

- **Statistical Inference**: Methods for making predictions on population based on data from a sample
- Goal: results from any sample would be identical or nearly identical to the results obtained from the population
- The population has its own distribution
- For a simple random sample, the sample distribution approximates the population distribution
- The larger the sample size, the better the approximation
- Statistics describes a sample. Parameters describe a population



Inference (continued)

- Statistics desribe sample characteristics, $\overline{y}$ and s are sample mean and sample standard deviation
- Parameters describe population characteristics. Usually trying to estimate these
- Proportions: are a relative frequency, can be for population or sample

Measure	Statistic	Parameter
Proportion	$\hat{p}$	p
Mean	$\overline{y}$	μ
Standard deviation	s	σ